Outil d'aide au diagnostic du cancer à partir d'extraction d'informations issues de bases de données et d'analyses par biopuces

Lyamine Hedjazi

Université de Toulouse

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :
Université Toulouse III Paul Sabatier (UT3 Paul Sabatier)

Discipline ou spécialité :

Systèmes Automatiques

Présentée et soutenue par :
Lyamine Hedjazi

le : jeudi 8 décembre 2011

Titre :

Outil d'aide au diagnostic du cancer à partir d'extraction d'informations issues de bases de données et d'analyses par biopuces

---

Ecole doctorale :
Systèmes (EDSYS)

Unité de recherche :
LAAS-CNRS

Directeur(s) de Thèse :

Florence Dalenc et Marie-Véronique Le Lann

Rapporteurs :

Sylvie Galichet
Sylvie Charbonnier

Membre(s) du jury :

Isabelle Bloch
Boutaib Dahhou
Gilles Favre
Yijun Sun
Joseph Aguilar-Martin (invité)

Aknowledgments

Les travaux présentés dans cette thèse ont été effectués au Laboratoire d'Analyse et d'Architecture des Systèmes du Centre National de la Recherche Scientifique (LAAS-CNRS) au sein du groupe Diagnostic, Supervision et Conduite (DISCO). Je tiens pour cela à remercier les directeurs successifs du LAAS, Messieurs Raja CHATILLA, Jean-Louis SANCHEZ et Jean ARLAT de m'avoir accueilli au LAAS-CNRS.

Je tiens à remercier aussi Madame Louise TRAVE-MASSUYES, responsable du groupe DISCO, de m'avoir accueilli dans son groupe de recherche pendant ces années de thèse.

J'exprime toute ma gratitude à ma Directrice de thèse, Madame Marie-Véronique LE LANN, pour avoir encadré mes travaux de recherche, et à qui cette thèse doit beaucoup. Je la remercie pour sa disponibilité constante et efficace, son soutien et ses encouragements. Ma reconnaissance va aussi à ma co-directrice de thèse, Madame Florence DALENC, pour ses conseils précieux, sa patience et la confiance qu'elle m'a accordée pour mener à terme ces travaux de recherche. Je tiens également à remercier Monsieur Joseph AGUILAR-MARTIN pour ses remarques, ses conseils, sa disponibilité et son soutien tout au long de cette thèse. Mes remerciements vont aussi à Madame Tatiana KEMPOWSKY-HAMON pour l'aide qu'elle a apportée pour la réalisation de ces travaux de thèse.

Lire et juger une thèse n'est pas une tâche aisée à accomplir. Aussi, je tiens tout particulièrement à exprimer ma reconnaissance à tous les membres du jury de ma thèse:

- Boutaib DAHHOU, professeur à l'Université Paul Sabatier, pour avoir présidé ce jury.
- Sylvie GALICHET, Professeur à l'Université de Savoie, et Sylvie CHARBONNIER, Maître de Conférences à l'Université Joseph Fourier, pour m'avoir fait l'honneur d'accepter la lourde tâche de rapporter sur cette thèse.
- Isabelle BLOCH, Professeur à Télécom Paris Tech, et Gille FAVRE, Professeur et Praticien hospitalier à l'Université Paul Sabatier, d'avoir accepté de juger mes travaux de thèses et pour leurs remarques pertinentes.
- Yijun SUN, assistant scientist (University of Florida), pour sa lecture pertinente du manuscrit de thèse et pour m'avoir accueilli dans son laboratoire (ICBR) dans le cadre d'un séjour de recherche.

Un GRAND merci à ma famille qui m'a soutenu pendant 28 ans de prés et de loin. Finalement, je ne saurai conclure sans remercier mes amis pour leur soutien inconditionnel.

Abstract

Cancer is one of the most common causes of death in the world. Currently, breast cancer is the most frequent in female cancers. Although the significant improvement made last decades in cancer management, an accurate cancer management is still needed to help physicians take the necessary treatment decisions and thereby reducing its related adverse effects as well as its expensive medical costs. This work addresses the use of machine learning techniques to develop such tools of breast cancer management.

Clinical factors, such as patient age and histo-pathological variables, are still the basis of day-to-day decision for cancer management. However, with the emergence of high throughput technology, gene expression profiling is gaining increasing attention to build more accurate predictive tools for breast cancer. Nevertheless, several challenges have to be faced for the development of such tools mainly (1) high dimensionality of data issued from microarray technology; (2) low signal-to-noise ratio in microarray measurement; (3) membership uncertainty of patients to cancer groups; and (4) heterogeneous (or mixed-type) data present usually in clinical datasets.

In this work we propose some approaches to deal appropriately with such challenges. A first approach addresses the problem of high data dimensionality by taking use of ℓ_1 learning capabilities to design an embedded feature selection algorithm for SVM (ℓ_1 SVM) based on a gradient descent technique. The main idea is to transform the initial constrained convex optimization problem into an unconstrained one through the use of an approximated loss function. A second approach handles simultaneously all challenges and therefore allows the integration of several data sources (clinical, microarray …) to build more accurate predictive tools. In this order a unified principle to deal with the data heterogeneity problem is proposed. This principle is based on the mapping of different types of data from initially heterogeneous spaces into a common space through an *adequacy* measure. To take into account membership uncertainty and increase model interpretability, this principle is proposed within a fuzzy logic framework. Besides, in order to alleviate the problem of high level noise, a symbolic approach is proposed suggesting the use of interval representation to model the noisy measurements. Since all data are mapped into a common space, they can be processed in a unified way whatever its initial type for different data analysis purposes. We particularly designed, based on this principle, a supervised fuzzy feature weighting approach. The weighting process is mainly based on the definition of a *membership margin* for each sample. It optimizes then a

membership-margin based objective function using classical optimization approach to avoid combinatorial search. An extension of this approach to the unsupervised case is performed to develop a weighted fuzzy rule-based clustering algorithm. The effectiveness of all approaches has been assessed through extensive experimental studies and compared with well-know state-of-the-art methods. Finally, some breast cancer applications have been performed based on the proposed approaches. In particular, predictive and prognostic models were derived based on microarray and/or clinical data and compared with genetic and clinical based approaches.

Résumé

Le cancer est l'une des causes les plus fréquentes de décès dans le monde. Actuellement, le cancer du sein est le plus répandu dans les cancers féminins. Malgré les avancées significatives faites ces dernières décennies en vue d'améliorer la gestion du cancer, des outils plus précis sont toujours nécessaires pour aider les oncologues à choisir le traitement nécessaire à des fins de guérison ou de prévention de récidive tout en réduisant les effets néfastes des ces traitements ainsi que leurs coûts élevés. Ce travail porte sur l'utilisation de techniques d'apprentissage automatique pour développer de tels outils de gestion du cancer du sein.

Les facteurs cliniques, tels que l'âge du patient et les variables histo-pathologiques, constituent encore la base quotidienne de prise de décision pour la gestion du cancer du sein. Cependant, avec l'émergence de la technologie à haut débit, le profil d'expression génique suscite un intérêt croissant pour construire des outils plus précis de prédiction du cancer du sein. Néanmoins, plusieurs challenges doivent être relevés pour le développement de tels outils, principalement: (1) la dimensionnalité des données issues de la technologie des puces, (2) le faible rapport signal sur bruit dans la mesure de biopuces, (3) l'incertitude d'appartenance des patients aux différents groupes du cancer, et (4) l'hétérogénéité des données présentes habituellement dans les bases de données cliniques.

Dans ce travail, nous proposons quelques approches pour surmonter de manière appropriée de tels challenges. Une première approche aborde le problème de haute dimensionnalité des données en utilisant les capacités d'apprentissage dit normé $\ell 1$ pour la conception d'un algorithme de sélection de variables intégré à la méthode SVM (machines à vecteurs supports), algorithme basé sur une technique de gradient. Une deuxième approche permet de gérer simultanément tous les problèmes, en particulier l'intégration de plusieurs sources de données (cliniques, puces à ADN, ...) pour construire des outils prédictifs plus précis. Pour cela, un principe unifié est proposé pour surmonter le problème de l'hétérogénéité des données. Pour tenir compte de l'incertitude d'appartenance et augmenter l'interprétabilité du modèle, ce principe est proposé dans le cadre de la logique floue. Par ailleurs, afin d'atténuer le problème du bruit de niveau élevé, une approche symbolique est proposée suggérant l'utilisation de la représentation par intervalle pour modéliser les mesures bruitées. Nous avons conçu en particulier, basée sur ce principe, une approche floue supervisée de

pondération de variables. Le processus de pondération repose essentiellement sur la définition d'une marge d'appartenance pour chaque échantillon. Il optimise une fonction objective basée sur la marge d'appartenance afin d'éviter la recherche combinatoire. Une extension de cette approche au cas non supervisé est effectuée pour développer un algorithme de regroupement automatique basé sur la pondération des règles floues.

L'efficacité de toutes les approches a été évaluée par des études expérimentales extensives, et comparée avec des méthodes bien connues de l'état de l'art. Enfin, un dernier travail est consacré à des applications des approches proposées dans le domaine du cancer du sein. En particulier, des modèles prédictifs et pronostiques ont été extraits à partir des données de puces à ADN et/ou des données cliniques, et leurs performances comparées avec celles d'approches génétiques et cliniques existantes.

Contents

List of Figures

List of Tables

Introduction- Résumé

Le cancer du sein est actuellement le plus fréquent des cancers féminins. Dans le monde, chaque année, l'on compte plus de 1 050 000 de nouveaux cas diagnostiqués et plus de 400 000 décès causés par le cancer du sein. Rien qu'en France, il est prévu que près de 53 000 nouveaux cas de cancer du sein seront diagnostiqués et que 11 500 patientes mourront du cancer du sein en 2011 (Institut de Veille Sanitaire, 2011). Malgré les avancées significatives faites ces dernières décennies en vue d'améliorer la gestion du cancer, des outils de diagnostic et de pronostic plus précis sont encore nécessaires pour aider les oncologues à choisir le traitement nécessaire à des fins de guérison ou de prévention de récidives

La gestion du cancer du sein peut se résumer en trois problèmes principaux: diagnostic, pronostic et prédiction de bénéfice thérapeutique. Bien que le diagnostic du cancer du sein puisse être entièrement assuré par des outils d'imagerie médicale, le pronostic et la prédiction du bénéfice thérapeutique semblent être des tâches plus difficiles. En effet, à cause de l'hétérogénéité et la complexité de la maladie du cancer, les patients avec les mêmes symptômes auraient des évolutions de cancer très différentes. Les approches traditionnelles sont basées principalement sur un petit ensemble de variables cliniques et histo-pathologiques. Cependant, ces outils de pronostique et de prédiction sont loin d'être parfaits et des modèles plus précis sont nécessaires pour améliorer la gestion du cancer du sein.

L'émergence de technologies à haut débit dans la dernière décennie, comme la technologie des biopuces (puces à ADN), a rendu possible la mesure simultanée de l'expression de milliers de gènes. Ces technologies ont apporté avec elles l'espoir de gagner de nouveaux aperçus sur la biologie du cancer et d'améliorer les outils actuels de gestion du cancer. Cependant, ces technologies ont aussi apporté avec elles de sérieux challenges liés aux caractéristiques intrinsèques des données produites telles que: (1) la grande dimensionnalité des données et (2) la nature bruitée des mesures. Toutefois, l'incertitude de mesure n'est pas le seul type d'incertitude auquel on est confronté lorsque l'on veut appliquer les méthodes d'apprentissage automatique à des problèmes réels. En raison de la grande complexité de la maladie du cancer du sein, la tumeur d'un patient peut en effet appartenir simultanément à des groupes moléculaires différents de cancer avec un certain degré d'appartenance. Par ailleurs, pour éviter le problème du faible nombre de patients sur lesquels on dispose d'informations, il serait préférable d'utiliser l'ensemble des bases de données issues de

biopuces disponibles. Néanmoins, cela soulève plusieurs problèmes tels que la différence entre les populations et les technologies biopuces utilisées nécessitant la prise en compte de l'incertitude d'appartenance dans le processus de décision. Vu que les méthodes statistiques traditionnelles sont mal adaptées pour faire face à de tels problèmes, les méthodes d'apprentissage automatique ont été choisies comme une bonne alternative pour surmonter ces challenges.

Des études récentes ont démontré la valeur potentielle de la signature d'expressions génétiques dans l'évaluation du risque de récurrence de la maladie post-chirurgicale. Cependant, ces études tentent de développer des outils de pronostique basés sur des marqueurs génétiques pour remplacer les critères cliniques existants, ce qui suggère que chaque approche doit être utilisée indépendamment. De plus, le fait qu'en procédant de cette manière nous occultons complètement la richesse des informations contenues dans les marqueurs cliniques établies durant des décennies de recherche sur le cancer, les cliniciens peuvent faire face à la situation critique où le patient a un critère pathologique clinique en contradiction avec le résultat fourni par la signature génétique. Une approche typique alors serait d'intégrer les deux types d'informations (cliniques et l'expression des gènes) dans le processus de prise de décision. Cependant, en plus des défis indiqués précédemment liés principalement aux données de biopuces, d'autres dilemmes tels que l'hétérogénéité des données caractérisant les données cliniques doivent être confrontés pour intégrer à la fois les deux types d'information. Les facteurs cliniques utilisés pour la description de l'état du patient sont en effet généralement représentés de différentes manières selon la perception des médecins.

Par conséquent, ce qui est vraiment nécessaire pour améliorer la gestion du cancer actuelle est le développement d'approches d'apprentissage automatique capables d'aborder tous les problèmes indiqués ci-dessus. Pour résumer, trois défis doivent être principalement confrontés: le premier est lié à la dimensionnalité élevée dans les données en particulier celles issues de la technologie des biopuces, le second est le problème du bruit et des incertitudes associés généralement aux données alors que le dernier est lié à la présence de données de type mixte dans les bases de données cliniques. C'est l'ensemble de ces problèmes que nous avons abordés dans cette thèse dans un cadre d'apprentissage automatique dans le but de concevoir des outils de gestion plus précis du cancer pour aider les médecins dans leur décision.

Introduction

Cancer is one of the most common causes of death in the world. Due to the rapid increase in cancer cases, cancer will soon replace heart disease as the leading cause of deaths worldwide. Currently, breast cancer is the most frequent in female cancers. In the world, each year, the there are more than 1 050 000 new diagnosed cases and more than 400 000 deaths caused by breast cancer. In France alone, it is expected that around 53 000 new breast cancer cases will be diagnosed and 11 500 will die from breast cancer in 2011 (Institut de Veille Sanitaire, 2011). Although the significant improvement made last decades in cancer management, an accurate cancer diagnosis and prognosis is still needed to help physicians take the necessary treatment decisions and thereby reducing its related adverse effects as well as its expensive medical costs.

Breast cancer management can be summarized by three main issues: diagnosis, prognostication and prediction of therapy benefit. An early breast cancer diagnosis improves the chances of cure and may avoid distant metastasis development, i.e. development of new tumors in different organs. A prognostic tool would enable the physicians to forecast the likely course of the disease (e.g. Relapse or Remission) and therefore spare patients from unnecessary anti-cancer toxic treatments such as chemotherapy. A predictive tool would enable however to predict the tumor response to a particular treatment and therefore to prescribe the optimal tailored treatment for each patient. Although breast cancer diagnosis can be fully assured by imaging modalities and computer-aided detection tools, breast cancer prognostication and prediction of therapy benefit seems to be more challenging tasks. Due indeed to the high cancer heterogeneity and complexity, patients with the same symptoms would have very different evolutions and outcome.

Traditional approaches are based mainly on a small set of clinical and histo-pathological variables (e.g. tumor size and lymph node status). However, these prognostic and predictive tools are far from perfect and more accurate models are needed to improve breast cancer management. Clinician practitioners have rapidly grasped the urgent need of new accurate tools as well as a good understanding of the biological mechanisms involved in breast cancer progression.

The emergence of high throughput technologies in the last decade, such as microarray technology, has made possible the simultaneous measurement of the expression of thousands

of genes. These technologies have carried with them the hope to gain new insights into cancer biology and improve current tools for cancer management. Meanwhile, these technologies have brought with them also serious challenges related to intrinsic characteristics of the issued data such as: (1) high data dimensionality (thousands of gene expressions for few a number of samples); and (2) the noisy nature of measurements. Since traditional statistical methods are ill-conditioned to deal with such problems, machine learning approaches have been picked up as a good alternative to overcome these difficulties.

However, measurement uncertainty is not the only type of uncertainty to be faced with in real-world problems by machine learning approaches. Due to the high complexity of breast cancer disease, a patient's tumor can belong simultaneously to many cancer groups with some degree of membership. Moreover, to alleviate the problem of small sample size, it would be preferable to use all available microarray datasets. Nevertheless, this raises several problems such as the difference among populations and the use of different microarray technologies requiring the consideration of membership uncertainty in the decision making process.

Recent studies have demonstrated the potential value of gene expression signature in assessing the risk of post-surgical disease recurrence. However, these studies attempt to develop genetic marker-based prognostic tools to replace the existing clinical criteria, suggesting that each approach should be used independently. Besides the fact that by doing so we are ignoring the rich information contained in clinical markers established over decades of cancer research, clinicians can face the critical situation where the patient has a clinical pathological criterion in contradiction with the gene signature outcome. One typical approach would be to integrate both types of information (clinical and gene-expression) in the decision-making process. However, in addition to the challenges stated previously related mainly to microarray data, other dilemmas should be faced to integrate both information, such as data heterogeneity in clinical data. Clinical features used for patient state description are generally represented in different ways according to the physician perception (one may note for example the age for a patient by a quantitative value (e.g. age= 35) whereas another prefers a symbolic value (e.g. age< 35)).

Therefore, what is really needed to improve current cancer management is developing machine learning approaches capable of handling all above stated challenges. To summarize, three challenges are mainly faced: the first one is related to high dimensionality in data especially issued from microarray technology, the second one is the problem of noise and

uncertainties associated usually to both data whereas the last one is related to the presence of mixed-type data in daily produced clinical datasets. Addressing efficiently those problems is urgently needed provided that in some cancer applications the three challenges can be even faced simultaneously (e.g. integration of clinical and microarray data). Indeed, in order to improve the accuracy of current predictive tools recent trends in bioinformatics and biomedicine are directed towards the integration of increasing numbers of sources of data. This thesis addresses such problems within a machine learning framework with the aim to design more accurate cancer management tools to help the physicians in their decision making process.

This work is the result of a collaboration which has been initiated 4 years ago between the group DISCO of LAAS and the Institut Claudius Regaud first of all through a common PhD scholarship obtained after competition from Université Paul Sabatier («bourse dite du Président» and the participation to the project named ONCOMATE, (labelled by the fondation INNABIOSANTE). This project aimed to develop a novel technological platform for the detection of cancer marker proteins, by combining three major technologies: molecular imprints of target marker proteins into sugar hybrid polymers, a label free sensor chip based on diffraction of light by nanoscale structures, and machine learning algorithms fed with the screening of a cancer tissues database with full anonymous patient records.

The manuscript is organized as follows:
The first chapter provides a brief overview about the most important tasks in breast cancer management: cancer diagnosis, prognosis and prediction of treatment benefit. We briefly describe their evolution over decades of cancer research and their challenging aspects from medical point of view. We explain their medical aspects and the approaches usually used to deal with them.

The second chapter reviews the state-of-the-art of machine learning in cancer research. We have described the three machine learning tasks mostly used in cancer management: supervised classification, clustering and feature selection. A few examples of the most known approaches for each task are briefly described by highlighting their advantages and drawbacks. Then some application examples of such approaches in breast cancer management are provided. This chapter ends with a description of the recent challenges that have to be faced to improve cancer management and treatment. In particular, we give further details

about the problems of data heterogeneity, high dimensionality, low signal-to-noise ratio and membership uncertainties.

The third chapter addresses the problem of data dimensionality by taking advantage of ℓ_1 learning capabilities. We particularly propose an embedded feature selection approach for SVM problem using gradient descent techniques without resorting to any dual formulation. The basic idea is the transformation of the initial SVM convex optimization problem into unconstrained non-convex one. The non differentiable property of the hinge loss function has been overcome by using its approximated Huber loss function. We show that this approach guarantees the global optimality of the solution while exhibiting a good computational efficiency compared to other approaches solving the same problem. Large-scale numerical experiments have been conducted to demonstrate these claims.

In chapter four we consider to deal simultaneously with the problems of data heterogeneity and membership uncertainty. In this order a unified principle, referred to as SMSP (Simultaneous Mapping for Single Processing), is introduced to cope with the problem of data heterogeneity within a fuzzy logic framework. This principle is based on a simultaneous mapping of data from initially heterogeneous spaces into only one homogeneous space using an appropriate measure of typicality (or membership). Once the heterogeneous data are represented in a unified space, only a single processing for various analysis purposes such as machine learning tasks can be performed. We considered the three most used types of features: (1) quantitative; (2) interval; and (3) qualitative.

In chapter five the problem of supervised learning based on the SMSP principle is addressed. A new feature weighting method is proposed for mixed-type and high dimensional data based on a *membership margin* to improve the performance of fuzzy-rule based classifiers. For this reason, a weighted fuzzy rule concept is introduced and a membership margin-based objective function is defined. A classical optimization approach is used to avoid heuristic combinatorial search. Large-scale experiments have been also conducted to compare the proposed approach with some well-known feature selection approaches on three state-of-the-art classifiers.

In chapter six the problem of unsupervised learning based on the SMSP principle is considered. We propose a novel approach based on online feature weighting for clustering of heterogeneous data. The proposed algorithm is an extension of our supervised feature weighting algorithm. To cope with the problem of data heterogeneity, the SMSP principle is

extended here also to reason in a unified way about heterogeneous data in an unsupervised framework. An extensive experimental study has been then performed on artificial and real-world problems to prove the effectiveness of the proposed approach.

Finally, some breast cancer applications of the proposed approaches are shown in chapter seven. In particular, the works presented here develop (1) Cancer prognosis based only on clinical data (2) Derivation of 20 genes signature for cancer prognosis based on microarray data (3) Derivation of a hybrid signature for cancer prognosis based on the integration of clinical and microarray data (4) Derivation of a more robust prognostic signature (referred to as GenSym) based on a symbolic approach by modeling the different noises as symbolic intervals (5) Derivation of 4-markers signature for the prediction of neoadjuvant treatment benefit in HER2 over-expressed breast cancer patients.

CHAPITRE 1-Résumé
Gestion et traitement du cancer

Le cancer est l'une des causes les plus fréquentes de décès dans le monde. Selon la dernière édition du rapport mondial sur le cancer (World Cancer Report WCR) de l'Agence internationale de recherche sur le cancer, dû à l'augmentation rapide des cas de cancer, le cancer va bientôt remplacer les maladies cardiaques comme la principale cause de décès dans le monde. WCR prévoit que 12,4 millions de personnes seront diagnostiquées avec certaines formes de cancer chaque année et que 7,6 millions de personnes en mourront.

Les cancers les plus courants dans le monde en termes d'incidence ont été: poumon (1,52 millions de cas), sein (1,29 million) et colorectal (1,15 millions). En raison de son mauvais pronostic, le cancer du poumon a également été la cause la plus fréquente de décès (1,31 millions), suivi par le cancer de l'estomac (780 000 décès) et le cancer du foie (699 000 décès). Nous nous concentrons dans notre travail sur le cancer du sein comme l'une des tumeurs malignes les plus fréquemment diagnostiquées chez les femmes.

Des stratégies de gestion du cancer sont nécessaires de toute urgence pour réduire la morbidité et la mortalité par cancer, et améliorer la qualité de vie des patients atteints de cette maladie. Des travaux de recherche considérables ont été réalisés ces dernières décennies dans l'espoir d'apporter de nouvelles perspectives à la maitrise de la biologie du cancer et l'amélioration des approches utilisées actuellement pour la gestion du cancer en particulier celui du sein.

La gestion du cancer du sein peut se résumer en trois tâches principales successives:

1- *La détection précoce et le diagnostic efficace du cancer,*
2- *Une pronostication efficace pour prédire le risque de développer des métastases (de nouvelles tumeurs dans les différents organes) sans traitement systématique,*
3- *Le choix d'un traitement optimal et personnalisé en fonction de l'agressivité du cancer en prédisant le bénéfice thérapeutique.*

Le premier chapitre de cette thèse décrit chaque tâche et donne brièvement leur évolution au cours des décennies de recherche sur le cancer. Nous avons essayé de souligner leurs aspects

difficiles d'un point de vue médical en expliquant les questions d'intérêt et les approches habituellement utilisées pour les traiter.

Malgré les nombreuses tentatives effectuées en matière de recherche sur le cancer, il peut être constaté que la tâche de diagnostic de cancer du sein est toujours basée principalement sur des outils de détection par imagerie (Hayat, 2008). Cependant, contrairement au diagnostic de cancer, les tâches de pronostic et de prédiction du bénéfice d'un traitement ont connu une véritable révolution au cours des dernières décennies. Les approches traditionnelles utilisées pour effectuer ces deux tâches ont été basées essentiellement sur l'utilisation des connaissances qualitatives acquises au cours de plusieurs décennies de recherche sur le cancer. Cette connaissance est formulée généralement sous la forme de règles en fonction de certains facteurs cliniques tels que l'âge, le grade histologique et le statut des récepteurs hormonaux. On note parmi ces approches l'indice NIH adopté aux Etats Unis (Eifel et al., 2001) et le critère de St-Gallen en Europe (Goldhirsh et al., 2003). Des modèles plus sophistiqués ont été aussi proposés tels que Adjuvant! (Olivotto et al., 2005) et l'indice de pronostic de Nottingham (NPI (Galea et al., 1992)) et sa version améliorée (Belle et al., 2010). Ces approches ne parviennent pas néanmoins à fournir une gestion précise du cancer. Cependant, l'introduction des nouvelles technologies de pointe récemment ont permis d'obtenir quelques éclaircissements sur les processus biologiques qui sous-tendent la grande hétérogénéité du cancer du sein. En particulier, la technologie des biopuces a largement marqué la recherche sur le cancer pendant le siècle courant ouvrant la porte à une prise en charge adaptée et personnalisée du cancer du sein en se basant sur l'extraction des signatures génétiques moléculaires. Les travaux de grands impacts inclus mais ne se limitent pas aux signatures d'Amsterdam (Van't Veer et al., 2002), et de Rotterdam (Wang et al., 2005a) pour la tâche pronostic et la signature de prédiction de survie sans rechute pour la tâche de prédiction (Ma et al., 2004).

Toutefois, ces progrès significatifs en terme de technologie ont amené avec eux de sérieux défis liés à l'énorme quantité de données produites par ces technologies et ont requis également une révolution similaire en termes d'approches permettant de traiter ces données. Cela fera l'objet du chapitre suivant dans lequel nous nous concentrons sur l'analyse de l'une des approches les plus utilisées (méthodes d'apprentissage automatique) pour effectuer les trois tâches de gestion du cancer. Sur cette base, nous décrivons les enjeux récents liés à ce domaine qui feront les problématiques que nous abordons dans cette thèse.

CHAPTER 1
Cancer Management and Treatment

Cancer is one of the most common causes of death in the world. According to the last edition of the World Cancer Report (WCR) from the International Agency for Research on Cancer, due to rapid increase in cancer cases, cancer will soon replace heart disease as the leading cause of deaths worldwide. WCR projected that 12.4 million people will be diagnosed with some forms of cancer each year and 7.6 million people will die. WCR said: “The global cancer burden doubled in the last 30 years of the 20th century, and it is estimated that this will double again between 2000 and 2020 and nearly triple by 2030”.

According to WCR, 26.4 million people per year may be diagnosed with cancer by 2030, with 17 million people dying from it. There will be 1% increase in cancer incidences each year, with larger increases in China, Russia, and India. Adoption of tobacco use and higher-fat diets and demographic changes, including a projected population increase of 38% in less-developed countries between 2008 and 2030 are the main reasons of increase in cancer cases in these countries.

The most common cancers in the world in terms of incidence were: lung (1.52 million cases), breast (1.29 million) and colorectal (1.15 million). Because of its poor prognosis, lung cancer was also the most common cause of death (1.31 million), followed by stomach cancer (780 000 deaths) and liver cancer (699 000 deaths). We focus in our work on breast cancer as one of the most frequently diagnosed malignancy in women in the world.

Cancer management strategies are needed urgently to reduce the morbidity and mortality from cancer, and to improve the quality of life of cancer patients. Tremendous research works were performed last decades in the hope to bring new insights to cancer biology and improving current approaches for breast cancer management.

Breast cancer management can be summarized in three main successive tasks:

- Early detection and efficient cancer diagnosis,
- Efficient prognostication to predict the risk to develop metastases (new tumors in different organs) without systematic treatment,

- Selection of an optimal and personalized treatment according to cancer aggressiveness by predicting the therapy benefit.

Usually traditional clinical tools are used to perform such tasks based on histo-pathological factors such patient age and tumor size. However, recently new advanced high throughput technologies, such as gene expression profiling through microarray, have being introduced extensively in this field.

This chapter describes each task and gives briefly their evolution over decades of cancer research. We try highlights their challenging aspects from medical point of view. We explain the medical questions of interest and the approaches usually used to deal with them.

1.1 Cancer detection and diagnosis

Early cancer detection plays a key role in decreasing the death rates from cancer and achieves a better prognosis (Hayat, 2008). Indeed, different sources (Institut National du Cancer, 2011; Association pour la Recherche sur le Cancer, 2011) shows that breast cancer treatment in an early stage of development can increase significantly the patient's survival chance. Moreover, early breast cancer detection increases the chances for conservative surgery to be carried out instead of radical mastectomy, the only solution in advanced stage breast cancers (Haffty *et al.*, 1991). However, the main aim of this task should not be only to detect the existence of the cancer but also to identify the cancer class among the pre-established classes and discover new cancer subclasses. For decades many techniques were proposed to perform an accurate breast cancer diagnosis.

Usually, the most used technique is based on imaging detection tools (e.g. mammography is considered as the most cost-effective method for detecting breast cancer (Hayat, 2008). However, due to the complex structure of the breast, thousands of mammograms must be processed to detect a few cancers (Gallardo-Caballero *et al.*, 2007). This task can be tedious and stressful, and can cause radiologist confusion leading to diagnosis errors (Hayat, 2008). Moreover, despite the availability and recommended use of mammography as a routine screening method for women older than 50 years of age, it is still inefficient and insufficient to identify accurately the cancer class (Antman and Shea , 1999; Hayat, 2008). For that other techniques that could be used individually or in combination with existing modality for cost-effective screening of breast cancer have been investigated.

In addition, there is a wide spectrum in cancer morphology and many tumors are atypical or lack morphologic features that are useful for differential diagnosis. Therefore, with the increasing need of an accurate detection of cancer, the search is on for reliable markers that will be clinically helpful in the diagnosis of small tumors. To this end, a large number of blood tumor markers have been proposed for breast cancer detection, including CEA (CarcinoEmbryonic Antigen), ESR (Erythrocyte Sedimentation Rate) (Cheung *et al.*, 2000; Li *et al.*, 2002), but have not been well adopted in clinical practices. However, due to the high cancer heterogeneity the currently accepted clinical diagnostic markers fall short to classify the disease in subtypes and there is a critical need to identify novel diagnostic markers (Golub *et al.*, 1999). Golub and co-authors pointed out that cancer classification task can be divided into two challenges: class discovery and class prediction. Class discovery refers to defining previously unrecognized tumor subtypes whereas class prediction refers to the assignment of particular tumor samples to already-defined subtypes (or classes). Therefore, reliable markers are required to gain new insights into cancer biology and can be clinically helpful in the diagnosis of small tumors.

It has been found out recently that cancer diseases including breast cancer result from the accumulation of mutations, chromosomal instabilities and epigenetic changes that together facilitate an increased rate of cellular evolution and damage that progressively impairs the cell's detailed and complex regulation system of cell growth and death. This fact has motivated cancer researchers initially to investigate the importance of one or only few genes at a time in order to improve cancer detection and diagnosis (Matsumura and Tarin, 1992). Although hundreds of such studies have pointed out differences in the expression of one or few genes, no one of them have provided a comprehensive study of gene expression in cancer cells (Zhang *et al.*, 1997; Ramaswamy *et al.*, 2001). Recent advances in high throughput technologies, such as microarray and mass spectrometry (see Appendix 1), have made it possible to answer such questions through simultaneous analysis of the expression patterns of thousands of genes and proteins (Golub *et al.*, 1999; Ramaswamy *et al.*, 2001; Li *et al.*, 2002). These technologies are considered promising for gaining new insights into cancer biology.

1.2 Cancer prognosis

After the diagnosis of breast cancer, the next important step is the prognosis which aims to predict the survival of a patient, or her risk to develop metastases without treatment (Figure 1.1) (Haibe-Kains, 2009). Roughly speaking, prognosis attempts to accurately forecast the

evolution or outcome of a specific situation (e.g. Relapse or Remission) using input information obtained from a concrete set of variables that potentially describe that situation (Gómez-Ruiz *et al.*, 2004). Naturally, this task depends strongly of the diagnosis task presented previously, as an accurate diagnosis will allow giving some information about the likely evolution of the disease. Moreover, this can be also extremely important because it assists oncologists, as described in the next subsection, to select the optimal treatment required for a breast cancer patient chemo-, hormone-, or other systematic therapies; and which patient can be treated with loco-regional treatment alone (Haibe-Kains, 2009).

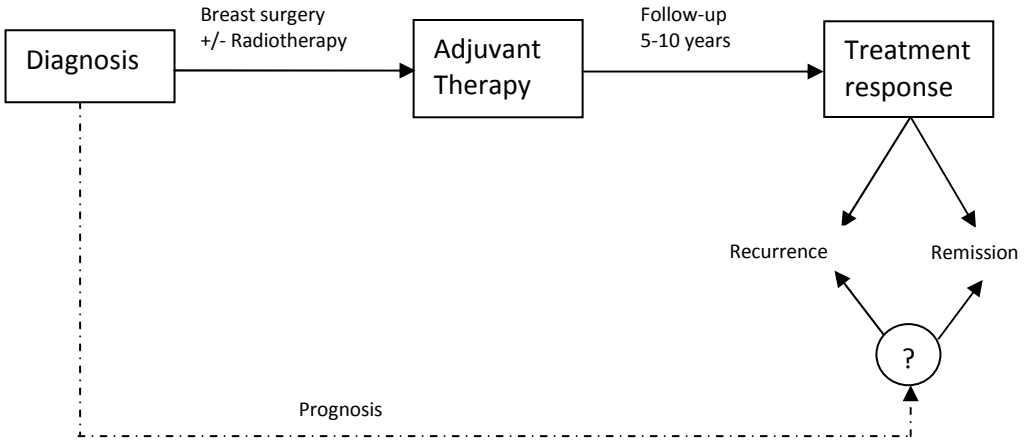


Fig. 1.1. Breast cancer prognosis.

Similarly to cancer diagnosis, many approaches were proposed in the literature to perform cancer prognosis. For a long time cancer prognosis was guided by the clinical and histo-pathological knowledge gained from many decades of cancer research. In this approaches, the risk of recurrence is primarily determined by the age of the patient, nodal status, tumor size, histological grade, the expression status of hormonal receptors, i.e. estrogen (ER) and the progesterone (PgR) as quantified by immunohistochemistry (IHC), the status of HER2 oncogene, vascular emboli, proliferation index and histologic type (Haibe-Kains, 2009) (See Glossary and Appendix 1 for definitions). Many cancer prognosis criterions were proposed based on these variables; among them we find the National Institute of Health index for USA (Eifel *et al.*, 2001) and the St Gallen consensus criteria (Goldhirsh *et al.*, 2003) for Europe in order to assist clinicians in their decision-making (see Figure 1.2). However, using only one variable at a time (e.g. histological grade) has been found insufficient and not accurate enough (Perez *et al.*, 2006). To improve prognosis accuracy, more sophisticated models based on a combination of these variables has been also proposed such as multivariable outcome prediction models (e.g. Adjuvant! (Olivotto *et al.*, 2005)) and the Nottingham Prognostic Index (NPI (Galea *et al.*, 1992)) and its improved version (Belle *et al.*, 2010)) (see Figure

1.2). However, the prognosis accuracy is far from perfect and more accurate models are needed before it will be possible to clearly identify whether a patient will relapse, especially patients with early breast cancer (node-negative, i.e. nodal status equal to 0), to spare them from receiving unnecessary systematic therapy as well as reduce its related expensive medical costs. It is reported that a third of breast cancer patients are over treated which makes them undergo its side effects in the short and long terms. On the contrary a more moderated number undergoes an under treatment by underestimating their distance recurrence and therefore they are wrongly spared from systemic adjuvant treatment. Moreover, two patients with exactly the same clinical and pathological characteristics can have different outcomes. Therefore, a more accurate prognosis could avoid any adverse side effects of adjuvant therapies and its related high costs.

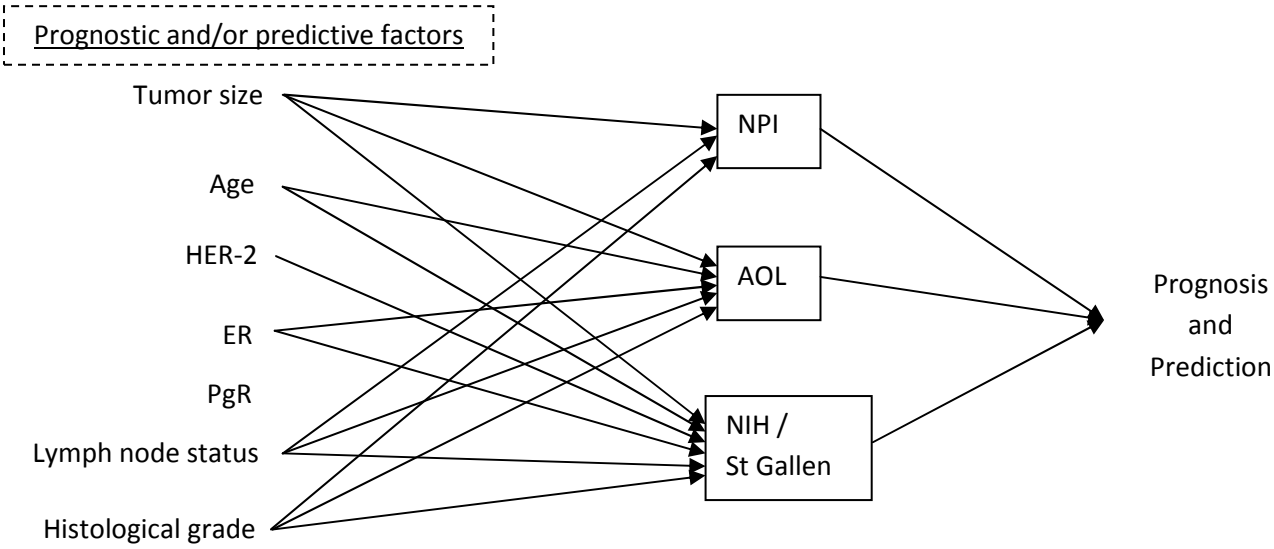


Fig. 1.2. Traditional prognostic and predictive tools for breast cancer.

To this aim, firstly two protein biomarkers known as uPA/PAI-1 (for respectively Urokinase-type Plasminogen Activator and Plasminogen Activator Inhibitor type1) have been shown to be effective to identify subclasses of patients as a function of their recurrence risk (Janicke *et al.*, 2001). It has been shown in a retrospective study that these biomarkers have a superior prognostic power than other classical factors (Age, hormone receptors, Grade). Meanwhile, recent advances in high throughput technologies have also open the door to new research orientations in this field aiming to achieve a more accurate cancer prognosis. Indeed, clinical investigators have found out that these technologies can not only be useful to gain new insights into cancer biology but can also be a powerful prognostic tool. Unlike traditional clinical variables which are usually limited to few, these technologies give the advantage to

provide simultaneously the expression differentiation of thousands of genes in the aim to derive prognosis models based only on a set of genetic markers.

A first outstanding work in this direction was performed by the Netherlands Cancer Institute (NKI) which has conducted a comprehensive study in order to derive a more accurate tool for early breast cancer prognosis (Van't Veer *et al.,* 2002). In this work the Agilent microarray technology was used to extract a set of genes differentially expressed among two groups of patients having different survival outcomes. First group include patients that have developed a distant metastases within five years from diagnosis whereas the second does not. In this study a set of genes was identified including mainly genes involved in the cell cycle, invasion, metastasis, and angiogenesis. This signature is known under the name of "Amsterdam genomic signature" and enables to classify node-negative breast cancer patients, with a tumor size inferior or equal to 5 cm (stage I or II) and aged less than 61, either in a high or a low risk group. A supplementary study was also performed on a new large population of patients from the same institution, including both node-negative, node-positive, treated and untreated breast cancers, to validate the predictive power of this signature (Van de Vijver *et al.*, 2009). This signature was also compared to classical clinical criterions (i.e. NIH, St Gallen). In this study the authors have shown the superiority of this genetic signature, compared to NIH and St Gallen criterions, in term of predictive power of patients' outcomes. In the conclusion of this work it has been pointed out that this predictive ability could spare a large number of patients to be over-treated or to receive unnecessary toxicity from chemotherapy.

In recent studies, many attempts were also performed in the same direction to identify new gene signatures. A gene-expression signature known as the Reccurence Score signature include only 21 genes has been derived allowing to refine the stratification of ER-positive and Node-negative breast cancer patients receiving tamoxifen in adjuvant setting (Paik *et al.*, 2004). Three risk levels have been defined: weak risk, intermediate risk and high risk. We distinguish also the signature known as the Rotterdam signature (Wang *et al.*, 2005a), where 76 genes have been identified for the same purpose of that designed by (Van't Veer *et al*, 2002) for node-negative patients who did not receive a systematic treatment. However, this study was performed using Affermetrix technology and has been shown to better identify patients with poor prognosis compared to classical clinical criterions.

Although the major contribution of such retrospective studies to open new directions for cancer practitioners, they should be still validated in prospective by randomized trials to

obtain a sufficient LOE (Level Of Evidence) and therefore be used in routine practices (Institut National du Cancer, 2009). A randomized clinical trial, called MINDACT for 'Microarray In Node negative Disease may Avoid ChemoTherapy', is ongoing to be performed in order to compare the predictive accuracy of the Amsterdam signature with clinical and pathological criteria such as adjuvant! Online, to identify women with node-negative breast cancer with a low risk of relapse. Another randomized trial called TAILORx is also now under consideration to validate the Recurrence Score signature by putting them in competition with classical factors having a level of evidence LOE 1 such as ER, HER-negative, and uPA/PAI-1. The results expected from these studies will attribute to these signatures a predictive power of level LEO 1 required for clinical implementations (Institut National du Cancer, 2009).

Although the potential of the studies presented above in breast cancer research, several critical reviews can be found in literature about such genomic approaches (Reis-Filho *et al.*, 2006; Koscielny, 2008). For instance, Reis-Filho *et al* (2006) have pointed out that the clinicians may face the situation where the patient has a clinical pathological criterion corresponding to poor prognosis and a good gene signature. One typical approach would be to integrate both types of information (clinical and gene-expression) in the decision-making process which has been shown recently effective in improving the prognosis tasks (Gevaert *et al.*, 2006; Sun *et al.*, 2007a).

1.3 Systemic treatment responsiveness prediction

The prediction task aims to predict the response of a breast cancer patient to a treatment. In other words, for each patient, we need to decide which therapy will be the most effective. To this end, as in the case of cancer prognosis and diagnosis, this task consists also to identify a set of markers that could predict response of a given patient to a particular drug (predictive factors). This would spare patients from receiving unnecessary treatment and decrease its associated medical and financial cost. We can distinguish two settings for treatment responsiveness prediction in breast cancer: adjuvant (Figure 1.3) and neo-adjuvant settings (Figure1.4) (Mauri *et al.*, 2005).

In the last decade, the systematic adjuvant treatment is usually prescribed in the aim to decrease the recurrence risk of breast cancer patients. An important consequence of such procedure is overtreatment resulting from the administration of adjuvant therapy to patients for whom only a surgery would be sufficient (Straver *et al.*, 2009). This leads mainly to

expose the patients to adverse side effect of the treatment while increasing its associated cost. In this case, the prediction is similar to the prognosis task as illustrated in Figure 1.1, except that a treatment is selected for each patient by the end. Precisely, in this setting we try to predict whether the administration of a particular adjuvant therapy to a patient after surgery will be beneficial after some years of follow-up (generally more than 5 years). However, although the response to a treatment in advanced breast cancer can be assessed by tumor measurement in this case, it is still relatively difficult to be characterized in the case of early stage breast cancers after surgery (Chang *et al.*, 2005). An accepted practice in this case is to administrate adjuvant chemotherapy even if we know that it is not beneficial for a significant number of patients (Chang *et al.*, 2005).

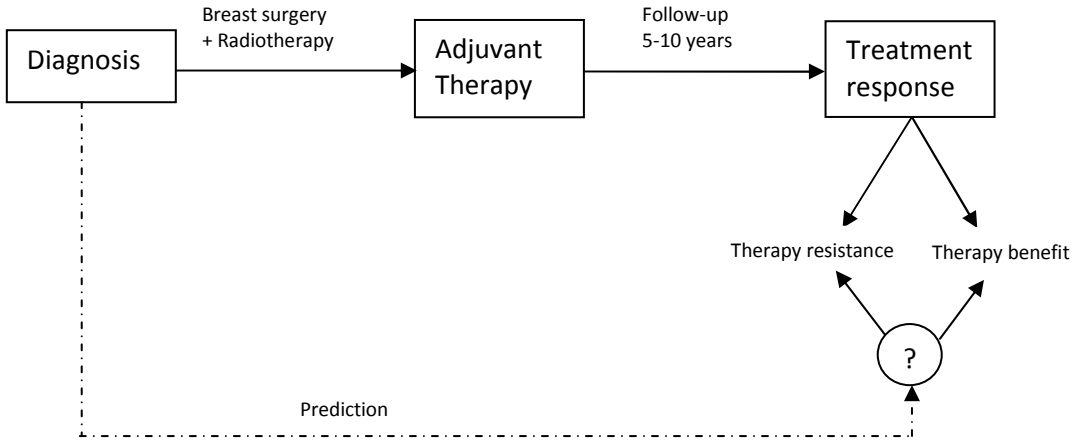


Fig. 1.3. Adjuvant setting for prediction of treatment benefit

With respect to the neo-adjuvant setting, a biopsy of breast cancer is firstly performed before the administration of the neo-adjuvant therapy (pre-operative therapy including chemotherapy and hormone therapy, Figure 1.4). Then the tumor is removed by surgery to assess the benefit from the treatment such as a decreased tumor size and axillary lymph nodes. Indeed, although the fact that both settings (adjuvant and neoadjuvant) were reported equivalent in terms of survival and overall disease progression, neoadjuvant therapy was found to be a safe approach allowing to avoid mastectomy in a significant number of women (Makris *et al.*, 1998; Cleator *et al.*, 2004; Mauri *et al.*, 2005). The benefit from a treatment for patients in these cases is usually characterized in terms of pathological complete response (cPR) defined as the complete disappearance of cancer cells in the breast and lymph nodes. Even of the fact that the concern in this case is to analyze the response or resistance to the treatment without paying much attention to the survival issue, it has been pointed out that the response to some neoadjuvant therapies (e.g. chemotherapy) correlates closely with improved clinical outcome (Fisher *et al.*, 1998).

However, in both settings the identification of a set of biomarkers that predicts the response to treatment accurately is not an easy task. The fact is that over 20 years of cancer research for new important markers, we still have very few biomarkers that predict accurately the response to particular therapies. Mainly there are two biomarkers used actually in the day-to-day clinical practice: Hormone Receptors (HR: Estrogen Receptor ER and Progestrone Receptor PgR) and HER2/ ERBB2 receptor (Chang *et al.*, 2005; Colozza *et al.*, 2005). Hormone receptors are effective factors for prediction of hormonotherapy response whereas HR-negative is considered as a powerful predictive factor of chemotherapy response in the neoadjuvant setting. HER2-positive enables to predict the patient responsiveness to anti-HER2 treatments. Although several attempts were also performed to identify additional biomarkers, they are still to date unconvincing due especially to the huge heterogeneity of breast cancer (Konecny *et al.*, 2004; Colozza *et al.*, 2005).

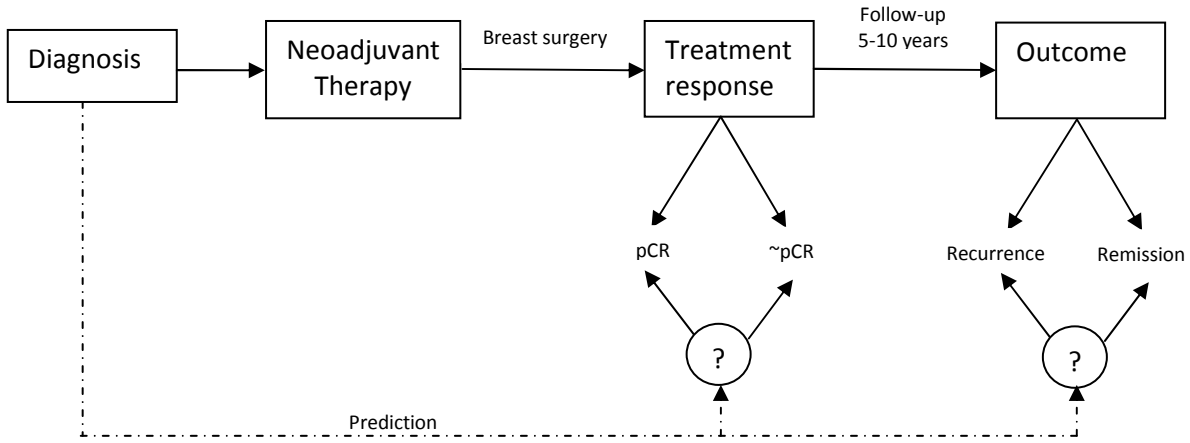


Fig. 1.4. Neoadjuvant setting for prediction of treatment benefit

Similarly to cancer prognosis and diagnosis, these limitations have pushed the cancer researchers to take advantage of the genomic approaches to develop more accurate markers that predict the response to particular regimens. For the adjuvant setting, at least two gene signatures can be found in literature derived by gene expression profiling. The, first one concerns the prediction of relapse-free survival (RFS, see Glossary for details) developed by (Ma *et al.*, 2004) whereas the second is the 16-gene signature which can predict the risk of recurrence in patients receiving adjuvant tamoxifen (Wang *et al.*, 2005a). Both signatures were derived using a set of patients treated with adjuvant hormonotherapy, i.e. treated after surgery, which enables to address the prognosis issue (appearance of metastases) as well as the prediction of response to treatment. Another gene-expression signature also to be mentioned is known by the Reccurence Score signature include only 21 genes for

hormonotherapy responsiveness prediction (Paik *et al.*, 2004). Concerning the neo-adjuvant setting, Chang *et al.* (2003a) have used gene expression profiling to derive 92-gene signature that predict the response to neoadjuvant docetaxel in primary breast cancer patients. A neoadjuvant approach was also assessed to analyze the change of gene expression during chemotherapy (Buchholz *et al.*, 2002). Another neoadjuvant study has also reported a 74-gene markers signature using microarray data to predict some therapies (Ayers *et al.*, 2004). These encouraging results have strongly suggested that microarray profiling will have a promised future in the optimal neoadjuvant treatment selection. Several works have been reported recently within the neoadjuvant setting framework (Lee *et al.*, 2007, Straver *et al.*, 2009). In (Straver *et al.*, 2009), the predictive capacity of the 70-gene signature (Van't Veer *et al.*, 2002) has been assessed on neoadjuvant chemotherapy treatment in breast cancer.

1.4 Conclusion

In this chapter we provided an overview about the main tasks in breast cancer management: diagnosis, prognostication and prediction of treatment benefit. We briefly described each task and its most important research works. Their challenging aspects have been also highlighted from medical point of view.

Although the many attempts performed in cancer research fields, breast cancer diagnosis task is still based mainly on imaging detection tools. However, unlike cancer diagnosis, prognosis and treatment response prediction tasks have known a real revolution over the last decades. Traditional approaches to perform both tasks have been based mainly on using the qualitative knowledge gained over many decades of cancer research. This knowledge is reformulated usually on the form of rules about some clinical factors but fails short to provide an accurate cancer management. However, advanced technologies have made it possible to get some insights into the biological process underlying the high heterogeneity of breast cancer. Particularly, microarray technology has widely marked the cancer research in the current century by opening the door to tailored and personalized management of breast cancer based on molecular signatures derivation.

However, such advancements have brought with them serious challenges related to the huge amount of data issued by these technologies, and thereby required also a similar revolution in terms of approaches enabling to process this data. For that, in the next chapter we focus on the reviewing of one of the most used approaches (Machine Learning approaches) to perform the

three cancer management tasks. Based on that, we describe the recent challenges related to this field which will make the concerns of the present thesis.

CHAPITRE 2- Résumé
Méthodes par apprentissage pour la gestion et le traitement du cancer

La gestion du cancer et le choix de son traitement adéquat ont été pour longtemps effectués sur la base de connaissances qualitatives retenues par des experts ou en utilisant les diverses directives médicales. Toutefois, la maladie du cancer s'est avérée complexe et très hétérogène ce qui rend l'approche qualitative insuffisante et le processus de prise de décision très compliqué. A titre d'exemple, la tâche de pronostic implique plusieurs oncologues utilisant différents bio-marqueurs et facteurs cliniques. Habituellement, dans de tels cas de nombreux types d'informations qualitatives sont intégrées pour arriver à une décision raisonnable sur le pronostic par les cliniciens participants. Ce n'est pas une tâche facile, même pour les cliniciens les plus qualifiés. Si l'on ajoute à cela le besoin accru d'explorer la grande quantité de données biologiques étant disponibles (mesures protéomiques et génomiques), des approches plus efficaces sont devenues indispensables pour aider les médecins dans leur décision. Récemment, les approches d'apprentissage automatique se sont montrées très efficaces pour aider à la prise de décision en fournissant une prédiction plus précise et des modèles de classification efficaces. La première utilisation de ce type d'approche dans le domaine du cancer date d'environ 25 ans par des méthodes populaires telles que les réseaux de neurones et les arbres de décision (Simes, 1985, Maclin et al., 1991). Avec l'introduction de la technologie à haut débit, le recours à des méthodes de calcul plus intensif est indispensable.

Dans ce chapitre nous décrivons l'état de l'art sur l'utilisation des méthodes d'apprentissage automatique dans le domaine du cancer en soulignant leurs avantages et inconvénients. Cette utilisation peut être résumée en trois tâches principales:

- *Classification de nouveaux patients en des classes de cancer prédéfinies en utilisant un modèle obtenu par apprentissage, connue sous le nom de classification supervisée.*
- *Regroupement des patients ayant des propriétés similaires en sous-groupes, connu sous le nom de classification non supervisée. Les approches utilisées pour effectuer cette tâche peuvent être divisées en deux catégories: hiérarchiques et en se basant sur la partition de l'espace.*

Puis quelques applications de ces approches dans la gestion du cancer du sein sont rapportées. Malgré leur utilisation réussie dans la gestion du cancer du sein en se basant sur les facteurs cliniques classiques, il a été remarqué que la plupart d'entre elles ne parviennent pas à faire face aux défis récents apportés par l'introduction des données issues de technologies avancées. Nous pouvons par exemple mentionner le problème de sur-apprentissage dans les méthodes de classification supervisée en raison souvent du faible ratio attribut/échantillon (nombre de patient). Cela nécessite un recours aux méthodes de sélection largement étudiées et développées pour surmonter ce problème. Nous avons examiné brièvement les travaux de recherche considérables effectués dans cette direction. Il a été constaté cependant que la sélection de variables n'est pas seulement utile pour la réduction de la dimensionnalité du problème, mais permet des progrès majeurs pour acquérir de nouvelles connaissances sur la biologie du cancer en utilisant les profils d'expression génétique. Grace à ces approches, des méthodes adaptées et personnalisées sont aujourd'hui en cours de développement en se basant sur l'extraction de plusieurs signatures génétiques afin d'améliorer la précision de gestion du cancer. Nous avons enfin décrit les approches d'apprentissage non supervisées et leurs applications dans la gestion du cancer de sein en particulier à travers leur utilisation dans l'identification de groupe de gènes co-exprimés.

Ce chapitre se termine par une description des défis récents auxquels il faut faire face pour améliorer la gestion et le traitement du cancer. Nous avons considéré principalement les problèmes d'hétérogénéité des données, la dimensionnalité élevée, le faible rapport signal-bruit et les incertitudes d'appartenance. L'hétérogénéité des données est liée à l'utilisation quotidienne de variables de type mixte dans la création des bases de données, une pratique courante dans de nombreux problèmes de cancer. Malgré le nombre important de travaux consacrés à résoudre le problème de la dimensionnalité élevée des données, il est toujours considéré comme un problème de recherche ouvert et l'un des principaux défis dans la théorie de l'apprentissage statistique. Alors que le problème du faible rapport signal sur bruit est lié au problème de la reproductibilité des technologies à haut débit (puces à ADN, spectrométrie de masse), dû principalement aux variations de conditions expérimentales et biologiques. Au mieux de notre connaissance, ce problème n'a jamais été abordé par la communauté d'apprentissage automatique. Nous avons aussi noté que les bruits ne sont pas les seules incertitudes dans les données du cancer, l'incertitude d'appartenance d'une tumeur à chacun des sous-types de cancer est une réalité évidente, et elle gagne une attention croissante dans les études récentes qui utilisent des données recueillies à partir de différentes technologies par les différents centres médicaux (Haibe-Kains et al., 2010).

CHAPTER 2
Machine Learning for Cancer Management and Treatment

For a long time cancer management and treatment were performed based on expert qualitative knowledge held by individuals or using diverse medical guidelines. However, cancer disease has been shown to be complex and very heterogeneous which make the qualitative approach insufficient and the decision-making process very complicated. Breast cancer diagnosis for instance is based on the analysis of thousands of mammograms issued by imaging detection tools. This important task seems to be very complex and tiring, and can even lead the radiologists to commit some diagnosis errors. Furthermore, the prognosis task involves usually multiple physicians with different skills using different biomarkers and clinical factors. Typically in such cases many types of qualitative information are integrated to come up with a reasonable decision about the prognosis by the attending physicians based on their own intuition. This is not an easy task even for the most skilled clinicians. If we add to that the increased need to explore the large amount of biological data being available (proteomic and genomic measurements), more efficient approaches to help physicians in their day-to-day practices have become indispensable. Recently, machine learning has been shown very effective to help physicians in their decision making by constructing more accurate prediction and classification models. Machine learning is a branch of artificial intelligence that employs a variety of statistical, probabilistic and optimization techniques that allows computers to "learn" from past examples and to detect hard-to-discern patterns from large, noisy, heterogeneous or complex datasets (Baldi and Brunak, 2001; Cruz and Wishart, 2006). Although machine learning was basically much related to statistics, it offers nowadays a powerful mean to deal with statistically ill-posed problems such as curse of dimensionality (small sample size characterized by a high feature dimensionality (Bellman, 1961)) and noisy measure (Mitchell, 1997; Duda *et al.*, 2001)). Since nearly 25 years artificial neural networks (ANN) and decision trees (DTs) have been widely used for cancer detection and diagnosis (Simes , 1985, Maclin *et al.*, 1991, Cicchetti, 1992). More recently, machine learning methods are being also used increasingly for cancer prognosis and treatment planning. Firstly, Machine learning approaches have been used mainly to perform cancer prognosis and diagnosis, as explained in the previous chapter, based on some clinical and histo-patholgical factors,

including histological grade, size of tumor and the age of the patient (Cochran, 1997; Gómez-Ruiz *et al.*, 2004). With the development of high throughput technologies (DNA microarray, sequencing), proteomic (protein chips, immune-histology), physicians have find themselves faced to thousands of genetic, cellular and clinical markers. In this situation, for which human intuition and traditional statistics fails, the resort to more intensively computational methods is unavoidable, such as machine learning approaches. This helps physicians to analyze and interpret data, and gain new insights into cancer biology. To this end, machine learning approaches have known recently a wide spread use in cancer research to scale with such complex experimental data for different purposes (diagnosis, prognosis and treatment planning) (Khan *et al.*, 2001; Guyon *et al.*, 2002).

Generally, machine learning methods are used to analyze medical datasets organized in table form containing a set of patient (individuals or patterns) in term of their properties (attributes, features, variables). The use of machine learning methods in cancer research can be summarized in three main tasks:

- Classifying new patients based on trained models to already-defined cancer classes, known as supervised classification within machine learning community
- Regrouping patients having similar properties into subgroups, known as unsupervised classification or clustering within machine learning community
- Selecting relevant biomarkers using feature selection approaches either in a supervised or unsupervised context.

However not every machine learning method is appropriate for any cancer research problem. For instance some machine learning methods scale very well to the size of data, others do not. Likewise some methods may have some data requirements and assumptions that render them inappropriate to the problem under investigation. This is not necessarily a weakness to machine learning, it is only to highlight the attention should be paid to choose a suitable method for a particular problem (Cruz and Wishart, 2006).

This chapter describes each task and gives briefly their associated challenging aspects in the bioinformatics context. We explain the medical questions of interest, the approaches usually used, and the state of bioinformatics research. This chapter ends with a description of the main challenges that have to be faced to improve cancer management and treatment.

2.1 Supervised classification

Classification is considered as one of the fundamental problems in machine learning. Duda and Hart (2001) define it as the problem of assigning an element or instance to one of several pre-specified categories. Only available information is a set of patterns characterized by a set of features each of them assigned to a predefined class. Each pattern is classified based on a set of classification rules which are often unknown in many real-life situations (Baldi and Brunak, 2001). As a simple example, we can cite the problem of breast cancer diagnosis as a supervised classification problem (Wolberg *et al.*, 1994). The elements to be classified form a set of patients as shown in Table 2.1.

Tab. 2.1 Cancer diagnosis dataset used for supervised classification.

ID number	Clump thickness	Uniformity of cell size	…	Mitoses	Class
842302	17.99	10.38	…	0.11890	Malignant
842517	20.57	17.77	…	0.08902	Malignant
…	…	…	…	…	…
…	…	…	…	…	…
926954	16.6	28.08	…	0.78200	Malignant
927241	20.6	29.33	…	0.12400	Malignant
92751	7.76	24.54	…	0.07039	Benign

The attributes (features) of a given patient are some variables including around thirty features computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. The outcome of each patient is taken as either diagnosed to have a breast cancer or not, representing its predefined class. This simple example has been for a long time used to assess the performance of newly proposed machine learning approaches. Compared to other fields, oncology is possibly the area in which more applications of machine learning have been performed (Vellido and Lisboa, 2007). Almost all machine learning approaches applied on this problem employ supervised learning such as artificial neural networks (Rumelhart *et al.*, 1986), decision trees (Quinlan, 1986), discriminant analysis (Fisher, 1936), *k*- nearest neighbor (Cover and Hart, 1967) and Support Vector Machines (Vapnik, 1998). We list below some of the most used supervised machine learning approaches in cancer research.

2.1.1 Artificial neural networks

Artificial neural networks (ANN) were originally inspired from the human-being brain which works with interconnected neurons (Figure 2.1). The strength of neural connection is

determined through a learning process on labeled data characterized by weights (Cruz and Wishart, 2006). In an ANNs, the neurons are organized in layers, in such a way that usually only neurons belonging to two consecutive layers are connected.

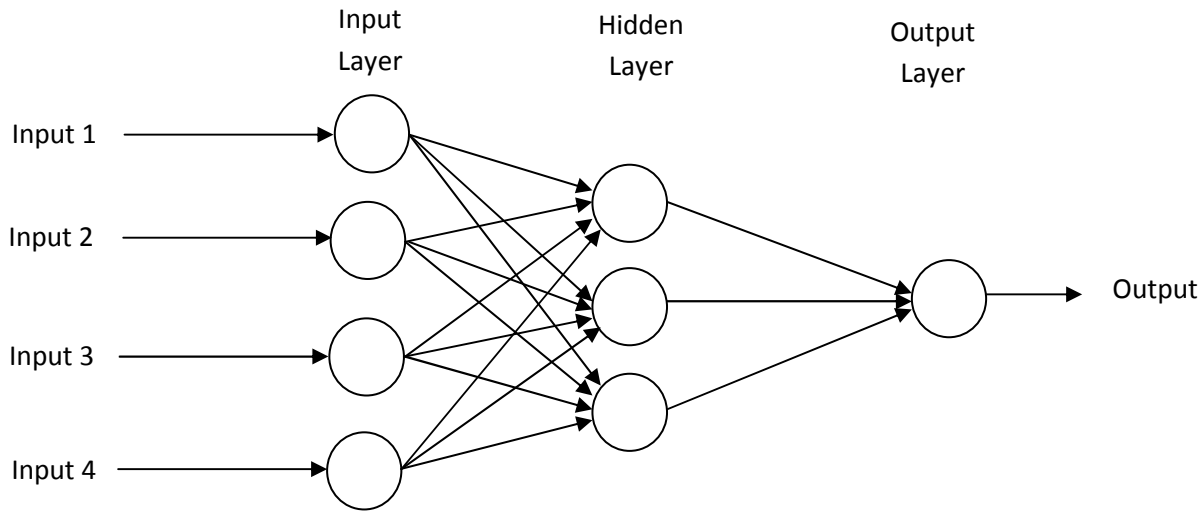


Fig. 2.1. Artificial Neural Network

During the classification process, ANNs enable to perform statistical operations (linear, logistic, and non linear regression) and logical operations or inferences (AND, XOR, NOT, IF-THEN) (Mitchell, 1997; Rodvold *et al.*, 2001). The perceptron (Rosenblatt, 1962) is the simplest neural network that, using a threshold activation function, enables to separate two classes by a linear discrimination function. Adjustment of connection strength is usually based on an optimization approach called *backpropagation* algorithm (Rumelhart, 1986). One of the first applications of machine learning approaches in cancer research were through neural networks (Maclin *et al.*, 1991; Cicchetti, 1992). Recently, their use has also been extended to other cancer applications such as cancer prognosis and treatment planning (Gómez-Ruiz *et al.*, 2004; Jerez *et al.*, 2004; Ripley *et al.*, 2004; Mian *et al.*, 2005). Some limitation of the ANNs is the lack of interpretability and the problem of overfitting especially when a high dimensional data is faced (e.g. microarray data) (Cruz and Wishart, 2006).

2.1.2 Decision trees

A decision tree is a structured graph or flow chart of decisions (nodes) and their possible consequences (leaves or branches) used to create a plan to reach a goal (Quinlan, 1986) (Figure 2.2). In a classification tree, pattern classification starts from the root node by successively asking questions about each of its properties (features). Different exclusive links from a root node correspond to the different possible values of the property (feature).

According to the answer, this process is followed until arriving to a leaf node which has no further question. The pattern is finally assigned to the class represented by this node.

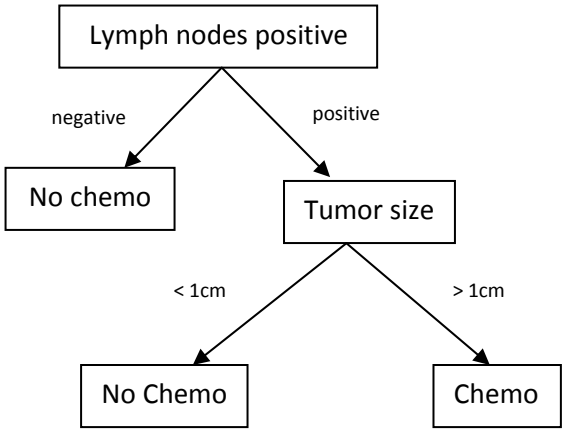


Fig. 2.2. Decision tree

A variety of approaches can be found for choosing the appropriate order of features in the decision tree and how possibly make reduce the large trees. Decision trees are very well accepted in medical applications owing to its high model transparency and comprehensive interpretability. This is argued by the fact that decision trees are a sort of rule-based methods which provide a comprehensive interpretation. Indeed, the factor of interpretability should not be underestimated in the real medical practice where "most physicians are not even accustomed to the idea of computer-aided problem solving" (Lucas, 1997). Decision trees are also one of the first methods applied in breast cancer research such as predicting breast cancer survivability (Delen *et al.*, 2005), diagnosis (Lee *et al.*, 2010a) and treatment planning (Khan *et al.*, 2008). Some potential limitations affecting the application of decision trees in cancer research is its difficulty to scale with high dimensional data (e.g. microarray data) and the strong assumption on mutual exclusivity of classes (Cruz *et al.*, 2006).

2.1.3 Discriminant analysis

Fisher linear discriminant analysis (Fisher, 1936) constructs a linear hyperplan based on the maximization of between-group to within-group ratio. Assuming a multivariate normal distribution and homogeneity of covariance matrices, the hyperplan is described by a linear discriminant function which equals zero at the hyperplan. In this case, the hyperplan is defined by geometric means between the centroids (i.e. the center of each classe) (Baldi and Brunak, 2001). Recently, a variety of non linear discriminant analysis approaches were proposed based on kernel concept to improve its classification performance (Mika *et al.*, 1999). This approach has found its place in some breast cancer applications (Miller *et al.*,

2005; Michiels *et al.*, 2005; Reid *et al.*, 2005; Sun *et al.*, 2007a). However, this approach suffers from several limitations such as the *small sample size problem* due to within-class matrix singularity (Fukunaga, 1990). This problem arises whenever the number of samples is smaller than the dimensionality of samples (the case of cancer classification with gene expression profiling characterized by thousands of genes and less than one hundred patients).

2.1.4 *k*- nearest neighbor

The *k*- nearest neighbor method classifies each unlabelled sample by the majority label among its *k* nearest neighbors in the training set (Cover and Hart, 1967). This makes it very well suited for non-linear classification problems. One potential of this approach is that it does not make any assumption on data distribution. A variety of breast cancer studies can be found in literature based on this approach (Parry *et al.*, 2002; Olshen and Jain, 2002; Zheng *et al.*, 2010). Though simple, however, it is known that *k*-NN classifier is very sensitive to the presence of irrelevant features. Moreover, this method tends to be slow for large training dataset because the nearest neighbors should be searched over all instances (Baldi and Brunak, 2001).

2.1.5 Support vector machines

The key idea of this approach is that by an appropriate mapping into sufficiently high dimensional space, it is always possible to define a hyperplan that separates the data from two categories (Vapnik, 1998) (Figure 2.3).

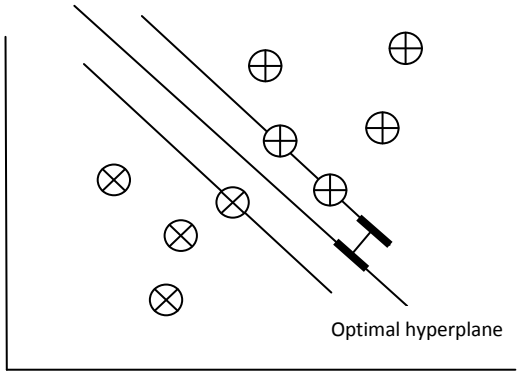


Fig. 2.3. Support Vector Machines

The mapping is performed using some specific functions (known as kernel functions) which are chosen by the user among a variety of functions (Gaussian, polynomial, linear,…) according to the problem under investigation. The goal in all cases is to find the separating hyperplan in the resulted space with the largest margin, expecting that the larger is the margin,

the better is the generalization of classifier (Vapnik, 1998). This problem is generally reformulated as a constrained optimization problem and solved generally by resorting to its dual reformulation. Various applications using SVM has been performed on breast cancer research (Liu *et al.*, 2003; Chang *et al.*, 2003b; Land and Verheggen., 2009). SVM approach is known to be very robust to noisy features and the overffiting problem is unlikely to occur. This has encouraged recently its use in many breast cancer studies using microarray data (Guyon *et al.*, 2002; Buness *et al.*, 2009; Lee, 2010b). Although its demonstrated efficiency in wide range of classification problems, SVM presents major limitations such as the problem of selecting a suitable kernel function, its parameters and penalties (Baldi and Brunak, 2001).

2.2 Unsupervised classification (clustering)

Clustering is considered as one of the fundamental research problems in various data analysis fields such as machine learning and pattern recognition (Jain and Dubes, 1988; Jain *et al.*, 1999; Xu and Wunch, 2005). Cluster analysis seeks to organize a set of patterns (e.g. patients or genes) into clusters such that patterns within a given cluster have a high degree of similarity, whereas patterns belonging to different clusters have a high degree of dissimilarity (Duda *et al.*, 2001). Unlike supervised classification, the outcome of each element in the unsupervised context is unknown making the learning task more challenging.

One typical example in cancer research is the clustering of genes expression data (Belle *et al.*, 2010). In microarray experiment, the expression value of thousands of genes is obtained for only few patients. Extracting co-expressed genes in different samples from this data is of great importance as it may allow gaining new insights into cancer biology. This is typically a clustering problem where co-expressed genes should be grouped into the same cluster (Baldi and Brunak, 2001).

Many algorithms have been proposed to address this problem for different purposes (Jain and Dubes, 1988; Jain *et al.*, 1999; Baraldi *et al.*, 1999; Xu and Wunch, 2005). Clustering techniques can be roughly divided into two main categories: Hierarchical and partitioning.

2.2.1 Hierarchical clustering

Hierarchical clustering produces a nested series of partitions on the form of tree diagram or dendogram (Jain and Dubes, 1988; Jain *et al.*, 1999). In hierarchical clustering we can distinguish two situations between two groups from different partitions: either they are disjoint or one group wholly contains the other (Figure 2.4). Two clusters are merged in

hierarchical measure based on a distance or dissimilarity measure such as Minkowski and Mahalanobis measures (Jain and Dubes, 1988; Jain *et al.*, 1999). It exist several algorithms to establish a hierarchical tree: agglomerative and divisive. Hierarchical clustering is the most commonly used method to summarize data structures in bioinformatics generally and in breast cancer specifically (Baldi and Brunak, 2001). Many studies can be found in cancer research literature about the use of this clustering approach, especially for microarray data analysis. In (Sotiriou *et al.*, 2003), the use of hierarchical cluster analysis has led to distinguish between two groups of patients based on their ER status. This approach has also been used in the famous Stanford study to identify subgroups of cancers with separate gene expression profiles (Perou *et al.*, 2000). Alizadeh *et al.* (2000) were able to identify formerly unknown types of B-cell lymphoma with distinct clinical behaviour by using hierachical clustering of expression data. The use of this approach, however, was not only limited to cancer class discovery, prognosis and treatment responsiveness prediction were respectively targeted in (Belle *et al.*, 2010) and (Rouzier *et al.*, 2005).

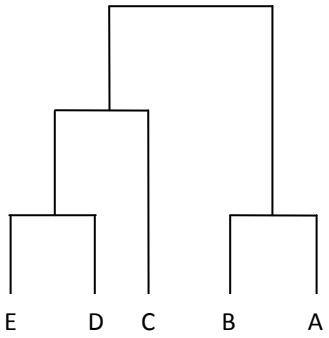


Fig. 2.4: Hierarchical clustering

2.2.2 Partitioning clustering

Partitioning clustering identifies only one partition of the data that optimizes an appropriate objective function (kernel, spectral, fuzzy and classical) (Jain and Dubes, 1988; Jain *et al.*, 1999; Xu and Wunch, 2005) (Figure 2.5).

The clustering can be either hard (each pattern belongs to only one class) or fuzzy (where each pattern belongs with a certain degree of membership to each resulting cluster) (Jain *et al.*, 1999). Fuzzy clustering offers the advantage to provide a basis for constructing rule-based fuzzy model that has simple representation and good performance for non-linear problems (Yao *et al.*, 2000).

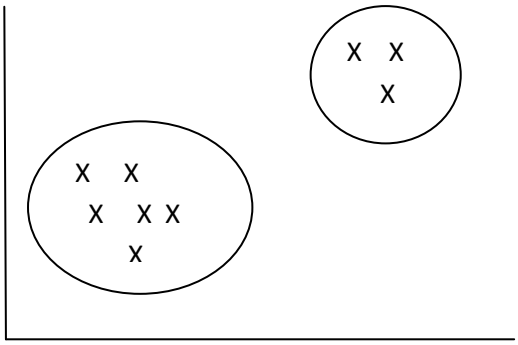


Fig. 2.5 Partitioning clustering

The *k*-means algorithm (MacQueen, 1967) is one of the most popular partitioning clustering algorithms. This algorithm is based on a "hard" partition of the data into *k* clusters based on the minimization of the within-group sum of squares. A direct extension of the *k*-means algorithm is the Fuzzy C-means (FCM) (Bezdec, 1981), where the fuzzy set notion is introduced into the class definition. In this case, each element belongs to a given class with certain membership degree. Likewise, FCM minimizes the within-group sum of squares but by taking into account the membership degrees of each element. Another interesting clustering approach is the Self-Organizing feature Maps (SOM) (Kohonen, 1982). In this approach the data are represented by means of codevectors on a grid with fixed topology. Codevectors are adaptive according to input distribution, but adaptation is propagated along the grid to neighborhood codevectors, according to a specific neighborhood function (Filippone *et al.*, 2008).

These clustering approaches are widely used in breast cancer research. For instance, a molecular classification of tumor samples can be achieved using either unsupervised methods like *k*-means clustering (Bertucci *et al.*, 2002; Wang *et al.*, 2003; Wiseman *et al.*, 2005) or 'SOMs' (self organizing maps) (Covell *et al.*, 2003). Tamayo *et al.* (1999) have also used SOMs on DNA array data to differentiate subtypes of acute leukaemia. Clustering approaches have been also used to cluster the gene in groups and establish the relation between the co-expressed genes in each group (De Souto *et al.*, 2008). Many studies can be found also where the clustering is performed in both directions, i.e. patients and genes, called biclustering (Cheng and Church, 2000; Sheng *et al.*, 2003). However, the use of different methods may yield different results. Therefore, those approaches should be used with caution according to the problem under consideration.

2.3 Feature selection

Usually, for many learning domains potential useful attributes, also called features, for pattern description are defined randomly. However, not all of these features are important for learning task (i.e. supervised or unsupervised learning): some of them can be irrelevant, some may be redundant, and some can even misguide learning results. The problem of selecting important features is known in the literature as feature selection. Feature selection is defined as the problem of choosing a small subset of features that ideally is necessary and sufficient to describe the target concept (Kira and Rendell, 1992a). Cancer research is the field in which feature selection is being extensively employed. With the involvement of high throughput technology in breast cancer management, feature selection has become a necessary step in order to discard the huge number of irrelevant genes. The most important objectives of feature selection are: (a) to avoid overfitting and improve model accuracy, i.e. classification performance in the supervised learning and better clusters detection in the case of clustering, (b) reducing training time of the model, (c) to gain deeper insight into the underlying processes that generated the data (Saeys *et al.*, 2007). A typical feature selection task consists of four basic steps: subset generation, subset evaluation, stopping criterion and result validation (Liu and Yu, 2005). Subset generation produces candidate feature subsets for evaluation based on a certain search strategy. According to the evaluation criterion, this new subset can be either retained to replace the previous best subset or rejected. This process is repeated until a given stopping criterion is satisfied. Then the winner feature subset is validated finally via a real world dataset (see (Liu and Yu, 2005) and reference therein for review). Many research efforts have been directed in the last two decades towards developing efficient feature selection methods in a supervised framework (Kira and Rendell, 1992a; Weston *et al.*, 2001; Gilad-Bachrach *et al.*, 2004). However, only few works have been devoted to address this problem in the unsupervised learning and clustering. This is mainly due to the absence of class labels, unlike in supervised learning, to assess the importance of a subset of features. Most of unsupervised feature selection algorithms are based on information or consistency measures (Mitra *et al.*, 2002; Dy and Brodley, 2004; Wei and Billings, 2007).

In the context of classification, existing feature selection methods are traditionally categorized as filter, wrapper, hybrid or embedded methods, with respect to the criterion used to search for relevant features (Kohavi and John, 1997; Guyon and Elisseeff, 2003). We describe below the three approaches and review some of their advantages and drawbacks.

2.3.1 Filter methods

In filter methods an independent evaluation function based generally on a measure of information content is used to select a set of features that maximizes this function, regardless of their effects on model performance. Then different classification methods can be applied using only this subset of features. Filter approaches are computationally very efficient and can scale well with high dimensional data. Thanks to its computational properties, many cancer research studies have resorted to use filter approaches especially for microarray data analysis. In these approaches all the genes are evaluated individually, e.g. through t-test and Fisher score (Dudoit *et al.*, 2002; Li *et al.*, 2004). However, filter approaches presents some limitations related to the problem of interactions between features. Furthermore, they do not often guarantee a maximum classification performance because they totally ignores the effects of the selected subset of features and thereby sometimes perform very poor.

2.3.2 Wrapper methods

Wrapper methods use the performance of a learning method to assess the relative usefulness of the selected feature subset (e.g. by cross validation) (Kohavi and John, 1997; Guyon and Elisseeff, 2003). In other words, wrapper method requires one learning method (e.g. decision trees, SVM, k-NN,…) and uses its performance as the evaluation criterion. For feature subset search step, an exhaustive procedure can be performed, if the number of features is not too large. But, with ten thousands of features, the search becomes quickly intractable to perform the combinatorial searching required in wrapper methods. A wide range of search strategies can be used, including best-first, branch-and-bound, simulated annealing, genetic algorithms (see (Kohavi and John, 1997) for review). With the aim to improve the classification accuracy in cancer applications, wrapper methods have known also rapidly a wide spread use (Blanco *et al.*, 2004; Wang *et al.*, 2005b). In (Sun *et al.*, 2007a) a more advanced approach that avoids the computational issue by optimizing a margin-based objective function has been used for gene selection. A relevant comparative study between filter and wrapper methods for gene selection has been performed in (Inza *et al.*, 2004).

2.3.3 Hybrid methods

In hybrid feature selection approaches a filter feature selection method is firstly used to reduce the initial feature dimension and then a wrapper approach is applied on the reduced subset of features (Das, 2001). Nevertheless, the search in this approach is time consuming

and depending on the learning approach used by the wrapper method. Some attempts can be found in cancer literature using such approaches for microarray data analysis (Xing *et al.*, 2001).

2.3.4 Embedded methods

In these approaches the feature selection task is incorporated into the learning process. Just like in wrapper approaches, embedded approaches require therefore a learning algorithm. Embedded approaches have the advantage that they integrate the interaction with the learning method, while at the same time being less computationally intensive than wrapper methods. Embedded methods are not new in machine learning as some of the oldest decision trees such as CART (Breiman *et al.*, 1984) encompass a built-in mechanism to perform feature selection. Weston and his co-authors have proposed an embedded feature selection approach for SVM methods (Weston *et al.*, 2001). Recursive Feature Elimination RFE (Guyon *et al.*, 2002) is a well-known feature selection method designed specifically for microarray data analysis. It works by iteratively training an SVM classifier with a current set of features, and then heuristically removing the features with small feature weights.

2.4 Recent challenges in breast cancer management

In spite of the intensive research performed in the machine learning filed (see previous sections) in past decades, many challenges are still needed to be addressed seriously to improve cancer management. Challenges are mainly related to data characteristics used in decision-making process. Three challenges are mainly faced: the first one is related to the presence of mixed-type data in daily produced clinical datasets, the second one is related to high dimensionality in data especially issued from microarray technology and the last one is the problem of noise and uncertainties associated usually to both data. Addressing efficiently those problems is urgently needed provided that in some cancer applications the three challenges can be even faced simultaneously (e.g. integration of clinical and microarray data to improve breast cancer management (Sun *et al.*, 2007a, Gevaert *et al.*, 2006). We describe thereafter in detail the three challenges which will make the focus of the present thesis.

2.4.1 Data heterogeneity

Features used by physicians for patient state description are generally represented in different ways. The most used representation is the pure quantitative one which assumes a complete accuracy about the information. Taken as it appears, a real number contains an infinite

amount of precision whereas human knowledge is finite and discrete. So, there is a need to use data represented by symbolic values to fit with human perception. The representation of data can be therefore done in different ways: quantitative (e.g. Age=50), symbolic intervals (e.g. age belongs to the interval [40,60]) or qualitative values (e.g. old, young, menopause,…). Thus, the development of an automatic mechanism for medical support is faced with this problem of data heterogeneity (quantitative, qualitative and interval data). Indeed, daily produced medical datasets are commonly characterized by a subset of heterogeneous (mixed type) features. For instance, many datasets from the popular UCI machine learning repository (Blake and Merz, 1998) are described by heterogeneous features. During the last decades, few research works have been directed to defy the issue of representation multiplicity for data analysis purposes (Michalski and Stepp, 1980; Mohri and Hidehiko, 1994; Hu *et al.*, 2007). However, to the best of our knowledge, no standard principle has been proposed in the literature to handle in a unified way heterogeneous data. Indeed, a lot of proposed techniques process separately quantitative and qualitative data. In feature selection tasks for example, they are either based on distance measures for the former type (Kira and Rendell, 1992a) and on information or consistency measures for the later one (Dash and Liu, 2003). Whereas in classification and clustering tasks, eventually only a Hamming distance is used to handle qualitative data (Aha, 1989; Aha, 1992; Kononenko, 1994). Other approaches are originally designed to process only quantitative data and therefore arbitrary transformations of qualitative data into a quantitative space are performed without taking into account their nature in the original space (Cover and Hart, 1967; Kira and Rendell, 1992a; Weston *et al.*, 2001). Another inverse practice is to enhance the qualitative aspect and discretize the quantitative value domain into several intervals, then objects in the same interval are labeled by the same qualitative value (Liu *et al.*, 2002; Hall, 2000). Obviously, both approaches introduce distortion and end up with information loss with respect to the original data. Moreover, none of the previously proposed approaches combines in a fully adequate way, the processing of symbolic intervals simultaneously with quantitative and qualitative data. An interesting approach would be to unify the different heterogeneous spaces into one homogeneous space and then reason in a unified way about the whole data to make the appropriate decision. To avoid any type of distortion and/or information loss the space's unification process should be performed appropriately for each type of data.

2.4.2 High feature-to-sample ratio (curse of dimensionality)

The recent introduction of high throughput technology in breast cancer management has brought with it a new challenge related to the high dimensionality of microarray data. Indeed, this problem, known as curse of dimensionality (or high feature-to-sample ratio), is still considered as one of the principal challenges in statistical machine learning (Lafferty and Wasserman, 2006). As it has been pointed out in section 2.2, due to the presence of large amount of irrelevant genes, many traditional classification approaches either present some limitations (e.g. overfitting) or important computational time (e.g. *k*-NN). Even when the use of feature selection approaches can help to alleviate this problem, most of them become unpractical when the problem of dimensionality is associated with the problem of heterogeneity (section 2.4.1) or the noisy nature of microarray measurement (detailed in next section). Therefore, there is a need to develop new approaches enabling to deal efficiently and simultaneously with such problems.

2.4.3 Noise and uncertainty

From other side, the features used to describe a patient state can also be corrupted by several types of noise and uncertainties due to measurement, human approximations or biological interaction. For instance, it has been reported recently that the major difficulty in deciphering high throughput gene expression experiments comes from the noisy nature of the data (Tu *et al.*, 2002). Indeed, data issued from high throughput technology are not only characterized by dimensionality problem but present also another challenging aspect related to thier low signal-to-noise ratio. The noise in such type of data is multisource: Biological and noisy measurement, slide manufacturing errors, hybridization errors, scanning errors of hybridized slide (Tu *et al.*, 2002; Nykter *et al.*, 2006). Biological errors are typically due to internal stochastic noise of the cells and error sources related to sample preparation (Blake *et al.*, 2003). This type of intrinsic noise is present in all measurements, regardless of the measurement technology. Measurement errors, on the other hand, include error sources that are directly related to the measurement technology and its limitation (e.g. bias due to the used dyes) (Nykter *et al.*, 2006). The properties of this kind of extrinsic noise depend on the measurement technology (Blake *et al.*, 2003). Slide manufacturing errors are related to microarray slide images. These include variation in the spot position and size. In addition the marks done by a print tip and deformations in the spot shape can be produced (Nykter *et al.*, 2006). Hybridization errors include background noise, spot bleeding, scratches, and air

bubbles (Nykter *et al.*, 2006). Another possible source of error is the digitization of hybridized slide by scanning. The hybridized slide is read by scanning each dye color separately, it might be possible that channels do not align perfectly (Nykter *et al.*, 2006). Many studies were performed to study the different effects of experimental, physiological, and sampling variability (Lee *et al.*, 2000; Novak *et al.*, 2002). An interesting study has been performed in (Tu *et al.*, 2002) to analyze the quantitative noise in gene expression microarray experiments. The authors have shown through two illustrative concrete examples the difference in gene expression due to experimental noises. In the first example, a comparison between gene expression values measured on the same sample has been performed. Figure 2.6a shows the overall difference in two measured gene expression due to measurement error alone as provided in (Tu *et al.*, 2002). The deviation of the scattered points from the diagonal line represents the difference between the two measured transcriptomes. In the second example two samples from different cultures are compared as shown in figure 2.6 (b) so that the measured expression value differences contain the combined effect of the genuine gene expression differences caused by measurement error.

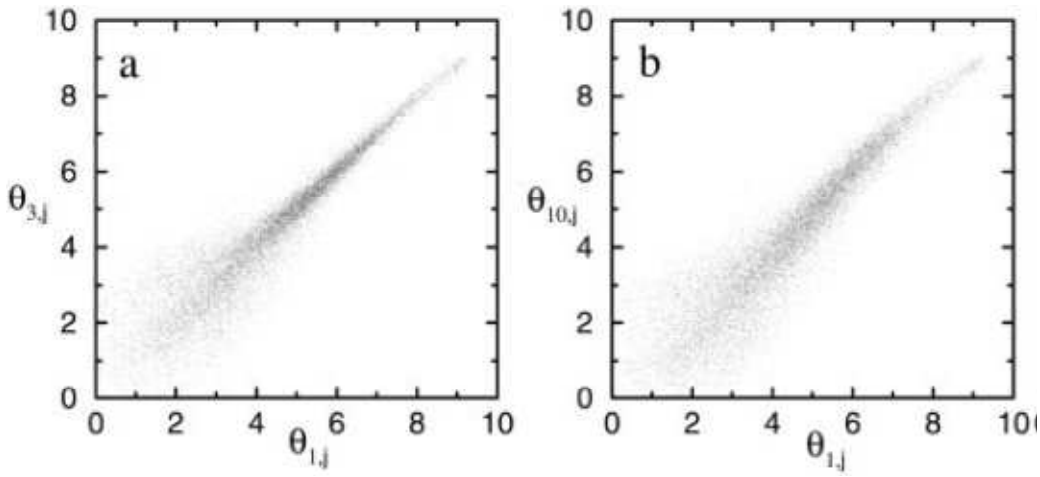


Fig. 2.6: The scatter plot of gene expression pairs (a) experiments pair on the same sample (b) experiment pair between two different samples. Figure taken from (Tu *et al.*, 2002).

Although Figures (a) and (b) appear similar, the deviations in the expression values from the diagonal line are completely different. The first one is due only to gene expression measurement error whereas the second is due to the combined effect of the gene expression differentiation and measurement error. Therefore, it is crucial to characterize the difference caused purely by experimental measurement from the expression differentiation due to the difference between the two cultures.

All existing feature and classification approaches assume that microarray data is perfect without wondering about its reliability. One common practice to deal with this problem is to transform in non-linear way the gene-expression levels in a preprocessing phase so that the variance across experiments becomes comparable for each gene (Huber *et al.*, 2002). A drawback with this approach is that a global transformation does not adequately account for the fact that the same gene may be measured with different precision in different experiments (Huber *et al.*, 2002). Another drawback with this approach is that a complex non-linear transformation of the data complicates the interpretation of measurement when compared with a global transformation. Machine learning approaches can offer also a powerful tool to tackle such problem. An interesting approach would be to use symbolic data analysis (SDA) popularized by Bock and Diday (Bock and Diday, 2000). Within this framework, interval data representation can be used to take into account the usually uncertainty and noise inherent to measurements (Billard, 2008). Symbolic interval features are extensions of pure real data types, in the way that each feature may take an interval of values instead of a single value (Gowda and Diday, 1992). In this framework, the value of a quantity x (e.g. gene expression value) is expressed as a closed interval $[x^-,x^+]$ whenever x is noised or uncertain; representing the information that $x^- \le x \le x^+$. However, the introduction of interval representation makes the data processing task more complex than when only a numerical value is considered, especially when high dimensionality problem is faced jointly. It is worthwhile to note that interval data presentation can be useful also for many other real world problems in cancer field.

Measurement uncertainty is not the only type of uncertainty to be faced in real-world problems generally and medical field specifically. Another uncertainty type of big interest is the membership uncertainty of patients to each class, i.e. a patient's tumor can belongs simultaneously to many cancer groups with some degree of membership, in a way that the decision making mechanisms become reproducible and robust, because clinically relevant cancer groups are identified in several public datasets using different populations of breast cancer patients. Indeed, breast cancer has been shown to be a highly heterogeneous disease requiring the consideration of such uncertainty in decision making process. Even in the day-to-day practice, physicians in their decision process incorporate naturally such uncertainty for any disease management. Fuzzy set theory, introduced by Lotfi Zadeh (Zadeh, 1965), represents an appropriate framework to deal with membership uncertainties. Medicine was one of the first fields in which Zadeh's fuzzy set theory was applied, to deal with vagueness in

perceptions of reality phenomena (Zadeh, 1969). Although fuzzy approaches have started to gain increasing attention in wide range of cancer applications (Ressom *et al.*, 2003; Andrews *et al.*, 2003; Haibe-Kains *et al.*, 2010), its scalability with recent challenges is still far to be convincing compared to other classical machine learning approaches. Therefore efficient fuzzy approaches to deal with such problems may make a major contribution in the improvement of cancer management.

2.5 Conclusion

In this chapter we have reviewed the state-of-the-art of machine learning in cancer research. We have described the main three machine learning tasks most used in cancer management: supervised classification, clustering and feature selection. A few examples of the most famous approaches for each task have been briefly described by highlighting their advantages and drawbacks. Then some applications of such approaches in breast cancer management have been provided. Although their successful use in breast cancer management based on traditional clinical factors, we have noticed that most of them fail to deal with the recent challenges brought by the introduction of data issued from advanced technologies. We can for instance mention the problem of overfitting in supervised classification methods due usually to the low feature-to-sample ratio. This requires a resort to feature selection approaches extensively studied and developed to overcome this problem. We reviewed briefly the tremendous research work have been made in that direction. We have noticed however that feature selection is not only useful for dimension reduction but has made major advancements to gain new insights in cancer biology by using gene expression profiles. Thanks to feature selection approaches a tailored and personalized cancer management is today underway by the derivation of several genetic signatures for different purposes. We have finally described the unsupervised learning approaches and their applications in breast cancer management especially through their use in the identification of group of coexpressed genes.

This chapter ends with a description of the recent challenges that have to be faced to improve cancer management and treatment. We considered mainly the problems of data heterogeneity, high dimensionality, low signal-to-noise ratio and membership uncertainties. Data heterogeneity is related to the use of mixed-type features in daily produced datasets, a common practice in many cancer problems. Although the important number of works devoted to address the problem of high data dimensionality, it is still considered as an open research problem and one of the principal challenges in statistical learning theory. Whereas the

problem of low signal-to-noise ratio is related to the problem of reproducibility in high-throughput technologies (microarray, mass-spectrometry), due mainly to the variations in experimental and biological conditions. To the best of our knowledge, this problem has never been addressed by the machine learning community. We noted also that the noises are not the only uncertainties in cancer; membership uncertainty of a tumor to cancer subtypes is an evident reality and is gaining increasing attention in recent studies using gathered datasets issued from different technologies by different medical centers.

In next chapter we address the problem of high dimensionality through the development of an embedded feature selection for SVM based on descent gradient method.

CHAPITRE 3- Résumé
Sélection de variables intégrée dans les machines à vecteurs supports par une méthode de gradient

Les technologies à haut débit fournissent régulièrement des bases de données caractérisées par un nombre sans précédent de variables pour représenter chacun des individus. Grace à sa capacité de fournir des solutions creuses, la régularisation d'apprentissage de type ℓ_1 a été montrée comme étant une méthode prometteuse pour la sélection de variables dans les problèmes de classification. Parmi le large éventail d'applications de régularisation de type ℓ_1, nous pouvons distinguer une régularisation ℓ_1 pour la régression logistique (Ng, 2004), LASSO (Tibshirani, 1996) et ℓ_1-SVM (Bradely et Mangasarian, 1998; Zhu et al, 2003). Nous nous concentrons dans ce chapitre sur le problème de régularisation ℓ_1-SVM afin de développer une approche de sélection de variables dite de type intégrée (ou «embedded» en anglais) pour surmonter le problème de dimensionnalité élevée.

En dépit de ses propriétés intéressantes, la mise en œuvre rapide des algorithmes ℓ_1-SVM pour des données de grande dimension a été considérée pendant longtemps comme un problème difficile, car la fonction objective ainsi obtenue est non-différentiable. Les méthodes génériques utilisées pour résoudre des problèmes convexes non-différentiables tels que les méthodes basées sur le gradient sont typiquement très lents. Diverses techniques d'optimisation avancées ont été exploitées pour développer des dizaines d'algorithmes capables de traiter des problèmes de moyenne et de grande échelle. Durant les années passées très peu de travaux ont été consacrés pour résoudre ce problème. En particulier, on peut distinguer le travail de Zhu et ses co-auteurs (Zhu et al., 2003) et plus récemment les travaux de Fung and Mangasarian (Fung and Mangasarian, 2004; Mangasarian, 2006). Dans le premier travail, le problème ℓ_1-SVM est formulé comme un problème de programmation dynamique afin d'utiliser les logiciels classiques pour le résoudre alors que dans le deuxième travail une méthode de Newton a été utilisée pour résoudre le problème dual comme un problème à pénalité extérieure. La dernière méthode est caractérisée cependant par une grande complexité en raison du nombre de paramètres à ajuster (cinq dont le paramètre de régularisation), ce qui la rend inutilisable par des utilisateurs non spécialistes. De plus on montre dans ce chapitre que la méthode de Newton ne garantit pas

toujours une solution optimale globale. Il a été souligné cependant par Chapelle (Chapelle, 2007) que le problème d'optimisation initial peut être résolu aussi efficacement sans passer par une formulation duale étant donné que dans les deux cas le même résultat est obtenu.
On propose ici d'utiliser une approche générique basée sur une technique de descente du gradient, notée ici DGM (Direct Gradient Method), pour résoudre le problème initial ℓ_1-SVM. Cette méthode s'est montrée efficace pour résoudre les problèmes de régression logistique normés ℓ_1 (Cai et al., 2010a). Cependant, elle suppose que la fonction objective soit différentiable, ce qui n'est pas le cas dans le cas du problème ℓ_1-SVM. Pour surmonter ce problème, la fonction de perte est remplacée par une fonction approximée dites de Hubber. Ensuite le problème d'optimisation convexe initial est transformé en un problème non-convexe sans contrainte, avec lequel, en utilisant une méthode de descente du gradient, une solution optimale globale est garantie. Cette méthode a été implémentée sur Matlab et comparée avec la méthode proposée par (Fung and Mangasarian, 2004; Mangasarian, 2006) dite LPNewton sur huit bases de données de grande dimension pour démontrer son efficacité. Il a été montré que cette méthode surpasse la méthode LPNewton en termes de temps d'exécution (CPU time) et en termes de précision atteinte en variant le paramètre de régularisation. A titre d'exemple, sur une base de données de cancer de la prostate (Stephenson et al., 2005) contenant 97 patients caractérisés par l'expression de 22291 gènes, la méthode proposée atteint le coût ciblé dans un temps d'exécution de 40.3 secondes alors que la méthode LPNewton demande 2147 secondes ; elle est donc 50 fois plus rapide. Il a été de plus constaté que la méthode LPNewton échoue à converger pour certaines valeurs de paramètre de régularisation dans la plupart des cas.

Il est à noté que ce travail a été réalisé dans le cadre d'un séjour de recherche au Laboratoire ICBR (Interdisciplinary Center for Biotechnology Research) à l'Université de Floride cofinancé par l'Ecole Doctorale EDSYS, l'Université Paul Sabatier et le groupe de recherche DISCO (Diagnostic et Conduite des Systèmes) du LAAS.

La grande dimensionnalité des données n'est cependant pas le seul problème rencontré dans les applications pratiques du cancer. Des problèmes tels que l'hétérogénéité des données, les incertitudes et les bruits peuvent également être rencontrés conjointement avec le problème de dimensionnalité élevée. Par conséquent, des méthodes plus efficaces sont nécessaires pour faire face simultanément à tous ces problèmes. Cette problématique représentera notre sujet d'intérêt dans les chapitres suivants afin de développer des approches appropriées capables de gérer de tels problèmes simultanément.

CHAPTER 3
Embedded Feature Selection for SVM by Gradient Descent Methods

High-throughput technologies produce routinely large datasets characterized by unprecedented number of features representing each data sample. ℓ_1 regularized learning, due to its ability to produce sparse solutions, has been shown to be a promising method for feature selection in classification problems. Among the wide range of ℓ_1 regularization applications, we can distinguish ℓ_1 regularized logistic regression (Ng, 2004), LASSO (Tibshirani, 1996) and ℓ_1-SVM (Bradely and Mangasarian, 1998; Zhu et al., 2003). We focus in the present work on the problem of ℓ_1 regularized SVM in the primal domain.

Despite its attractive properties, the fast implementation of ℓ_1-SVM algorithms for high-dimensional data has long been considered as difficult computational problem since the so-obtained objective function is non-differentiable. Generic methods for non-differential convex problems such sub-gradient methods are typically very slow. Various advanced optimization techniques were exploited to develop dozens of algorithms capable of handling medium and large scale problems. In the last few years only few works have been devoted to solve this problem. In (Zhu *et al.*, 2003) the ℓ_1-SVM problem is formulated as a linear programming problem and a standard software packages was used to solve it. Whereas in (Fung and Mangasarian, 2004; Mangasarian, 2006) a Newton method was used to solve the dual linear program formulation as an exterior penalty problem. The basic idea of this approach is to set up the 1-norm SVM problem as unconstrained minimization problem in the dual space. This method has been tested on a wide variety of data sets and compared with other methods (Fung and Mangasarian, 2004; Mangasarian, 2006). However, this method ends up with a high complexity due to the number of the resulted parameters to be adjusted (five including the regularization parameter), which makes it impracticable by non proficient users. Furthermore, as it is shown in this chapter, reaching an optimal global solution by the adopted Newton method is not always guaranteed. Nevertheless, it has been pointed out recently by (Chapelle, 2007) that dual and primal optimization problems are two equivalent ways of reaching the same result. Indeed, it has been shown that the primal problem can be solved efficiently without need to pass by the dual formulation.

A generic approach based on descent gradient technique has been recently proposed, referred to as DGM for Direct Gradient Method, capable to solve various ℓ_1 regularized learning problems, provided that the loss function is differentiable (Cai *et al.*, 2010a). It has been shown through an application on ℓ_1 regularized logistic regression that this method has the advantage to provide a simple and fast implementation. We show in the present work that the ℓ_1-SVM problem can be solved easily in the primal domain by using a generic gradient-descent technique based on DGM (Cai *et al.*, 2010b). The basic idea is to transform a convex optimization problem with a non-differential objective function into an unconstrained one. It has been proved theoretically therein that if the initial point is properly selected, DGM provides an optimal global solution. We take advantage here of this property to extend it to solve one of the important problems in machine learning; the ℓ_1-SVM problem in the primal domain. This approach is however not straightforward provided that the hinge loss function in the objective function is non-differentiable. To overcome this problem, we replaced the non-differentiable hinge loss function by its approximate differentiable Huber loss function. It has been pointed out indeed that the SVM using this loss function provides the same sparse solution as SVM with the hinge loss function within certain condition (Chappelle, 2007). We then transform the initial constrained convex optimization problem into an unconstrained problem in the primal domain. Some numerical experiments was performed to compare the proposed approach with the Newton family approaches proposed by (Fung and Mangasarian, 2004; Mangasarian, 2006). We demonstrate that our algorithm, though simple, outperforms this method in term of computational efficiency and the optimal quality of the obtained solution.

It is worthwhile to note that this work has been performed during a research stay in ICBR (Interdisciplinary Center for Biotechnology Research) at the University of Florida, under the supervision of Ph.D Yijun Sun. The chapter is organized as follows. Section 2 describes the main idea of the DGM approach. Section 3 presents the detailed implementation of the ℓ_1-SVM method. Section 4 presents some numerical experiments to compare the new approach with one of the well known state-of-art algorithm.

3.1 Gradient descent based method for solving ℓ_1 regularized problems

This section describes the main idea of the Gradient descent method for solving ℓ_1 regularized problems. This description has been taken from (Cai *et al.*, 2010a). Let $D = \left\{x^{(n)}, y_n\right\}_{n=1}^{N}$ denote

a training dataset, where $\mathrm{x}^{(n)} \in \Re^J$ is the n-th pattern and $\mathrm{y}_\mathrm{n} \in \Re$ is the corresponding class. We seek an optimal solution $(\mathrm{w}^*, \mathrm{b}^*)$ to the following ℓ_1 regularized learning problem:

$$\min_{\mathrm{w},\mathrm{b}} f_1(\mathrm{w},\mathrm{b}) = \frac{1}{N}\sum_{n=1}^{N} L\left(y_n, \mathrm{w}^T \mathrm{x}^{(n)} + \mathrm{b}\right) + \lambda \|\mathrm{w}\|_1 \tag{3.1}$$

Where $\|\mathrm{w}\|_1 = \sum_j |\mathrm{w}_\mathrm{j}|$, w_j is the j-th element of w, $L(\cdot)$ is a loss function and λ is a regularization parameter that controls the sparseness of the solution. We herein require that $L(\cdot)$ be a convex and differentiable function with respect to the second argument. The above formulation encompasses a wide range of learning algorithms, including LASSO (Tibshirani, 1996) and ℓ_1 regularized logistic regression algorithm (Ng, 2004). If a modified hinge loss is used (see, for example, (Rennie & Srebro, 2005; Chapelle, 2007)), equation (3.1) represents an approximate formulation of ℓ_1-SVM.

The above formulation has a very appealing property for high-dimensional data analysis. It has been proved in (Rosset *et al.*, 2004) that solving problem (3.1) leads to a globally optimal solution w^* with at most N non-zero elements. When $N << J$, it provides an explicit mechanism to perform feature selection to significantly reduce model complexity. This property, however, comes at a price. Unlike ℓ_2 regularization, $\|\mathrm{w}\|_1$ is a non-differentiable function of w. The efficient implementation of ℓ_1 regularized formulations poses a serious challenge to the machine learning community. We below show how a simple gradient descent technique can be used to efficiently solve ℓ_1 regularized learning problems.

Denote $\overline{\mathrm{x}}^{(n)} = \left[(\mathrm{x}^{(n)})^T, -(\mathrm{x}^{(n)})^T\right]^T$. Let us consider the following optimization problem:

$$\begin{aligned} &\min_{\overline{\mathrm{w}},\mathrm{b}} f_2(\overline{\mathrm{w}},\mathrm{b}) = \frac{1}{N}\sum_{n=1}^{N} L\left(y_n, \overline{\mathrm{w}}^T \overline{\mathrm{x}}^{(n)} + \mathrm{b}\right) + \lambda \sum_{\mathrm{i}=1}^{2\mathrm{J}} \overline{w}_i \\ &\textit{s.t.} \quad \overline{\mathrm{w}} \geq 0 \end{aligned} \tag{3.2}$$

The following lemma shows that the solution to (3.1) can be recovered from the solution to (3.2).

Lemma 3.1. Let $(\overline{\mathrm{w}}^*, \mathrm{b}^*)$ be an optimal solution to (3.2) where $\overline{\mathrm{w}}^* = \left[(\overline{\mathrm{w}}^{*(1)})^T, (\overline{\mathrm{w}}^{*(2)})^T\right]^T$ and $\overline{\mathrm{w}}^{*(1)}, \overline{\mathrm{w}}^{*(2)} \in \Re^J$. Then, $(\overline{\mathrm{w}}^{*(1)} - \overline{\mathrm{w}}^{*(2)}, \mathrm{b}^*)$ is an optimal solution to (3.1). Also, if $(\mathrm{w}^*, \mathrm{b}^*)$ is an optimal solution to (3.1), then there exists $\overline{\mathrm{w}}^{o(1)}$ and $\overline{\mathrm{w}}^{o(2)}$, so that $\mathrm{w}^* = \overline{\mathrm{w}}^{o(1)} - \overline{\mathrm{w}}^{o(2)}$ and $\left(\left[(\overline{\mathrm{w}}^{o(1)})^T, (\overline{\mathrm{w}}^{o(2)})^T\right]^T, \mathrm{b}^*\right)$ is an optimal solution to (3.2).

Proof. See Appendix 2.

The following lemma shows that at least half of the elements of the optimal solution to (3.2) are zero. We will exploit this property in our algorithm implementation in Section 3.

Lemma 2.2. Let $(\overline{\mathrm{w}}^*, \mathrm{b}^*)$ be an optimal solution to (3.2) and $\overline{\mathrm{w}}^* = [(\overline{\mathrm{w}}^{*(1)})^T, (\overline{\mathrm{w}}^{*(2)})^T]^T$. Then, $\forall j \in [\Im] = [1,..., J]$, either $\overline{w}_j^{*(1)}$ *or* $\overline{w}_j^{*(2)}$ or both equal to zero.

Proof. See Appendix 2.

The conversion from (3.1) to (3.2) is a standard step that has been previously used in many algorithms (e.g. (Schmidt *et al.*, 2007) and (Duchi *et al.*, 2008)). Note that Eq. (3.2) is a constrained convex optimization problem, with a differentiable objective function. In order to use gradient descent, we convert it into an unconstrained optimization problem.

Let $\overline{w}_j = v_j^2$, $\forall j \in 2[\Im]$. Then, (3.2) can be re-written as

$$\min_{\mathrm{v,b}} f(\mathrm{v,b}) = \frac{1}{N}\sum_{n=1}^{N} L\left(y_n, \sum_{j=1}^{2J} v_j^2 \overline{x}_j^{(n)} + \mathrm{b}\right) + \lambda \sum_{j=1}^{2J} v_j^2 \tag{3.3}$$

After the above transformation, the objective function of (3.3) is no longer a convex function, which is usually an undesitable property in optimization, except for some rare cases (Evtushenkjo and Zhadan, 1996; Faybusovich, 1991). We show by next that the transformation is beneficial in the sense that it not only preserves global convergence property of the original problem, but also enables removal of irrelevant features.

Taking the derivative of *f* with respect to v and b, respectively, yields

$$\begin{aligned} \frac{\partial f}{\partial \mathrm{v}} &= 2\left(\frac{1}{N}\sum_{n=1}^{N} \frac{\partial L(y_n, \sum_{j=1}^{2J} v_j^2 \overline{x}_j^{(n)} + \mathrm{b})}{\partial t} \overline{\mathrm{x}}_n + \lambda\right) \otimes \mathrm{v} \\ \frac{\partial f}{\partial \mathrm{b}} &= \frac{1}{N}\sum_{n=1}^{N} \frac{\partial L(y_n, \sum_{j=1}^{2J} v_j^2 \overline{x}_j^{(n)} + \mathrm{b})}{\partial t} \end{aligned} \tag{3.4}$$

where $\partial L(.)/\partial t$ is the derivative of L with respect to the second argument, and $\otimes$ is Hadamard operator.

For convenience, we denote $\overline{\mathrm{v}} = [\mathrm{v}^{\mathrm{T}}, \mathrm{b}]^{\mathrm{T}}$ and $g = [(\frac{\partial f}{\partial \mathrm{v}})^T, (\frac{\partial f}{\partial \mathrm{b}})^T]$. Let $\overline{\mathrm{v}}^{(k)}$ be the estimate of $\overline{\mathrm{v}}$ in the *k-th* iteration and $\mathrm{g}^{(k)}$ be the value of g at $\overline{\mathrm{v}}^{(k)}$. A gradient descent method uses the following updating rule

$$\overline{\mathrm{v}}^{(k+1)} = \overline{\mathrm{v}}^{(k)} - \eta\, \mathrm{g}^{(k)} \tag{3.5}$$

where η is determined via a line search.

Theorem 3.1. Let $f(\mathrm{w},\mathrm{b})$ be a differential convex function of w and b, where $\mathrm{w}\in\Re^{\mathrm{J}}, \mathrm{w}\geq 0$, and $\mathrm{b}\in\Re$. Let $G(\overline{\mathrm{v}}) = f(\mathrm{w},\mathrm{b})$ where $\mathrm{w} = [w_1,...,w_J]^T = [v_1^2,...,v_J^2]^T$ and $\mathrm{b} = v_{J+1}$. If $\left.\frac{\partial G}{\partial \overline{\mathrm{v}}}\right|_{\overline{\mathrm{v}}=\overline{\mathrm{v}}^+} = 0$, then $\overline{\mathrm{v}}^+$ is not a local minimizer, but a saddle point or a global minimizer of $G(\overline{\mathrm{v}})$. If the Hessian $\mathrm{H}(\overline{\mathrm{v}}^+)$ is positive semi-definite, then $\overline{\mathrm{v}}^+$ is a global minimizer.

Theorem 3.2. For $G(\overline{\mathrm{v}})$ and $\overline{\mathrm{v}}^+$ defined above, if $\overline{\mathrm{v}}^+$ is found through gradient descent with a line search satisfying the following conditions:

1. interval condition: a line searche splits the selection under search into a finite number of intervals,
2. Descending condition (see the definition below),
3. Greedy condition (see the definition below);

and an initial point $\overline{\mathrm{v}}^{(0)}$ staisfying $v_j^{(0)} \neq 0, \forall j\in[\Im]$, then with probability one, $\overline{\mathrm{v}}^+$ is a global minimizer of $G(\overline{\mathrm{v}})$.

Here we give the definitions of the descending and greedy conditions for a line search:

Definition 3.1. (descending condition). *Let $G(\overline{\mathrm{v}})$ be an objective function, $g(\overline{\mathrm{v}})$ be its gradient, $\overline{\mathrm{v}}^{(k)}$ be the solution obtained in the k-th iteration, and $-d^{(k)}$ be the descending direction, a line search is said to satisfy the descending condition if the chosen step length η satisfies*

$$d^{(k)T} g\left(\overline{\mathrm{v}}^{(\mathrm{k})} - \eta d^{(k)}\right) > 0,$$

Definition 3.2. (greedy condition). *Given $G(\overline{\mathrm{v}})$ and $g(\overline{\mathrm{v}})$ defined above, and $\varepsilon^{(k)}$ be the length of the intervals at the k-th iteration, a line search is said to satisfy the greedy condition if the step length η chosen satisfies*

$$G\left(\overline{\mathrm{v}}^{(k)} - \eta d^{(k)}\right) \leq G\left(\overline{\mathrm{v}}^{(k)} - (\eta+\varepsilon^{(k)}) d^{(k)}\right),$$

or $\left(\overline{\mathrm{v}}^{(k)} - (\eta+\varepsilon^{(k)}) d^{(k)}\right)$ is excluded from the line search;and

$$G\left(\overline{\mathrm{v}}^{(\mathrm{k})} - \eta d^{(k)}\right) \leq G\left(\overline{\mathrm{v}}^{(\mathrm{k})} - (\eta-\varepsilon^{(k)}) d^{(k)}\right),$$

or $\left(\overline{\mathrm{v}}^{(k)} - (\eta-\varepsilon^{(k)}) d^{(k)}\right)$ is excluded from the line search.

The proof of theorems 3.1 and 3.2 is given in Appendix 2.

The interval condition prevents the algorithms from over-exploring along a gradient descent direction. The descending and greedy conditions ensure that a line search approaches a local

optimum along the descending direction, but never hits or goes beyond it. With these conditions a gradient descent method can improve its quality step by step and to be immune from misleading gradient information. Golden section search (Kiefer, 1953) is an example that satisfies the greedy condition and splits the section into finite intervals according to the golden section rule.

3.2 Implementation details

We present below the detailed implementation of DGM for solving ℓ_1-SVM problem in the primal domain. In ℓ_1-SVM the following hinge loss function is usually used:

$$L(y,a) = \max(0, 1 - y\,a) \tag{3.6}$$

However, the hinge loss function is not differentiable and therefore the application of DGM method is not straightforward in this case. To overcome this problem, we replace the non-differential hinge loss function by its approximate differentiable Huber loss function as suggested by (Chapelle, 2007), given by

$$L(y,a) = \begin{cases} 0 & ya > 1+h \\ \dfrac{(1+h-ya)^2}{4h} & 1-h \le ya \le 1+h \\ 1-ya & ya < 1-h \end{cases} \tag{3.7}$$

where h is a tunable parameter. If h is sufficiently small, SVM using the Huber loss provides the same sparse solution as SVM with the hinge loss (Chapelle, 2007). Hence, DGM described in section 2 can be directly used to solve ℓ_1-SVM in the primal domain. The gradients of f in Eq.(3.3) with respect to **v** and b in this case is given as follows

$$\frac{\partial f}{\partial \mathrm{v}} = \begin{cases} 2\lambda \otimes \mathrm{v} & ya > 1+h, \\ 2\left(\lambda - \dfrac{1}{2Nh}\sum_{n=1}^{N} y_n\left(\left(1+h-y_n\left(\overline{\mathrm{w}}^T\overline{\mathrm{x}}^{(n)}+\mathrm{b}\right)\right)\mathrm{x}^{(n)}\right)\right)\otimes \mathrm{v} & 1-h \le ya \le 1+h, \\ 2\left(\lambda - \dfrac{1}{N}\sum_{n=1}^{N} y_n\overline{\mathrm{x}}^{(n)}\right)\otimes \mathrm{v} & ya < 1+h, \end{cases} \tag{3.8}$$

and

$$\frac{\partial f}{\partial \mathrm{b}} = \begin{cases} 0 & ya > 1+h, \\ \left(-\frac{1}{N}\sum_{n=1}^{N} y_n\right) & 1-h \le ya \le 1+h, \\ \left(-\frac{1}{N}\sum_{n=1}^{N} y_n\right) & ya < 1+h, \end{cases} \tag{3.9}$$

The gradient descent steps in (3.5) is then applied. In each step, we first apply back-tracking line search (Nocedal and Wright, 1999) to obtain an end point $\bar{\mathrm{v}}^{(e)}$ where $f(\bar{\mathrm{v}}^{(e)}) \le f(\bar{\mathrm{v}}^{(k)})$, then apply golden section line search on the section between $\bar{\mathrm{v}}^{(k)}$ and $\bar{\mathrm{v}}^{(e)}$.

3.2.1 Hybrid Conjugate Gradient

Because a simple gradient descent method is known to zig-zag in some function contours, the Fletcher-Reeves conjugate gradient descent method (Fletcher, 1997) can be used to enhance the performance of the algorithm:

$$\begin{aligned} \bar{\mathrm{v}}^{(k+1)} &= \bar{\mathrm{v}}^{(k)} - \eta\, \mathrm{d}^{(k)} \\ \mathrm{d}^{(1)} &= g^{(1)} \\ \mathrm{d}^{(k)} &= g^{(k)} + \beta\, \mathrm{d}^{(k-1)}, \qquad \forall k > 1, \\ \beta &= \frac{\left\langle g^{(k)}, g^{(k)} \right\rangle}{\left\langle g^{(k-1)}, g^{(k-1)} \right\rangle}, \end{aligned} \tag{3.10}$$

where $\mathrm{d}^{(k)}$ is the conjugate gradient, and $\langle . \rangle$ is the inner product. Note that conjugate gradient method does not ensure that the objective function decreases monotonically. Hence, when $\left\langle g^{(k)}, d^{(k)} \right\rangle \le 0$, one usually replaces $\mathrm{d}^{(k)}$ with $g^{(k)}$ as the search direction to ensure that the algorithm always proceeds in a descending direction. In all our implementation, we adopt a hybrid gradient descent scheme. Denote $f^{(k)}$ as the objective function obtained in the *k*-th iteration and $\angle(a, \mathrm{b})$ the angle between vectors a and b. If $(f^{(k)} - f^{(k-1)}) / f^{(k)} < \theta_1$ and $\angle(g^{(k)}, \mathrm{d}^{(k)}) < \theta_2$, we use $-d^{(k)}$ as the descending direction, and $-g^{(k)}$ otherwise. In all our implementations, we set $\theta_1 = 0.01$ and $\theta_2 = 5/12\pi$. It should be noted that with the descending condition, the global convergence property also holds for conjugate gradient descent.

We stated in the previous section that the solution $\bar{\mathrm{w}}^{*}$ has at most *min*(*N*, *J*) non-zero elements. We exploit this property to speed up the implementation. Note in (3.4) that if $v_j = 0$, then the gradient will be zero on the *j*-th element, and v_j will remain zero thereafter. Hence,

if some elements of v are extremely small, the corresponding features can be eliminated from further consideration with a negligible impact on the subsequent iterations and the final solution found. In all our implementations, the criterion for eliminating small-valued weights is $w_j < 10^{-10}\|\mathrm{w}\|_\infty$, where $\|\mathrm{w}\|_\infty = \max_j \{w_j\}$.

3.2.2 Computational Complexity

With conjugate gradient descent, in each iteration, the flops needed to compute gradient is *O*(*NJ*), and the memory required is *O*(*N* +*J*), where *N* is the sample size and *J* the data dimensionality. We give by the following the pseudo-code of DGM Algorithm.

DGM Algorithm

1. *Initiate* $\bar{\mathrm{v}}^{(0)} = 1/\sqrt{(2J)}, \mathrm{b}^{(0)} = 0, k = 0$, *stopping criteria* δ, *parameters* θ_1, θ_2
2. $\bar{\mathrm{v}}^{(0)} = [(\mathrm{v}^{(0)})^T, b]^T$
3. *Compute* $f^{(0)}$ *using Eq. (3.3)*
4. *Repeat*
   a- *k=k+1*
   b- Compute $g^{(k)}$ using Eq. (3.7)
   c- *If k>1 and* $\left\| f^{(k)} - f^{(k-1)} \right\| < \theta_1$ then
      Compute $d^{(k)}$ using Eq. (3.7)
      If $\angle(g^{(k)}, \mathrm{d}^{(k)}) < \theta_2$ then
         $d^{(k)} = g^{(k)}$
      End if
   Else
      $d^{(k)} = g^{(k)}$
   End if
   d- Update $\bar{\mathrm{v}}^{(k)} = \bar{\mathrm{v}}^{(k-1)} - \eta^{(k)}\, \mathrm{d}^{(k)}$, where $\eta^{(k)}$ is determined via line search.
   e- *If* $\bar{v}_j^{(k)} < 10^{-5}\left\|\bar{\mathrm{v}}^{(k)}\right\|_\infty, \forall j \in 2[\Im]$ then
      $\bar{v}_j^{(k)} = 0$
   End if
5. *until* $\left\| f^{(k)} - f^{(k-1)} \right\| < \delta$
6. $\bar{\mathrm{w}}^{(1)} = [(\bar{v}_1^{(k)})^2, ..., (\bar{v}_J^{(k)})^2]^T$
7. $\bar{\mathrm{w}}^{(2)} = [(\bar{v}_{J+1}^{(k)})^2, ..., (\bar{v}_{2J}^{(k)})^2]^T$
8. $\mathrm{w} = \bar{\mathrm{w}}^{(1)} - \bar{\mathrm{w}}^{(2)}$
9. $\mathrm{b} = v_{2J+1}^{(k)}$

3.3 Numerical experiments

We present in the following some numerical experiments to compare DGM-ℓ_1SVM with one recent state-of-the-art method, namely, generalized LPNewton family algorithms proposed in

(Fung and Mangasarian, 2004; Mangasarian, 2006). Indeed, the method described in (Mangasarian, 2006) is only a special case of the one proposed by (Fung and Mangasarian, 2004) where the penalty parameter α is fixed to be one. It was stated in (Mangasarian, 2006) that this value of α leads to an exact solution of the SVM problem. This method has been tested on a wide variety of data sets and compared with other methods such as standard software packages (Fung and Mangasarian, 2004; Mangasarian, 2006).

3.3.1 Experiment Setup

Each algorithm has been applied to eight datasets using a specified set of λ values. Each algorithm is stoped when the achieved objective function is within a desired precision of the optimal solution. However, only a locally optimum solution is may be achieved which makes the comparison in term of CPU time in this case unbalanced. In order to make a fair comparison, in our experiments, for every dataset and λ value, we first run both algorithms within 10^{-6} precision. At the end of this stage, each algorithm provided one solution (w^*, b^*). Obviously, the good solution is the one which provided the minimal cost value on the objective function of the original SVM problem which has to be minimized (Eq. 3.1). Then, we set the so-obtained minimal cost achieved over both algorithms as the target value and we run each algorithm so that it stopped when the achieved objective function was within10^{-6} precision of this target cost. However, it is possible that the algorithm diverges and never achieves the desired target cost. To overcome this problem, the maximum number of iteration for each λ value was fixed to be 5×10^3. The CPU time consumed to achieve the target cost was then recorded and compared. By using this experimental protocol, we verified that the solution obtained by DGM-ℓ_1SVM was, as proved theoretically, a global minimizer. LPNewton algorithm was programmed in Matlab as provided by (Fung and Mangasarian, 2004). For a fair comparison, we developped DGM-ℓ_1SVM also on Matlab. LPNewton algorithm requires the specification of many parameters including regularization parameter. It is worthwhile to note here that for DGM-ℓ_1SVM the only requisite is to specify the regularization parameter λ as the parameter h must be specified to be very small (here we take h=10^{-8}) , in order to guarantee the same sparse solution as that would be obtained when a hinge loss function is used (Chapelle, 2007). In our experiments the values of λ (or equivalently $1/\nu$ in LPNewton algorithm) are taken in the range [2^{-7}, 2^7] and $\varepsilon = 10^{-1}$, δ belongs to the interval [10^{-3}, 10^3] as suggested by (Fung and Mangasarian, 2004). For the parameter α, it must be noted that we have found out empirically that this method performed

poorly when the special case $\alpha = 1$ was considered. For that reason we opted to take $\alpha = 10^3$ as suggested also by (Fung and Mangasarian, 2004). It must be noted also that all experiments was performed on a personal computer with Intel Core 2, 2.26 GHZ CPU, 1.98 GB memory, and Windows operating system.

3.3.2 Experimental results

We have compared the two algorithms on eight medium and large-scale data sets with feature dimensionality ranging from 1,000 to 44,932. Seven among them are cancer microarray data: Colon, leukemia, internet Ads. (Koh *et al.*, 2007; Lee et *al.*, 2006), prostate cancer (Stephenson et *al.*, 2005), GSE4922 (Ivshina *et al.*, 2006), Arcene (Guyon *et al.*, 2005) , ETABM77 (Buyse *et al.*, 2006). The Linear data is an artificially generated binary classification problem, with each class having 200 samples characterized by 104 features. The first 500 features are drawn from two normal distributions N (−1, 1) and N (1, 1), depending on class labels. The rest of the features are drawn from the standard normal distribution, thus providing no discriminant information. The internet Ads Data has a sparse data matrix where only a few features have non-zero values, whereas all other datasets have a dense data matrix. The summary of the data is given in Table 3.1. For each dataset, standardization was performed on the data matrix so that the effect of mean shift in microarray profiling is reduced.

Table 3.1 Summary of datasets

Dataset	No. of features	No. of samples
Colon cancer	2000	62
Leukemia	7129	72
Internet Ads.	1430	2359
Prostate cancer	22291	79
TABM77	1145	291
GSE4922	44932	249
Arcene	10000	200
Linear	10000	400

We have applied the two algorithms on each dataset and recorded in Table 3.2 the total running time summed over fifteen λ values uniformly spaced on a logarithmic scale over interval $[2^{-7}, 2^{7}]$. Indeed, the regularization parameter λ is usually estimated in practical applications, through a cross validation procedure. Hence, the total running time summed over all possible λ values is an important criterion to evaluate an algorithm. We plot also the CPU

time and the corresponding precision in term of cost for each λ value as shown in Figures 3.1 and 3.2. It can be observed that:

1. Figure 3.1 shows that DGM-ℓ_1SVM outperforms the LPNewton algorithm for all λ values on all datasets. The overall CPU time reported in Table 3.2 confirms this result.
2. Precision and the CPU time plotted in Figures 3.1 and 3.2 show that the minimal cost value is always achieved by DGM-ℓ_1SVM, which is consistent with our claim supported by a well founded theoretical demonstration that it provides a global minimum.
3. Except for Linear data set, when λ is large, LPNewton algorithm fails to converge to the target optimal cost which justifies the obtained poor precision generally when $\lambda > 2^3$ as shown in Figure 3.2.
4. DGM-ℓ_1SVM converges always in a finite time to a solution corresponding to the minimal cost whatever the value of λ.

To further demonstrate the effectiveness of our approach, we report also in Table 3.3 the overall summed CPU time for only the eleven λ values $[2^{-7},2^3]$ for whom both methods converge (before $5x10^3$ iterations). These results confirm that the proposed approach outperforms LPNewton approach even in the cases when this last converges to a finite solution. One possible explanation is that the solution provided by DGM-ℓ_1SVM is more optimal than the so-obtained by LPNewton approach.

Tab. 3.2. CPU time (in seconds) of the two algorithms performed on the eight data sets for all λ values. The algorithm stops when the achieved objective function is within 10^{-6} precision of the target cost.

Data/Method	Colon	Leuki.	Internet Ads.	Prostate	ETABM77	GSE4922	Arcene	Linear
DGM	**4.1**	**11**	**57.7**	**40.3**	**21.3**	**497.9**	**2149**	**86.8**
LPNewton	317	806	3012	2147	1105.4	11532	41631	5051.1

Tab. 3.3. CPU time (in seconds) of the two algorithms performed on the eight data sets for only eleven λ values. The algorithm stops when the achieved objective function is within 10^{-6} precision of the target cost.

Data/Method	Colon	Leuki.	Internet Ads.	Prostate	ETABM77	GSE4922	Arcene
DGM	**3.6**	**10**	**50.5**	**36.2**	**20.3**	**462.2**	**1794.6**
LPNewton	114.8	132	766.1	438.5	366.9	2865.9	2228.6

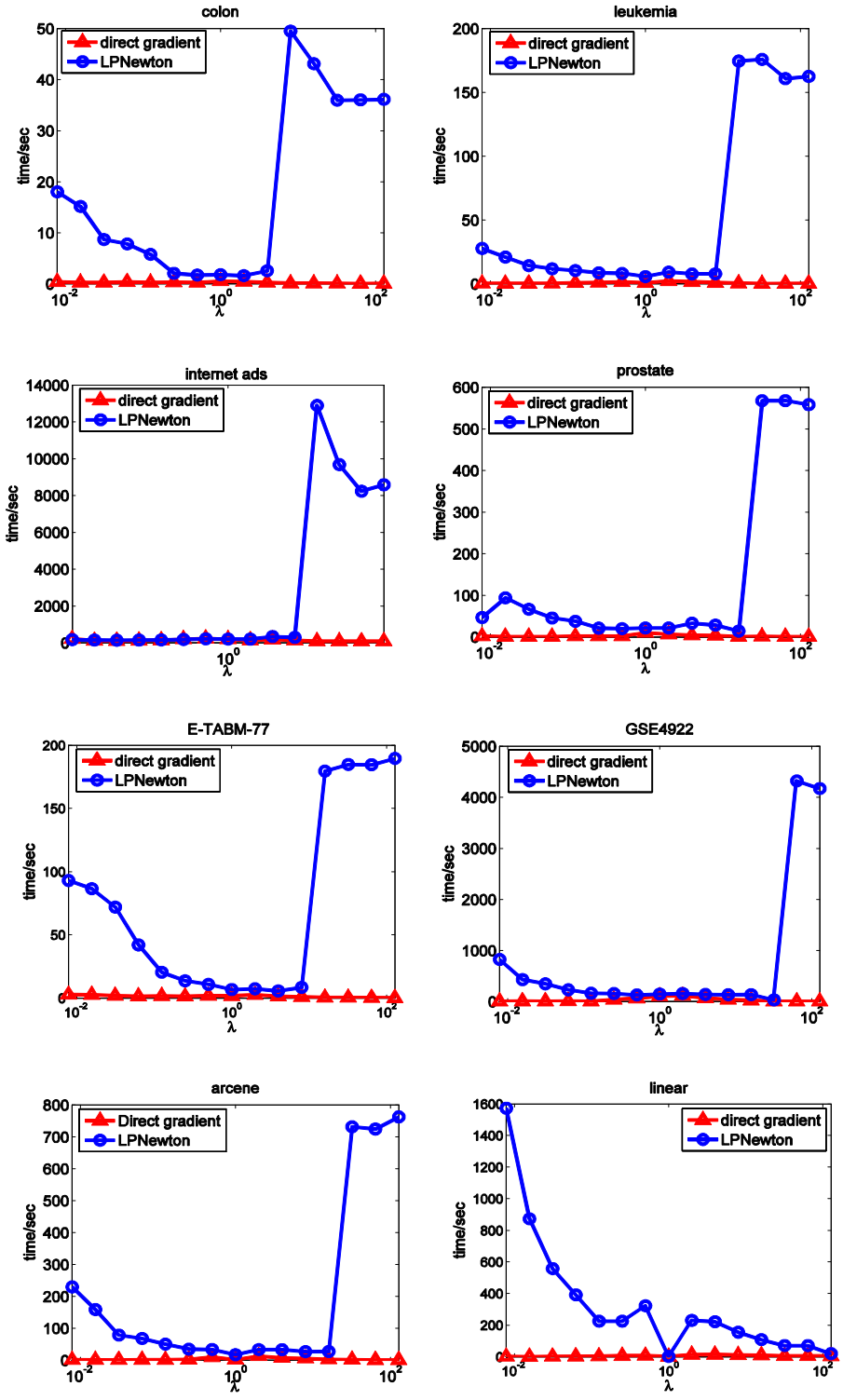


Fig. 3.1. Running time (in seconds) of DGM-ℓ_1SVM and LPNewton performed on eight benchmark data sets using different λ values.

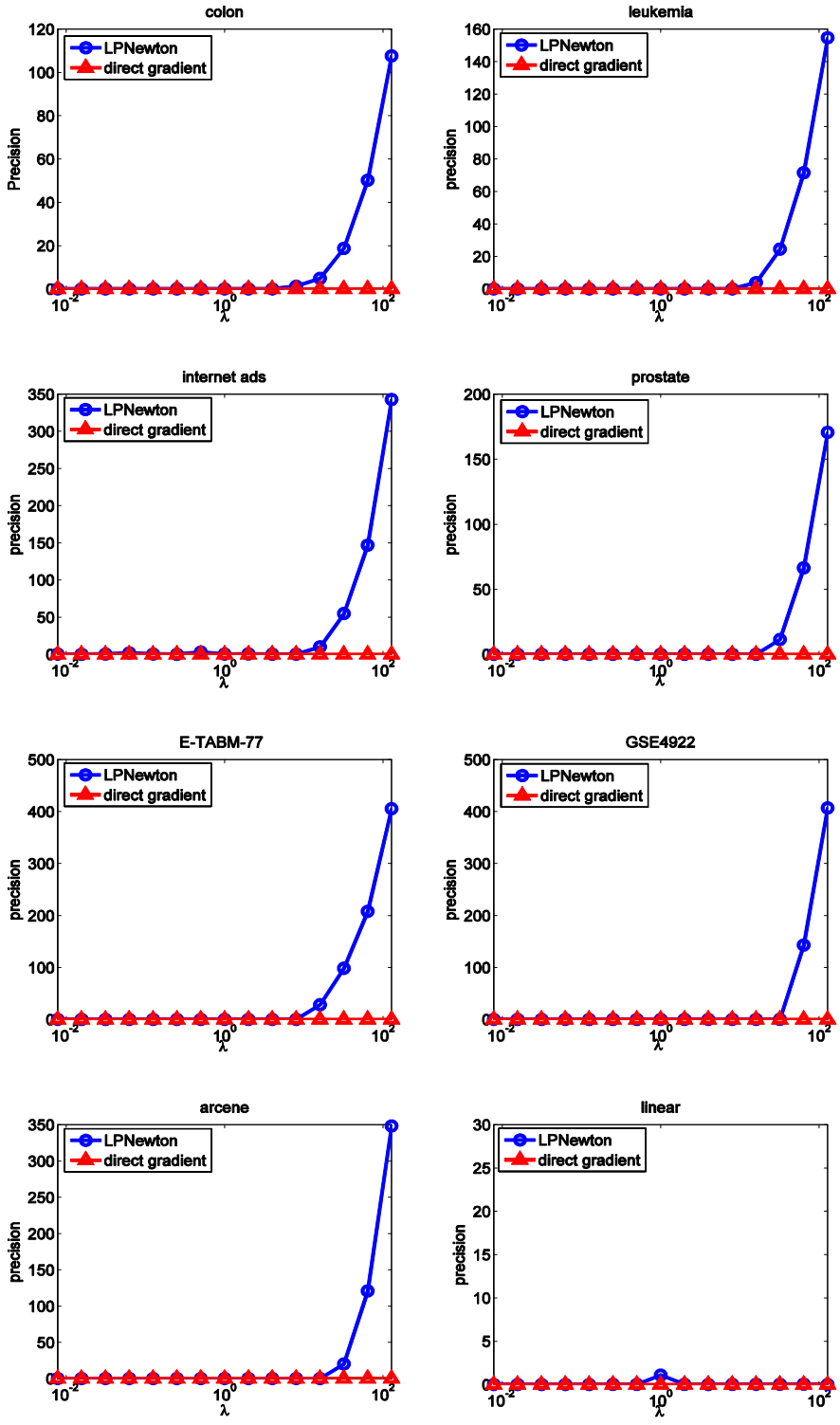


Fig. 3.2. Precision of DGM-ℓ_1SVM and LPNewton performed on eight benchmark data sets using different λ values.

3.4 Conclusion

We have proposed in this chapter an efficient approach to solve the ℓ_1SVM problem in the primal. We have shown that the proposed method, though simple, perform very well in practical situations. The basic idea is to take advantage of the global solution optimality which can be achieved using gradient descent techniques. Firstly, the hinge loss function is replaced by its approximated Huber loss function to overcome its non-differentiable property. Then, the initial convex optimization problem is transformed into an unconstrained non-convex problem, upon which, via gradient descent, reaching a globally optimum solution is guaranteed. We have conducted large-scale numerical experiments to demonstrate the theoretical claim and prove the computational efficiency over a well known state-of-art method.

High data dimensionality however is not the only problem to be faced in cancer applications. Other major issues such as data heterogeneity, uncertainties and noises can also be encountered jointly with high dimensionality problem. Therefore, more efficient methods are urgently needed to cope simultaneously with all above stated problems. This problematic will present our subject of interest in next chapters in an attempt to develop appropriate approaches capable of handling such problems.

CHAPITRE 4- Résumé
Vers un principe unifié pour le raisonnement sur des données hétérogènes

Pour une bonne compréhension du comportement d'un processus, le raisonnement humain traite habituellement des connaissances incomplètes et hétérogènes. Par conséquent, des méthodes appropriées pour représenter le processus avec des connaissances partielles sont nécessaires. La représentation la plus utilisée est celle purement quantitative qui suppose une exactitude complète de l'information. Cependant, un nombre réel contient une quantité infinie de précision alors que la connaissance humaine est finie et discrète. Ainsi, il est nécessaire d'utiliser les données représentées par des valeurs symboliques pour s'adapter à la perception humaine. La représentation des connaissances incomplètes sur les données peut être effectuée de différentes manières: intervalles symboliques ou valeurs qualitatives. Par conséquent, le développement d'un mécanisme automatique de raisonnement sur les données est confronté à cette multiplicité de représentations possibles.

Nous abordons dans ce chapitre l'une des principales difficultés rencontrées dans les tâches d'analyse de données: la diversité des types d'information. Une telle information est représentée par des données qualitatives, nominales ou ordinales, mélangées avec des données quantitatives et intervallaires. Notre objectif est de proposer un principe unifié pour établir différents mécanismes de raisonnement en utilisant simultanément trois types de données: purement quantitative, intervallaire symbolique et qualitatives. De nombreuses situations menant à des algorithmes bien conditionnés pour les données quantitatives, deviennent très complexes lorsque certaines informations sont sous forme qualitative. Dans une liste non exhaustive, on peut citer, déduction basée sur les règles, classification, «clustering», la réduction de dimensionnalité ....

Pour surmonter ce problème, une approche classique consistera à raisonner sur chaque type de données séparément pour déduire des décisions partielles. Cependant, cela représente un autre problème sérieux similaire à notre problème initial lié à la façon dont on doit procéder pour intégrer de telles décisions partielles et finir avec une décision globale pour l'ensemble des données. Une autre approche intéressante serait d'unifier les différents espaces

hétérogènes dans un espace homogène, puis raisonner de manière unifiée sur l'ensemble des données pour prendre la décision appropriée. Afin d'éviter tout type de distorsion et/ou perte d'information, le processus d'unification de l'espace devrait être effectué avec précaution pour chaque type de données. Dans ce but, nous introduisons ici un principe unifié permettant de raisonner sur des données hétérogènes, dénommé SMSP pour Simultaneous Mapping for Single Processing. Le principe est basé initialement sur une projection appropriée et simultanée des données hétérogènes dans un espace unifié. Cette projection peut être obtenue en utilisant une fonction caractéristique pour chaque type de données pour les amener dans un espace homogène. Ces fonctions peuvent être conçues de telle façon qu'elles expriment une mesure relative comme par exemple la mesure du degré d'adéquation (ou typicité) de chaque valeur d'une variable à des partitions existantes. Par exemple, dans le cadre de la théorie des ensembles flous, cette mesure est techniquement synonyme du terme de mesure d'appartenance qui est un nombre réel dans l'intervalle unitaire I = [0,1]. Quand une mesure relative différente, autre que l'adéquation à chaque classe, doit être considérée d'autres solutions alternatives peuvent être envisagées. Une solution possible est d'utiliser le concept de fonction noyau (Atkeson et al., 1997), popularisé dans le cadre de la théorie d'apprentissage statistique, pour la conception des fonctions caractéristiques adaptées à chaque type de données. Une fois que ces fonctions appropriées ont été choisies et que toutes les données sont représentées dans un espace homogène, un traitement unique peut être effectué en utilisant un mécanisme de raisonnement unique. Afin de prendre en compte l'incertitude d'appartenance, le principe SMSP est proposé ici dans le cadre de la théorie des ensembles flous. Une fois que les fonctions d'appartenance adaptées caractérisant l'adéquation à chaque classe sont choisies en fonction des types de variable, une partition floue des variables peut être effectuée à partir des données empiriques. Comme il est montré dans ce chapitre, chaque individu de la base de donnée initiale, décrite par m variables de plusieurs types (qualitative, quantitative, intervallaire), sera représenté par m degrés d'appartenance, c.-à-d. m nombres de l'intervalle unitaire. Les données transformées sont donc inclues dans un espace homogène isomorphe à un hypercube unité. Par conséquent, un mécanisme flou de raisonnement simple et unique peut être utilisé pour raisonner sur les données obtenues quel que soit leur type initial. On démontrera dans les chapitres suivants qu'en utilisant ce principe, il est possible d'effectuer une variété de tâches d'analyse de données (classification, réduction de dimensionnalité, regroupement ...).

CHAPTER 4
Towards a Unified Principle for Reasoning about Heterogeneous Data: A Fuzzy Logic Framework

For a good understanding of any process behavior, human reasoning deals usually with incomplete and heterogeneous knowledge. Therefore, appropriate methods for representing the process with partial knowledge are required. The most used representation is the pure quantitative one which assumes a complete exactitude about the information. Taken as it appears, a real number contains an infinite amount of precision whereas human knowledge is finite and discrete. So, there is a need to use data represented by symbolic values to fit with human perception. The representation of incomplete knowledge about the data can be done in different ways: symbolic intervals or qualitative values. Thus, the development of an automatic mechanism for reasoning about the data is faced with this multiplicity of possible representations.

We address here one of the main difficulties encountered in data analysis tasks: the diversity of information types. Such information is given by qualitative valued data, which can be nominal or ordinal, mixed with quantitative and interval data. Our focus is to propose a unified principle to establish various reasoning mechanisms using simultaneously three types of data: pure quantitative, symbolic interval and pure qualitative modalities. Many situations leading to well conditioned algorithms for quantitative valued information, become very complex whenever there are several data given in qualitative form. In a non exhaustive list, we can mention, rule based deduction, classification, clustering, dimensionality reduction… Although the problem of representation multiplicity has been addressed within the machine learning framework in some works (Michalski and Stepp, 1980; Mohri and Hidehiko, 1994; Hu *et al.*, 2007), no standard principle has been proposed in the literature to handle in a unified way heterogeneous data. The proposed methods respectively use distance and information content measures to process separately quantitative and qualitative in dimension reduction tasks (Kira and Rendell, 1992a; Dash and Liu, 2003), whereas a Hamming distance is usually used to handle qualitative data in classification and clustering tasks (Aha, 1989; Aha, 1992; Kononenko, 1994).

Other approaches are originally designed to process only quantitative data and therefore arbitrary transformations of qualitative data into a quantitative space are proposed without taking into account their nature in the original space (Cover and Hart, 1967; Kira and Rendell, 1992a; Weston *et al.*, 2001). For example, the feature *color* can take values in a discrete unordered set {red, black, green, white}. These values are transformed respectively to quantitative values 1, 2, 3 and 4. However, we can also choose to transform them to 4, 1, 2 and 3. This can represent a potential source of information loss.

In the opposite, the transformation of quantitative values in qualitative objects by discretizing the quantitative value domain into several intervals (Hall, 2000; Liu *et al.*, 2002) introduce also distortion and information loss with respect to the original data since objects in the same interval are labeled by the same qualitative value.

Although extensive studies were performed to process interval type data in the Symbolic Data Analysis framework (Bock and Diday, 2000), they were focused generally more on the clustering tasks (Gowda and Diday, 1992; De Carvalho *et al.*, 2010). Indeed, no standard principle has been proposed in the literature to handle in a unified way heterogeneous data and combine furthermore in a fully adequate way, the processing of symbolic intervals simultaneously with quantitative and qualitative data for different analysis purposes.

In this chapter we present a general principle, introduced here as "Simultaneous Mapping for Single Processing (SMSP)", which enables reasoning in a unified way about heterogeneous data for several data analysis purposes (Hedjazi *et al.*, 2010a; Hedjazi *et al.*, 2011a). The only requisite is to define characteristic functions that characterize a relative measure based on available knowledge about each feature. Once these functions are chosen appropriately, the initial heterogeneous space, where the information is of mixed nature, is transformed into a homogeneous space. Consequently, only a unique reasoning mechanism can be used to reason about the resulted data whatever its initial type.

We introduce below this principle noted SMSP principle and an example of simultaneous mapping of mixed features into a common space is presented within a fuzzy logic framework.

4.1 Simultaneous mapping for single processing principle

Many learning problems involve usually data of mixed type characterized especially within different heterogeneous spaces. The lacks of analogy between such spaces make the reasoning task to extract a reliable knowledge rather complex. To overcome this issue, due mainly to space's heterogeneity, one typical approach is to reason about each type of data separately to

derive separate partial decisions. However, this brings another serious issue similar to the initial one related to the way to integrate such partial decisions to end up with only a global decision for the whole data. Another interesting approach would be to unify the different heterogeneous spaces into one homogeneous space and then reason in a unified way about the whole data to take the appropriate decision. To avoid any type of distortion and/or information loss the space's unification process should be performed appropriately for each type of data. In this aim, we introduce here a unified principle for reasoning about heterogeneous data, referred to as Simultaneous Mapping for Single Processing (SMSP). The principle is based initially on an *appropriate* simultaneous mapping of heterogeneous data into a unified space. This mapping can be obtained by using a characteristic function for each type of data to bring them into a homogeneous space. These functions can be designed in such way that they express a relative measure as for example the measure of the appropriateness (adequacy, typicality) of each feature value of patterns to existing partitions. For instance, in the fuzzy set theory framework, this measure is technically synonymous to the term of membership measure which is a number of the real unit interval I= [0,1]. When a different relative measure other than the pattern appropriateness to each class is considered, other alternative solutions can be envisaged. One possible solution is to use the *kernel function* concept (Atkeson *et al.*, 1997), extensively studied in statistical learning theory, for designing suitable characteristic functions for each type of data. Once suitable functions are chosen and all data are represented in a homogeneous space, a single processing can be performed using a unique reasoning mechanism. The general concept of the SMSP principle is illustrated in Figure 4.1.

In order to take into account the membership uncertainty, the SMSP principle is proposed in the present work within the fuzzy set theory framework to reason about heterogeneous data. Once suitable membership functions that characterize the adequacy to each class are chosen according to feature types, a fuzzy partition of features can be performed based on empirical data. As it will be shown hereafter, each pattern of the initial data, described by *m* features having several types, (qualitative, quantitative, symbolic intervals), will be represented by *m* membership degrees, i.e. *m* numbers of the unit interval; therefore the transformed data set is included in a homogeneous space isomorph to an unit hypercube. Thus, a unique and simple fuzzy reasoning mechanism can be used to reason about the resulting data whatever its original type. It will be shown by next chapters that based on this principle it is possible to perform a wide variety of analysis (classification, dimensionality reduction, clustering…).

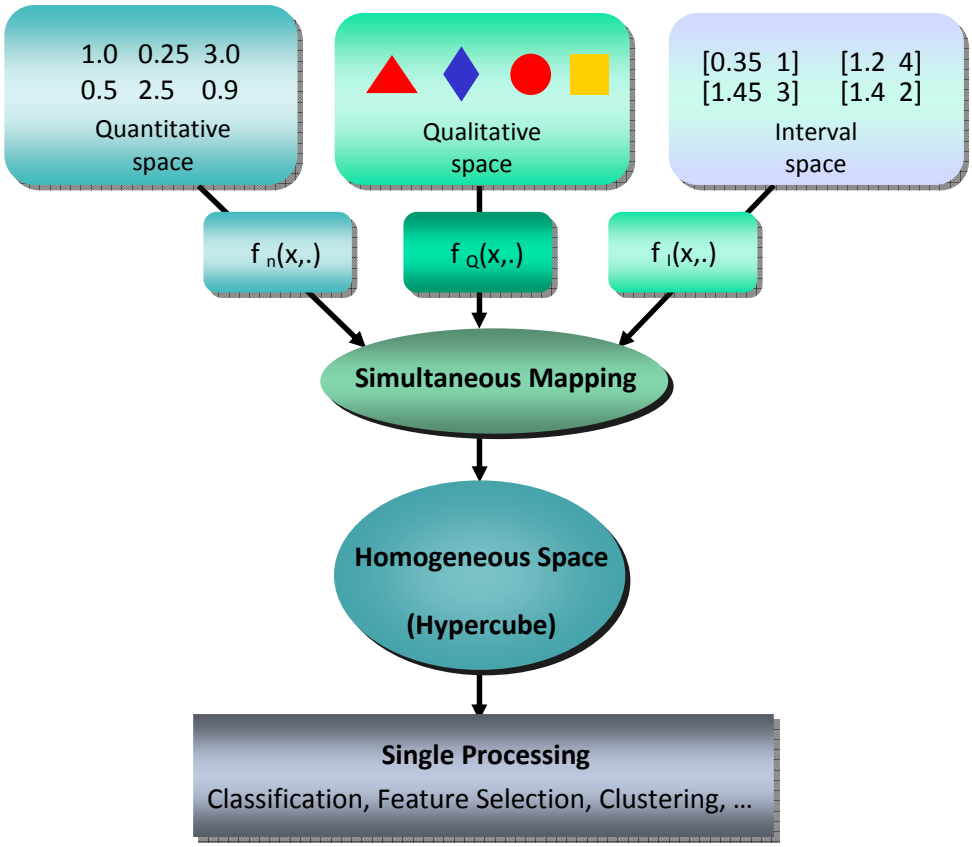


Fig. 4.1. SMSP principle

4.2 Homogeneous space of features

Basically we consider the three above mentioned types of data:

(a) *Quantitative features*: real numbers that can be normalized into the unit interval [0,1].

(b) *Symbolic intervals of the real line*: with no restriction of relative position (regular or overlapped).

(c) *Qualitative features:* that can be ordinal or nominal modalities.

According to (Dubois and Prade, 1997), three main semantics of fuzzy membership functions can be distinguished in the fuzzy literature. Among them we find similarity (or distance) and uncertainty, widely used in fuzzy pattern recognition, applied to the estimation of membership functions from data (Medasani and Kim, 1998). For instance, Bezdek (Bezdek, 1981) take use of the similarity semantic to define a relation r between two objects $x^{(1)}$ and $x^{(2)}$ as fuzzy (i.e $r \in [0,1]$ if $r = \rho(x^{(1)}, x^{(2)})$ where ρ is a metric (distance measure) otherwise it is considered as crisp (i.e $r \in \{0,1\}$). Regarding uncertainty semantic, it is reported in (Dubois and Prade, 1997) that uncertainty is often measured in terms of frequency of observed situations in a

random experiment. However, this approach leads to probability theory when the repeated observations are precise. In that sense, probability assignments to the elements of referential set U can be viewed as special membership functions such that the sum of membership grades is one (Dubois and Prade, 1997). In our work both semantics have been used to define an adequate membership function according to the type of data. However, in mixed data types the features are images of unrelated concepts. In order to bypass this difficulty, we take advantage of the commensurability assumption in the framework of fuzzy logic (Dubois and Prade, 1997) to end up with a unique space (unit hypercube) where all the features are represented by their memberships to a reference fuzzy partition. Therefore, a single processing of their membership degrees for data analysis purpose is straightforward based on aggregation in the resulting space.

Let $D = \{x^{(n)}, C_k\}_{n=1}^{N} \in X \times C$ be a dataset, where $x^{(n)} = [x_1^{(n)}, x_2^{(n)}, ..., x_m^{(n)}]$ is the nth pattern (item) and *N* is the number of patterns. Each pattern is represented by *m* features possibly of different types (quantitative, qualitative or symbolic interval), and C_k is the class label assigned to each pattern in the pre-established partitions*: k=1,2,…,l*.

Based on an appropriate data-driven process using the training dataset D, to each feature correspond *l* fuzzy sets representing the membership functions to each class. Namely, let $\{mff_1^i, mff_2^i \ldots, mff_l^i\}$ be the *l* fuzzy sets that form a fuzzy partition for the i^{th} feature.

The fuzzy set mff_k^i *is* defined by its membership function μ_k^i in the rank X_i of the i^{th} feature depending on a parameter θ_{ki} as follows:

$$\mu_k^i(x_i) = f_i(x_i, \theta_{ki}); \qquad k=1,2,\ldots,l \tag{4.1}$$

where θ_{ki} represents the i^{th} prototype of class C_k and can be estimated from the i^{th} feature values of patterns belonging to class C_k in the training dataset D. For each feature type, a particular learning process can be adopted to estimate its membership functions from data.

We present by next how this fuzzy partition is performed here using a particular membership functions to each type of feature and we support it by the following toy example.

Example: Consider the set of samples shown in Table 4.1. This set is classified into two classes $\{C_1, C_2\}$ and is described by three types of features: x_1 (quantitative feature), x_2 (interval feature) and x_3 (qualitative feature).

Tab. 4.1. Group of patterns characterized by three mixed-feature types

Samples	x_1 (Quantitative feature)	x_2 (Interval feature)	x_3 (Qualitative feature)	Class
$x^{(1)}$	5	[1.5 , 2.5]	Red	C_1
$x^{(2)}$	5.22	[3.5, 4.5]	Yellow	C_1
$x^{(3)}$	6.78	[5.5, 6.5]	Red	C_1
$x^{(4)}$	7	[7.5, 8.5]	Black	C_1
$x^{(5)}$	6	[9.5, 10.5]	Yellow	C_1
$x^{(6)}$	5.5	[11.5, 12.5]	Red	C_1
$x^{(7)}$	7	[13.5 , 14.5]	Black	C_1
$x^{(8)}$	9	[7.5 , 8.5]	Black	C_2
$x^{(9)}$	8	[9.5, 10.5]	White	C_2
$x^{(10)}$	10.5	[11.5, 12.5]	Red	C_2
$x^{(11)}$	8.5	[13.5, 14.5]	Black	C_2
$x^{(12)}$	9.5	[15.5, 16.5]	Black	C_2
$x^{(13)}$	10	[17.5, 18.5]	Red	C_2
$x^{(14)}$	11	[19.5, 20.5]	Black	C_2

4.3 Membership functions

4.3.1 Quantitative type features

It will be generally assumed that the universe of discourse of each quantitative feature is included in a compact interval; either the bounds of this interval are known, or they can be induced by the dataset. Therefore, without loss of information, its numerical values can be normalized within the interval $[x_{i\min}, x_{i\max}]$. This linear re-scaling of the feature into the interval [0,1] is performed by:

$$x_i = \frac{\hat{x}_i - \hat{x}_{i\min}}{\hat{x}_{i\max} - \hat{x}_{i\min}} \tag{4.2}$$

where $\hat{x}_i$ is the i^{th} raw feature value and x_i is its normalized value.

In the case of quantitative features, several membership functions proposed by (Aguado and Aguilar-Martin, 1999) can be used for $\mu_k^i(.)$. Among them we find:

a. Gaussian-like membership function

$$\mu_k^i(x_i) = e^{-\frac{1}{2}(x_i - \varphi_k^i)^2/\sigma_i^2} \tag{4.3}$$

where φ^i_k and σ_i are respectively the mean and the standard deviation of the i^{th} feature values based on the samples belonging to the class C_k. Therefore, the resulted prototype of class C_k is the mean vector of dimension m noted $\varphi_k=[\varphi^1_k, \varphi^2_k,..., \varphi^m_k]$. In case of a too small number of samples provided in real applications, the standard deviation vector $\sigma=[\sigma_1, \sigma_i,..., \sigma_m]$ may be estimated over all the training samples.

b. Binomial membership function

$$\mu_k^i(x_i) = \varphi_k^{i\,x_i}\left(1 - \varphi_k^i\right)^{1-x_i} \tag{4.4}$$

where φ^i_k is the mean of the i^{th} feature values based on the samples belonging to the class C_k.

c. Centered binomial membership function

$$\mu^i_k(x_i|\vartheta^i_k,\varphi^i_k)=\varphi_k^{i\,1-|x_i-\vartheta^i_k|}\left(1-\varphi^i_k\right)^{|x_i-\vartheta^i_k|} \tag{4.5}$$

where ϑ_k is a prototype for class C_k, and parameter φ_k measures the proximity to the prototype so that $\forall x_i \neq \vartheta^i_k : \mu^i_k(\vartheta^i_k|\vartheta^i_k,\varphi^i_k) \geq \mu^i_k(x_i|\vartheta^i_k,\varphi^i_k)$ and for $\varphi_1 \leq \varphi_2$ $\forall x_i \neq \vartheta^i_k$ we have the ordered memberships $\mu^i_k(x_i|\vartheta^i_k,\varphi_2) \geq \mu^i_k(x_i|\vartheta^i_k,\varphi_1)$.

An example of resulting fuzzy partition for quantitative features using Gaussian-like membership function is illustrated in the following example.

Example: If we consider the Gaussian-like membership function (4.3) for the quantitative feature x_1, the obtained parameters of membership functions with respect to the two classes after normalization are $\varphi^1_1 = 0.1786$, $\varphi^1_2 = 0.7979$ and $\sigma = 0.1123$.

4.3.2 Interval type features

To take into account the various uncertainties (noises) and/or to reduce large datasets, the interval representation of data has seen widespread use in recent years (Billard, 2008). In this work, a fuzzy similarity measure is proposed to handle this type of features in such a way that their symbolic nature is preserved.

The membership function for interval type features is taken as the similarity between the symbolic interval value of the i^{th} feature x_i and the interval $\rho^i_k = \left[\rho^{i-}_k, \rho^{i+}_k\right]$ representing class C_k as:

$$\mu^i_k(x_i) = S(x_i, \rho^i_k) \tag{4.6}$$

Symbolic interval features are extensions of pure real data types, in the way that each feature may take an interval of values instead of a single value (Gowda and Diday, 1992). In this framework, the value of a quantity x is expressed as a closed interval $[x^-,x^+]$ whenever only an incomplete knowledge is available about it; representing the knowledge that $x^- \leq x \leq x^+$ (Kuipers, 1994).

Definition 4.1: Let us consider a universe of discourse V as a compact subset of the real line R; that can be continuous or discrete. Any fuzzy subset will be defined as $X = \mu_X(\xi); \xi \in V \subset R$.

We denote *measure* ϖ of a fuzzy set on a discrete universe its scalar cardinal. Here the sigma-count $\varpi[X] = \sum_{\xi_i \in V} \mu_X(\xi_i)$ has been chosen; its extension to a continuous universe is $\varpi[X] = \int_V \mu_X(\xi).d\xi$.

Let us define a fuzzy interval $\hat{A} = \mu_A(\xi)$ as a fuzzy set such that $\forall \xi \notin A; \mu_A(\xi) = 0$, where A is a crisp interval called the *base* of $\hat{A}$. It must be noticed that for a non fuzzy interval X its measure is given by its length $\varpi[X] = \textit{upper .bound}(X) - \textit{lower .bound}(X)$.

Given two crisp intervals $A = [a^-, a^+]$ and $B = [b^-, b^+]$ let us define their *distance* ∂ as:

$$\partial[A, B] = \max\left[0, \left(\max\{a^-, b^-\} - \min\{a^+, b^+\}\right)\right] \tag{4.7}$$

Then the definition of the *similarity measure* between two fuzzy intervals $\hat{A}$ and $\hat{B}$ is given in the discrete case by:

$$S(\hat{A}, \hat{B}) = \frac{1}{2}\left(\frac{\sum_V \mu_{\hat{A}\cap\hat{B}}(\xi_i)}{\sum_V \mu_{\hat{A}\cup\hat{B}}(\xi_i)} + 1 - \frac{\partial[A, B]}{\varpi[V]}\right) \tag{4.8}$$

and its extension to the continuous case is:

$$S(\hat{A}, \hat{B}) = \frac{1}{2}\left(\frac{\int_V \mu_{\hat{A}\cap\hat{B}}(\xi_i).d\xi}{\int_V \mu_{\hat{A}\cup\hat{B}}(\xi_i).d\xi} + 1 - \frac{\partial[A, B]}{\varpi[V]}\right) \tag{4.9}$$

This similarity measure combines two terms. The first term corresponds to the well known Jaccard's similarity measure (Jaccard, 1908) which computes the similarity when the intervals are overlapped; We add to it the second term which allows to take into account the similarity when the intervals are not overlapped.

It shall be remarked that if only crisp intervals are considered this similarity measure can be written as given in (Hedjazi *et al.*, 2011b):

$$S(A, B) = \frac{1}{2}\left(\frac{\varpi[A \cap B]}{\varpi[A \cup B]} + 1 - \frac{\partial[A, B]}{\varpi[V]}\right) \tag{4.10}$$

For the learning step, let us consider a class C_k having N_k samples. The parameters that characterize this class, for the interval type features, are estimated based on an appropriate learning procedure such that the class is represented by a vector whose components are

intervals. The bounds of the interval of each component of this vector are given by the following arithmetic means:

$$\rho_k^{i^-} = \frac{1}{N_k}\sum_{j=1}^{N_k} x_i^{(j)^-}, \text{ and } \quad \rho_k^{i^+} = \frac{1}{N_k}\sum_{j=1}^{N_k} x_i^{(j)^+} \tag{4.11}$$

Where $x_i^{(j)^-}$ is the i^{th} feature lower bound of the j^{th} sample and $x_i^{(j)^+}$ is its upper bound. Therefore, the class C_k is then represented by the interval $\rho_k^i = \left[\rho_k^{i^-}, \rho_k^{i^+}\right]$, and its similarity to the i^{th} interval feature value of the n^{th} sample is given by $S[x_i^{(n)}, \rho_k^i]$ according to formulas (4.8), (4.9) or (4.10).

Consequently, the resulted class prototype for the r interval features is given by the vector of intervals $\rho_k = [\rho_k^1, \rho_k^2, ..., \rho_k^r]^T$.

For a better conditioning of magnitudes and processing time minimization, a normalization within the interval [0,1] is also performed:

$$x_i^- = \frac{\hat{x}_i^- - \hat{x}_{i\min}^-}{\hat{x}_{i\max}^+ - \hat{x}_{i\min}^-}, \quad x_i^+ = \frac{\hat{x}_i^+ - \hat{x}_{i\min}^-}{\hat{x}_{i\max}^+ - \hat{x}_{i\min}^-} \tag{4.12}$$

Where $x_i = [x_i^-, x_i^+]$ is the normalized interval value of $\hat{x}_i = [\hat{x}_i^-, \hat{x}_i^+]$. Consequently, the domain V^i becomes the unit interval [0, 1] for all features. This normalization does not introduce distortion on behalf of the linearity of the normalizing transform.

It is worthwhile here to note that the function S (A, B) fulfills the properties commonly used to characterize a similarity measure :

i. $0 \leq S(A,B) \leq 1$;
ii. S(A,B) = 1 if and only if A equals to B ;
iii. S(A,B) = S(B,A).

Example: If we consider the interval feature x_2 in the set of patterns described in Table 4.1, the resulting parameters of classes for the interval feature after normalization are: $\rho_1^2 = [0.3158, 0.3684]$ and $\rho_2^2 = [0.6316, 0.6842]$.

4.3.3 Qualitative type features

In the qualitative case, the possible values of the i^{th} feature form a set of modalities:

$$D_i = \left\{Q_1^i, \ldots Q_j^i, \ldots Q_{Mi}^i\right\} \tag{4.13}$$

Frequency is a quantity that has been used for measuring fuzzy set membership in several fuzzy applications (Dubois and Prade, 1997). Let Φ_{kj}^{i} be the frequency of modality Q_j^i for class C_k. The membership function of qualitative feature x_i can be specified as:

$$\mu_k^i(\chi_i) = \left(\Phi_{k1}^i\right)^{q_{i1}} * \cdots * \left(\Phi_{kMi}^i\right)^{q_{iMi}} \tag{4.14}$$

where: $$q_j^i = \begin{cases} 1 \;\; if \;\; x_i = Q_j^i \\ 0 \;\; if \;\; x_i \neq Q_j^i \end{cases}$$

Obviously the class parameters are represented by $\Omega_k^i = \left[\Phi_{k1}^i, \ldots, \Phi_{kj}^i, \ldots, \Phi_{kMi}^i\right]$ and the resulting fuzzy partition of the qualitative feature case is described by the following example.

Example: The resulting parameters of fuzzy partitions for the qualitative feature x_3 in the set of samples given in Table 4.1 are: $\Omega_1^3 = [0.2857, 0.4286, 0.2857, 0]$ and $\Omega_2^3 = [0, 0.2857, 0.5714, 0.1429]$.

4.4 Common membership space

A consequence of the fuzzy partition described previously is the mapping of different types of features from completely heterogeneous spaces into a common space which is the membership space. Thus, having a *v*-dimensional quantitative space, a *q*-dimensional qualitative space and an *r*-dimensional interval space, the resulting membership space with the respect each class is *m*-diemnsional (R^m) with $m = v+q+r$ which is the total number of features. In case of dichotomy problems, only one R^m space is necessary as the other can be obtained by complementary of membership.

Definition 4.2: Membership Degree Vector

A Membership Degree Vector (MDV) of dimension *m*, can be associated for a given pattern $x^{(n)}$ to each class as follows:

$$U_{nc_k} = \left[\mu_k^1(x_1^{(n)}), \mu_k^2(x_2^{(n)}), \ldots, \mu_k^m(x_m^{(n)})\right]^T \;\; ; k = 1,2,\ldots,l \tag{4.15}$$

Where $\mu_k^i(x_i^{(n)})$ (i.e. $\mu_k^i(x_i = x_i^{(n)})$) is the membership function of class C_k evaluated at the given value $x_i^{(n)}$ of the i^{th} feature of pattern $x^{(n)}$.

If we consider the previous example, using the definition 3 two MDVs are obtained for the fifth pattern $x^{(5)}$ with respect to its class C_1 and alternative class C_2 as follows:

$$U_{5c_k} = \left[\mu_k^1(x_1^{(5)} = 6), \mu_k^2(x_2^{(5)} = [9.5,10.5]), \mu_k^3(x_3^{(5)} = Yellow)\right]^T ; k = 1,2.$$

which yields

$$U_{5c_1} = [0.6370 , 0.4703 , 0.2857]^T \ \textit{and} \quad U_{5c_2} = [0.3000 , 0.4208 , 0.0000]^T$$

MDV is a m^{th} dimensional image of pattern x_n with respect to the considered class. All the components of the MDV are positive numbers in the unit interval [0,1], therefore U_{nc_k} can be considered as a discrete fuzzy subset and the function $\psi(U_{nc_k}) = \sum_i \mu_k^i(x_i^{(n)})$ represents its scalar cardinality (power or sigma count) as defined in (Zwick *et al.*, 1987) and (Wygralak, 2000). Once all features are simultaneously mapped into a common space, they can be henceforth processed similarly either for classification, feature selection or clustering. We show by next chapters the usefulness of the SMSP principle to perform those data analysis tasks.

4.5 Conclusion

In this chapter a unified principle is introduced to cope with the problem of data heterogeneity. This principle is based on a simultaneous mapping of data from initially heterogeneous spaces into only one homogeneous space using appropriate characteristic functions. Once the heterogeneous data are represented in a unified space, only a single processing for various analysis purposes such as machine learning tasks can be performed. We considered here the three most used types of features which are quantitative, interval and qualitative.

In the present work, this principle is proposed within the fuzzy set theory framework to reason about heterogeneous data. Once suitable membership functions that characterize the adequacy (typicality, appropriateness) of a pattern to each class are chosen according to feature types, a fuzzy partition of features can be performed based on empirical data. In the present work two well-known semantics (similarity and uncertainty semantics) have been adopted to define an adequate membership function according to the feature type. The first one has been used for quantitative and interval data whereas the later one has been adopted for the qualitative data. We take advantage of the commensurability assumption in the framework of fuzzy logic to end up with a unique space (unit hypercube) where all the features are represented by their memberships to a reference fuzzy partition. We show by next chapters that by employing this principle within a fuzzy logic framework, only a simple fuzzy reasoning mechanism can be used to perform several machine learning tasks such as classification, feature selection and clustering.

CHPITRE 5- Résumé
Apprentissage supervisé basé sur le principe «SMSP»

Il est reconnu dans la pratique que la plupart des connaissances médicales employées pour la prise de décision sont généralement exprimées sous la forme de règles qualitatives. Ceci est principalement la raison qui rend les systèmes à base de règles bien acceptés par les praticiens. Les systèmes flous à base de règles peuvent être particulièrement d'un grand intérêt car ils offrent une grande transparence et interprétabilité tout en permettant de traiter des informations bruitées, imprécises ou incomplètes présentes souvent dans de nombreux problèmes du monde réel.

La relation entre le résultat de la classification et la variable originelle est généralement non linéaire et complexe. Cependant, si la variable originelle est correctement «fuzzifiée», la relation peut être approchée par une fonction linéaire et un classifieur simple peut être utilisé (Li et Wu, 2008). Récemment, des systèmes basés sur des règles floues (Si-Alors) ont été appliqués à des problèmes de classification où les vecteurs de données non-floues (ou numériques) d'entrée doivent être attribuées à l'une des classes existantes (Ishibuchi et al, 1992; Chiu, 1997; Abe et Thawonmas, 1997). Toutefois, cette classe de classifieurs devient inutilisable dès qu'un problème de dimension élevée et/ou présentant une hétérogénéité des données est rencontré. Ce cas est fréquent dans les applications du cancer qui représente notre sujet d'intérêt. Nous montrons tout d'abord dans ce chapitre qu'un simple classifieur basé sur des règles floues peut être conçu selon le principe SMSP introduit dans le chapitre précédent pour faire face à l'hétérogénéité des données. Ensuite, en se basant toujours sur le même principe, une approche de pondération de variables est conçue et intégrée dans le classifieur flou dans le but de l'adapter à des problèmes de dimension élevée.

Dans ce travail, chaque ensemble flou de la prémisse de chaque règle floue (Si-Alors) est pondéré afin de caractériser l'importance de chaque proposition et donc de la variable correspondante. Pour justifier une telle opération, le processus d'estimation du poids est basé sur la maximisation des marges d'appartenance afin d'estimer un poids flou de chaque variable dans l'espace d'appartenance. Il est montré aussi que la définition de la fonction objective en se basant sur le concept de marge peut réduire efficacement la complexité de

calcul grâce à l'utilisation de techniques d'optimisation standards, qui permettent d'éviter une recherche heuristique combinatoire. Une extension de la méthode pour traiter les problèmes multi-classes est aussi proposée. Une étude expérimentale extensive a été menée pour démontrer l'efficacité de la méthode proposée sur deux ensembles de bases de données. Le premier est caractérisé par l'hétérogénéité des données et le deuxième par la dimension élevée. Cette méthode a été comparée avec des méthodes de pondération de variables bien connues dans la littérature: Relief (Kira and Rendell, 1992a), I-Relief (Sun, 2007b) and Simba (Gilad-Bachrach et al., 2004). Afin d'assurer une comparaison sans biais, les deux classifieurs populaires k-NN (Cover and Hart, 1967) et SVM (Vapnik, 1998) ont été aussi utilisés en plus du classifier flou que nous proposons. Les résultats obtenus montrent que la méthode proposée apporte des améliorations significatives combinée avec le classifieur flou. En particulier, nous avons observé que l'approche par pondération floue proposée améliore significativement les performances du classifieur sur presque l'ensemble des bases de données hétérogènes. Un gain significatif de performance est obtenu en ne conservant que quelques variables plutôt que l'ensemble des variables originelles. Par exemple, près de 5% de gain de performance est réalisé en utilisant uniquement les quatre premières variables au lieu des neuf variables originelles du jeu de données de Ljubljana sur le pronostic du cancer du sein. Il a été constaté aussi que la méthode de pondération floue fournit des résultats comparables ou même meilleurs que les autres méthodes de pondération classiques sur presque toutes les bases de données hétérogènes en utilisant les deux autres classifieurs (k-NN et SVM). Pour une comparaison plus rigoureuse entre les trois méthodes de sélection de variables, une analyse statistique a été aussi effectuée. En ce qui concerne les expériences sur le deuxième ensemble de bases de données caractérisé par la présence d'un nombre important de variables non pertinentes, les résultats fournis par la méthode proposée sont encourageants. En particulier il a été constaté que, bien que cette méthode possède une complexité numérique faible, elle permet de réduire significativement le nombre de gènes nécessaires pour effectuer les tâches de diagnostic et/ou pronostic. A titre d'exemple, sur la base de données du cancer de la prostate caractérisée par la mesure de l'expression de 10509 gènes, la méthode proposée surpasse les autres méthodes de pondération en fournissant une erreur de classification minimale de 5% pour juste 10 gènes en utilisant le classifieur flou. Ce résultat suggère que l'utilisation des 10 gènes sélectionnés au lieu de l'ensemble des 10509 gènes permet d'atteindre la performance de classification maximale en vue de pronostiquer ce type de cancer.

CHAPTER 5
Supervised learning based on SMSP principle

It is recognized in medical practice that most of physicians' knowledge employed for decision are usually expressed in the form of rules. This is mainly the reason that makes the rule-based systems very well accepted by medical practitioners. Fuzzy-rule based systems can be particularly of big interest as they offer a high transparency and comprehensive interpretability while they allow dealing with noisy, imprecise or incomplete information often present in many real world problems. They provide indeed a good trade-off between the empirical precision of traditional engineering techniques and its high interpretability. Fuzzy-rule based systems have been widely used in control problems (Lee, 1990; Sugeno, 1997). From this point of view, fuzzy logic can be seen as more appropriate rather than other classical methods which fail when the system model is highly dimensional and non linear. This is mainly due to its attractive properties that enable to handle imprecise and noisy data.

Usually the relationship between the result of classification and the original feature is nonlinear and complicated. However, if the original feature is appropriately fuzzified, the relationship may be approximated by a linear function and a simple classifier may be used (Li and Wu, 2008). Recently, fuzzy rule based systems have often been applied to classification problems where nonfuzzy (or numerical) input vectors have to be assigned to one of the given set of classes (Ishibuchi *et al.*, 1992; Chiu, 1997; Abe and Thawonmas, 1997). However, this class of classifiers becomes impracticable whenever high dimensional and/or heterogeneous problems have to be faced. This case is common to occur in cancer applications that are our subject of interest. Traditional fuzzy classifiers are commonly based on arbitrary choice to determine the number of linguistic terms of the fuzzified features, which is not always possible and accurate enough whenever a huge number of features is encountered. We show firstly in this chapter that a simple fuzzy rule based-classifier can be designed based on the previously introduced SMSP principle to deal with data heterogeneity. Then, based on the same principle, a feature weighting approach is designed and integrated into the fuzzy rule-based classifier in the aim to make it scalable with high dimensional problems. Indeed, weighting fuzzy if-then rules to improve classification performance is a common practice in fuzzy rule-based classifier systems (Ishibuchi and Nakashima, 2001; Ishibuchi and

Yamamoto, 2005; Jahromi and Taheri, 2008; Sanz *et al.*, 2010). However, this weighting aims usually to characterize the importance of each fuzzy rule by a scalar weight. For example, Ishibuchi and Yamamoto (2005) have proposed a heuristic automatic way to estimate the rule weights based on sample membership to each class in a supervised context. In the present work, each antecedent fuzzy set in the fuzzy if-then rule is weighted to characterize the importance of each proposition and therefore of the corresponding feature (Hedjazi *et al.*, 2011c). To justify such an operation, weight estimation process is based on membership margin maximization to estimate a fuzzy weight of each feature in the membership space. As it will be shown, the margin concept can efficiently decrease the computation complexity through the use of standard optimization techniques avoiding combinatorial search. Experiments on high and low dimensional datasets are performed in order to demonstrate that the proposed approach can improve significantly the performance of fuzzy rule-based as well as state-of-the-art classifiers and can even outperform classical feature selection approaches.

We start first by describing the fuzzy-rule based classifier for mixed-type data and then we describe the weight integration process into the classifier.

5.1 Fuzzy rule-based classifier for mixed-type data

In this section, we illustrate the problem of heterogeneous data classification as a reasoning problem in a common space based on the SMSP principle. Indeed, once the different types of features have been mapped into a common space it is possible to establish a unified reasoning scheme for a classification purpose. This approach is based on using the fuzzy partitions, resulted from the mapping described in chapter 4, to establish a fuzzy inference engine.

We describe by next the fuzzy-rule based classifier for mixed type data. We consider the following type of fuzzy if-then rules for *m*-dimensional problem:

R_k: If x_1 is A_1 and x_2 is A_2...and x_m is A_m then x belongs to class C_k
where the antecedent fuzzy sets A_i correspond to membership functions $\mu_k^i(x_i)$ for each class C_k defined in section 4.3 according to the type of i^{th} feature. It must be noticed here that the set of features used to evaluate each fuzzy if-then rule can possibly be of mixed types (quantitative, qualitative or interval-valued).
Then, the truth value of the consequent of each rule is determined by a fuzzy logic implication function which consists in a linear interpolation between a (t-norm) and a (t-conorm). Finally,

the sample is assigned to the class corresponding to the maximum membership obtained using the following fuzzy inference engine:

$$R^* = \underset{R_k}{Arg\max} \left\{ \alpha \left| \mu_k^1(x_1),..., \mu_k^m(x_1) \right| + (1-\alpha)\beta \left| \mu_k^1(x_1),..., \mu_k^m(x_1) \right| / k = 1,..., l \right\}$$

where γ and β are dual fuzzy aggregation functions T-norm and its dual T-conorm that combine memberships (given by the components of the MDV U_{nc_k}) of features values of a sample $x^{(n)} = [x_1^{(n)}, x_2^{(n)},..., x_m^{(n)}]$ to a class C_k (Piera and Aguilar, 1991). The parameter α, called *exigency* allows to adjust the compensation between the union and the intersection operators which can be pre-specified by the user or estimated through a cross-validation using training data.

Without the unification of the space of features, this simple inference mechanism could not be applied, and the influence of the different types of features would not be equal. Such a classifier is referred to here as LAMDA (Aguilar and Lopez De Mantaras, 1982; Isaza *et al.*, 2004; Hedjazi *et al.*, 2009; Hedjazi *et al.*, 2010b).

5.2 Weighted fuzzy rule-based classifier for mixed-type data

For many learning domains potential useful features, for sample description are defined randomly. Nevertheless, not all of the features have equal importance for classification task, some of them can be irrelevant and can even hurt classification performance. We describe in this section how a feature weighting process can be easily integrated in the previously described fuzzy-rule based classifier, through a weighted fuzzy rule concept in the aim to improve its performance. The concept of fuzzy weighted rule introduced here consists of weighting each proposition of the fuzzy rule to characterize the importance of each feature.

Definition 5.1: Weighted Fuzzy If-Then Rules (WFR)

A weighted If-then rule is similar to a conventional rule with the exception that a weight is assigned to each antecedent proposition. A WFR is defined as:

R: IF a THEN c, w_f ,

Where $a=\{a_1,a_2,\ldots,a_m\}$ is the antecedent portion which is composed of one or more propositions connected by “AND” or ”OR”. Each proposition a_i ($1 \leq i \leq m$) can have the format “ x_i is F_i”, where F_i is a fuzzy set corresponding to the type of the i^{th} feature established in the learning step. The feature value x_i can be quantitative, qualitative or

interval-valued type. Whereas, $w_f=\{w_{f1}, w_{f2}, \ldots, w_{fm}\}$ is a weight vector. The weight w_{fi} of a proposition a_i shows the degree of importance of a_i to contribute to the consequence c and therefore the importance of the i^{th} feature value x_i to the classification task.

Thus the classification rule becomes:

R_k: If x_1 is (w_{f1}) A_1 and x_2 is (w_{f2}) A_2 ...and x_m is (w_{fm}) A_m then x belongs to class C_k
where the antecedent fuzzy sets A_i correspond to the established fuzzy set in the universe of discourse of the i^{th} feature.
The remaining issue is to evaluate appropriately the weights of each feature, taking into account that they will be used to modify the membership to antecedent fuzzy sets of each classification rule. A natural idea is to estimate these weights in the membership space based on SMSP principle to justify such an operation.

5.3 Membership margin

In classical feature weighting methods, the feature relevance is estimated in a space assumed to be quantitative. This requires that other feature types must be transformed arbitrarily, without taking any consideration about their original space. While, based on SMSP principle, an appropriate mapping of different features into a common space is achieved; this allows bypassing the assumption of pure quantitative features. In recent machine learning theory, margin concept plays an important role to estimate the decision making confidence (Vapnik, 1998). In the following, we define a Membership Margin which enables to estimate the features weight in the membership space whatever their type and number.

Definition 5.2: *Membership Margin (MM).*
Let us consider class $c=C_1$, and its complement $\tilde{c}=C_2$. We assume that the n^{th} data sample $x^{(n)}=[x_1^{(n)}, x_2^{(n)}, \ldots, x_m^{(n)}]$ is labeled by class c. Let's define the membership margin for sample $x^{(n)}$ by:

$$\beta_n = \psi(U_{nc}) - \psi(U_{n\tilde{c}}) \tag{5.1}$$

Where U_{nc} and $U_{n\tilde{c}}$ are respectively the membership degree vectors of sample $x^{(n)}$ to classes c and $\tilde{c}$, computed with respect to all samples contained in D excluding $x^{(n)}$ ("leave-one-out margin") and ψ is an aggregation function. We define here $\Psi(Y)=\sum_i Y_i$, which can be extended to any other aggregation function.
Thus, in our case the function ψ is given as follows:

$$\psi(U_{nk}) = \sum_i \mu_k^i(x_i^{(n)}) \tag{5.2}$$

The feature membership can be seen as the *contribution* (relevance) of this feature to the rule's consequence of a given class. Consequently, when an assignment decision is necessary, if the contribution average of all features for the sample $x^{(n)}$ to its class is greater than its average contribution to the alternative class it is clear to assign it to the class with maximum contribution (which corresponds to its correct class). Therefore, sample $x^{(n)}$ is considered correctly classified if $\beta_n > 0$.

The arithmetic sum given by (5.2) defines a compromise aggregation between membership functions and lies in equal way between union and intersection (Dubois et Prade, 1988). On the other hand, if the MDV U_{nc_k} is considered as a discrete fuzzy subset, function ψ represents the scalar cardinality (power or sigma count) of U_{nc_k} as defined by (Zwick *et al.*, 1987; Wygralak, 2000). Therefore, the membership margin β_n given by (5.1) is the scalar cardinalities difference of these resulted fuzzy subsets.

Intuitive interpretation: This membership margin is a measure of how much the features memberships can be modified in the membership space before a sample $x^{(n)}$ being misclassified. According to the margin types described in (Grammer *et al.*, 2002), this margin can be also considered as an hypothesis-margin. Note that the membership margin is affected by the selected subset of features through the function ψ. It is worthwhile to note that our feature weighting approach is also based implicitly on maximum membership rule to label an pattern by an existing class. Membership margin for pattern $x^{(n)} \in c$ is based on the aggregation $\Psi(U_{nc})$ computing its global membership to the class *c*. Of course, other alternatives can be opted also using different types of aggregation functions. We stated previously that $x^{(n)}$ is considered correctly classified if $\beta_n > 0$. This is equivalent to write:

$$C_n = \underset{C,\tilde{C}}{\arg\max}\{\psi(U_{nc}), \psi(U_{n\tilde{c}})\} \tag{5.3}$$

which is equivalent to the maximum membership rule, that $x^{(n)}$ belongs to the class with maximum global membership. Therefore, this approach encompasses implicitly the decision process in the feature selection task.

5.4 MEmbership Margin Based feAture Selection: MEMBAS

Similarly to the classification task, since all features are simultaneously mapped into a common space thanks to SMSP principle, they can be henceforth processed in unified way for

feature weighting task. In the case of the compromise aggregation via the arithmetic sum described by (5.2), importance assignment is easily incorporated in the aggregation through a weighted sum (Dubois et Prade, 1988; Cross and Sudkamp, 2002).

Definition 5.3: *Fuzzy Feature Weight. FFW* is defined as the relative degree of usefulness of each feature in the membership space for the discrimination between two classes. These fuzzy feature weights are non-negative numbers expressing the discriminative power of the fuzzy sets between existing classes. It results that *FFW* is a vector, referred to as $W_f = \left[w_{f1}, \ldots w_{fm} \right] \in R^m$, assigned in the membership space, where the term 'fuzzy weight' comes from.

Definition 5.4: Weighted adequacy of a pattern.

Given a vector of positive fuzzy feature weights $W_f = \left[w_{f1}, \ldots w_{fm} \right] \in R^m$, the weighted adequacy of the n^{th} pattern is defined by the cardinality of the new fuzzy set that takes into account the weight of each feature in the membership space. It is given by the scalar product:

$$\psi(U_{nk}/w_f) = \sum_i w_f^T \, U_{nk} = \sum_i w_{fi} \, \mu_k^i(x_i^{(n)}) \tag{5.4}$$

5.4.1 Fuzzy Feature Weight Estimation

The basic idea to calculate the fuzzy feature weights is to scale feature memberships in the membership space by minimizing the leave-one-out error. Therefore, the margin given by (5.1) in the weighted membership space becomes:

$$\beta_n(w_f) = \psi\left(U_{nc}/w_f\right) - \psi\left(U_{n\bar{c}}/w_f\right) \tag{5.5}$$

However, the problem which remains is to find a procedure to estimate the weight vector w_f. One approach among others would be to take advantage of the membership margin definition (5.5) to define a margin-based objective function and then reformulate this problem as an optimization problem in the membership space as it is usually performed in the large margin theory framework.

a) Problem statement

It has been proved recently, within the margin theory framework (Vapnik, 1998), that a classifier based on minimizing a margin-based error function generalizes well on unseen test data. For this reason, it has also been extended for feature selection purposes (Weston *et al.*, 2001; Freund and Schapire, 1997). The present work takes its originality in the use of the

membership margin concept. To solve the above described problem, one can transform it to the following optimization problem in the feature membership space:

$$\underset{w_f}{\text{Min}} \quad \sum_{n=1}^{N} h\left(\beta_n\left(w_f\right) < 0\right) \tag{5.6}$$

Where $\beta_n(w_f)$ is the $x^{(n)}$ margin computed with respect to w_f and h is an indicator function. To solve the above problem, we define an objective function so that the averaged membership margin in the resulted weighted feature membership space is maximized:

$$\underset{w_f}{\text{Max}} \sum_{n=1}^{N} \beta_n(w_f) = \sum_{n=1}^{N} \left\{ \sum_{i=1}^{m} w_{fi} \mu_c^i(x_i^{(n)}) - \sum_{i=1}^{m} w_{fi} \mu_{\tilde{c}}^i(x_i^{(n)}) \right\}$$

Subject to the following constraints : (5.7)

1. $\| w_f \|_2^2 = 1$,
2. $w_f \geq 0$.

The first constraint is the normalized bound for the modulus of w_f so that the maximization ends up with non infinite values, whereas the second guarantees the nonnegative property of the obtained weight vector. Then (5.7) can be simplified as:

$$\begin{aligned} &\underset{w_f}{\text{Max}} \ (w_f)^T s \\ &\text{Subject to} \| w_f \|_2^2 = 1, w_f \geq 0 \end{aligned} \tag{5.8}$$

where

$$s = \sum_{n=1}^{N} \left\{ U_{nc} - U_{n\tilde{c}} \right\} \tag{5.9}$$

In the statement of this optimization problem, we must assume that there exists at least one feature $i \leq m$, such that $s_i > 0$.

b) Lagrangian optimization approach

This is a classical optimization problem stated in the framework of Lagrange multipliers (see Appendix 3). Therefore, taking in advantage that it provides an analytical solution, we get finally a closed form for w_f:

$$w_f^* = \frac{s^+}{\| s^+ \|} \tag{5.10}$$

with $s^+ = [\max(s_1,0), \ldots, \max(s_m,0)]^T$

5.4.2 MEMBAS Algorithm

We present bellow the algorithm of the proposed approach. We consider here the online learning version of Membas rather than the batch one due to its attractive properties. Although both approaches are equivalent in terms of the final result, it is known that an online algorithm

is computationally more efficient than its batch version when the amount of training data is large. Moreover, it enables also to update weights by the information brought by a new sample which was not available when starting the training. The computational complexity of Membas is O (*N m*), where *N* is the sample size and *m* the data dimensionality, and it can be summarized by the following algorithm:

1. *Initiate the fuzzy weight vector to zero, T number of iterations (T=N when the training is performed over all patterns).*
2. *For t=1…T*
   a) *Select randomly a sample* $\mathrm{x}^{(n)}$ *from D*
   b) *Determine the fuzzy partition of each feature according to its type with respect to* $D\backslash\{\mathrm{x}^{(n)}\}$.
   c) *Calculate the membership degree vectors MDVs* U_{nc} *and* $\mathrm{U}_{\mathrm{n}\tilde{c}}$ *for sample* $\mathrm{x}^{(n)}$.
   d) *Update vector* s *as*

$$\mathrm{s} = \mathrm{s} + \{\mathrm{U}_{\mathrm{nc}} - \mathrm{U}_{\mathrm{n}\tilde{c}}\}$$

3. *Calculate the optimal fuzzy weight vector as*

$$\mathrm{w}_{\mathrm{f}}^{*} = \frac{\mathrm{s}^{+}}{\|\mathrm{s}^{+}\|},$$

with $\mathrm{s}^{+} = [max(s_1,0), \ldots, max(s_m,0)]^T$

5.4.3 MEMBAS for multiclass problems

The extension of the MEMBAS method for multiclass problems is considered in this section. Once the membership function parameters of each class have been determined from the training dataset, the feature space is partitioned into a number of fuzzy sets equal to the number of classes. Consequently an equal number of membership degree vectors is resulted. Therefore, a similar margin definition for multiclass problems to the one given in (Sun, 2007b) can be used for the same purpose, by taking the maximum marginal membership with respect to all classes other than class *c*:

$$\beta_n = \min_{\{\tilde{c}\in C, \tilde{c}\neq C(\mathrm{x}_n)\}} \{\psi(\mathrm{U}_{\mathrm{nc}}) - \psi(\mathrm{U}_{\mathrm{n}\tilde{c}})\} \tag{5.11}$$

Thus (5.9) becomes

$$\mathrm{s} = \sum_{n=1}^{N} \min_{\{\tilde{c}\in C, \tilde{c}\neq C(\mathrm{x}_n)\}} \{\mathrm{U}_{\mathrm{nc}} - \mathrm{U}_{\mathrm{n}\tilde{c}}\} \tag{5.12}$$

By using this last expression of s and following the same steps we arrive to a form of $\mathrm{w}_{\mathrm{f}}^{*}$ similar to the one obtained in (5.10).

Finally, the estimated w_f^* in the membership space can be used now to weight each proposition of the fuzzy rule-based classifier with objective of improving its performance. Nevertheless, albeit a direct feature-weight assignment to each proposition can be very useful to improve the performance of the fuzzy rule-based classifier for relatively low dimensional problems, it becomes undesirable for high dimensional problems, such as bioinformatics problems characterized by thousands of features. However, feature-weight assignment process can be regarded as a generalization of feature selection (Wang *et al.*, 2004). In the present work we focus on feature selection problem rather than a direct feature-weight assignment. This is equivalent to activate or deactivate the proposition corresponding to each feature in the fuzzy if-then rule according whether it was deemed important or not by Membas. From other side, we formulated the weight computation procedure in the way that the proposed approach approximates the leave-one-out cross validation error. Therefore, this approach chooses features only if they contribute to the overall classification performance regardless of their redundancy or correlation. It is reported that often redundant features can deteriorate classification performance and removing them is necessary. However, it has been pointed out recently in some applications such as DNA microarray, that the ultimate goal is not always the identification of a small gene subset with good predictive power, but to help the physicians to have a good insight about the relationship between genes and certain diseases (Jenssen and Hovig, 2005). Discovering redundant (or coregulated) genes may provide some useful information about their interactions.

5.5 Experiments and Comparisons

In the present section, we show how the proposed method can improve the performance of the fuzzy rule-based classifiers as well as other well known state-of-the-art classifiers on some real-world problems. To further demonstrate its effectiveness, several comparisons have been performed: Membas versus three well-known feature selection approaches using three different classifiers to avoid biased comparison. They concern experiments on low-dimensional datasets (less than 50 features) and high-dimensional datasets (more than 1000 features).

5.5.1 Feature selection methods

For comparison purposes we used three methods: Relief (Kira and Rendell, 1992a), I-Relief (Sun, 2007b) and Simba (Gilad-Bachrach *et al.*, 2004), widely used for the validation of

newly proposed feature selection approaches. We give below a brief description of these three methods.

Relief is considered as one of the most successful feature selection methods due to its simplicity and effectiveness (Dietterich, 1997). In Relief the feature weights are estimated iteratively according to their discrimination ability based on samples neighbouring. For each iteration, a sample is selected randomly and its two nearest neighbours are found: one from the same class (nearest hit) and the other from the alternative class (nearest miss). An extension of Relief to multiclass problems has been presented in (Kononenko, 1994). Moreover, it has been proven recently that Relief is not an heuristic filter method as it has been long time considered but an online learning algorithm that solves an optimization problem (Sun, 2007b). In the same work, (Sun, 2007b) have proposed an efficient iterative version of Relief, referred to as I-Relief, by using an Expectation-Maximization algorithm. I-Relief searches the real nearest neighbor in the weighted feature space, unlike Relief which makes the assumption that the nearest neighbor in the original space is the same one in the weighted feature space. Further theoretical convergence analysis of I-Relief and its online version have been also provided to prove its superiority over the Relief family algorithms. I-Relief has one free parameter, the width of kernel Gaussian function, to be defined by the user. This parameter should be selected properly to guarantee the I-Relief convergence.

Concerning the Simba method, a gradient ascent to maximize a margin based evaluation function is performed. Simba also is based on samples neighboring as Relief. At each iteration, for a given randomly selected sample, the feature weight is updated by using the rule obtained by the gradient ascent procedure. I-Relief, Relief and Simba are recognized by machine learning community as efficient wrapper approaches, and widely used in literature to prove the effectiveness of recently proposed feature selection approaches (Dietterich, 1997). The three approaches are distance-based methods that maximize a 1-NN margin. Relief algorithm used here for comparison is also multiclass as proposed and used by (Gilad-Bachrach *et al.*, 2004). However, most of existing probabilistic and information-theoretic based approaches are of filter type (Wettschereck and Aha, 1995; Mitra *et al.*, 2002). It is a recognized fact now within the machine learning community that such filter approaches are computationally more efficient but perform worse than wrapper methods (Kohavi and John, 1997; Guyon and Elisseeff, 2003). Moreover, extensive comparative studies performed in last decades have proved their superiority against filter approaches on wide range of real-world problems. For example, Gilad-Bachrach *et al.* (2004) have compared Relief and Simba with a

mutual-information based approach. Many works can be also found in literature comparing Relief and Simba with information-theoretic and probabilistic approaches (Wettschereck and Aha, 1995; Wettschereck *et al.*, 1997; Robnik-Šikonja and Kononenko, 2003; Li and Lu, 2009;).

5.5.2 Experimental setup

Here, the three feature selection methods are evaluated based on the classification error obtained using their selected feature subset. Besides the fuzzy rule based classifier LAMDA, we used also the well known *k*-NN classifier (Cover and Hart, 1967). Moreover, in order to achieve a more accurate classification performance, they are also compared using SVM classifier (Vapnik, 1998). The *k*-NN method classifies each unlabelled sample by the majority label among its *k* nearest neighbors in the training set (Cover and Hart, 1967). It is known that *k*-NN classifier is very sensitive to the presence of irrelevant features and therefore adequate to compare feature selection methods (see chapter 2). While the Support Vector Machine method finds the separating hyper-plane with the largest sample-margin (Vapnik, 1998). Unlike *k*-NN, it is well known that SVM is very robust against noise, and that the presence of a few irrelevant features in the original feature set should not significantly affect its performance (see chapter 2). Consequently, SVM may perform similarly with the different feature selection methods in this case.

The main reason of using these three different classifiers is to assess whether this approach can, in addition to the fuzzy rule based classifier LAMDA, improve other state of the art classifiers. This comparative study was performed on two dataset collections. The first collection concerns six UCI Repository datasets (Blake and Merz, 1998): Diabetics, Thyroid, WDBC, Ljubljana, Twonorm and Heart. Two datasets (Ljubljana, Heart) among them have mixed feature types (quantitative, qualitative as well as interval). Moreover, 50 independently normal distributed irrelevant features with zero mean and unit variance were added to the original features of all datasets to assess the robustness of the newly proposed method against irrelevant features. The second collection concerns four DNA microarray datasets: DLBCL (Shipp *et al.*, 2002), Lung cancer (Bhattacharjee *et al.*, 2001), prostate cancer (Singh *et al.*, 2002), SRBC (Khan *et al.*, 2001). The main characteristic of these datasets is their high feature dimensionality (several thousands to ten thousands) and the small sample size (ten to one hundred). It must be noted also that some of datasets are multiclass. Additional information about each dataset is given in Table 5.1.

Tab. 5.1. Summary of used Datasets

Dataset	No. Train	No. Test	No. Feature	Class	Task description
Diabetics	615	153	8	2	Diabetes onset forecast
Thyroid	172	43	5	3	Thyroid disease diagnosis
WDBC	456	113	30	2	Breast cancer diagnosis
Ljubljana	222	55	9	2	Breast cancer prognosis
Twonorm	371	7029	20	2	Artificial dataset
Heart	216	54	13	2	Heart disease diagnosis
DLBCL	77	/	7129	2	Outcome prediction of Diffuse Large B-cell
Lung cancer	203	/	12600	5	Diagnosis of four lung cancer types
Prostate cancer	102	/	10509	2	Prostate cancer prognosis
SRBC	83	/	2308	4	Small, Round Blue-Cell tumors diagnosis

In all cases, the classification error was used as the criterion to evaluate the performance of the compared methods. I-Relief have one free parameter to be defined by the user. Sun (2007b) suggested that this parameter should be selected superior to 0.5 in order to guarantee the I-Relief convergence (we set it to 0.7 in the present experiments). As mentioned in the previous section, Simba suffers of local maxima problem because it performs a gradient ascent. To overcome this problem Simba performs a gradient ascent from several initial points predefined by the user. The number of points is set here to the Simba default value, 5 (Gilad-Bachrach *et al.,* 2004). Concerning SVM, the supervised binary classifier was used for binary class problems whereas a multiclass SVM "one against one" is used in the case of multiclass problems (Vapnik, 1998). As the focus of this work is the comparison between the feature selection methods, only a simple linear kernel for binary class problems and polynomial kernel for the multiclass ones have been used.

5.5.3 Experiments on low-dimensional datasets

As mentioned above, firstly the experiments have been performed on the six UCI datasets to compare Membas with Simba, Relief and I-Relief methods. These datasets contain mixed feature-type data, and each classifier has one parameter which has been adjusted through a cross-validation process (i.e. the exigency index assuring a linear interpolation between the fuzzy logic connectives for the fuzzy-rule based classifier LAMDA, the number of nearest neighbours k for the k-NN method and the regularization parameter for SVM). For this purpose each dataset was randomly partitioned into two subsets training and test data as it is detailed in Table 5.1. The three parameters are estimated through a cross validation using the training dataset (70% vs 30%). The optimal parameters values are taken according to the smallest classification error obtained on the remaining 30% of the training subset. Then, the classification error is calculated on the test subset consisting in unseen samples for the three classification methods. To eliminate any statistical variation and make the comparisons

between different methods more balanced, this procedure was repeated 20 times for each dataset. The averaged error over the 20 runs is considered as the classification error for a given feature subset. The averaged testing error for LAMDA, k-NN and SVM methods is plotted as a function of the top ranked features respectively in Figure 5.1, Figure 5.2 and Figure 5.3. Moreover, the optimal obtained averaged classification errors with the three classifiers and the corresponding number of selected features by each feature selection method are reported in Tables 5.2 to 5.4.

Tab. 5.2. Optimal Testing Errors (%) and corresponding number of features on the Ten Data Sets with LAMDA. Last raw (W/T/L) summarizes Win/Tie/Loss in comparing MEMBAS with other approaches based on significance level 0.05.

Dataset	MEMBAS	SIMBA	RELIEF	I-RELIEF	P-value (MEMBAS-RELIEF)	P-value (MEMBAS-SIMBA)	P-value (MEMBAS-IRELIEF)
Diabetics	**25.5 (5)**	27.2 (5)	28.2 (2)	25.6 (3)	**0.00**	**0.00**	0.31
Thyroid	**4.9 (4)**	5.6 (3)	6 (5)	5.3 (5)	0.25	0.89	0.90
WDBC	**5 (21)**	6.7 (30)	7.3 (29)	5 (28)	**0.00**	**0.00**	**0.01**
Ljubljana	**24.6 (4)**	29.9 (8)	34.9 (8)	25.3 (4)	**0.00**	**0.00**	0.09
Twonorm	**2.4 (20)**	4.4 (19)	3.4 (20)	2.5 (20)	0.91	0.80	0.97
Heart	**15.3 (7)**	25.8 (13)	23.2 (13)	28 (11)	**0.00**	**0.00**	**0.00**
DLBCL	5.2 (80)	8.5 (25)	6.5 (60)	**3.4 (300)**	**0.00**	**0.00**	0.27
Lung cancer	**4.4 (70)**	5.5(70)	4.4(100)	7.4 (100)	0.17	0.52	**0.03**
Prostate cancer	**5(10)**	13.3(40)	11(20)	6.8 (50)	**0.00**	**0.00**	0.07
SRBC	**0 (20)**	0(100)	0(160)	0 (60)	0.19	0.53	0.38
					W/T/L= 6/4/0	**W/T/L= 6/4/0**	**W/T/L= 3/7/0**

Tab. 5.3. Optimal Testing Errors (%) and corresponding number of features on the Ten Data Sets with *k*-NN. Last raw (W/T/L) summarizes Win/Tie/Loss in comparing MEMBAS with other approaches based on significance level 0.05.

Dataset	MEMBAS	SIMBA	RELIEF	I-RELIEF	P-value (MEMBAS-RELIEF)	P-value (MEMBAS-SIMBA)	P-value (MEMBAS-IRELIEF)
Diabetics	**24.8 (5)**	26.5 (6)	25.3(8)	26.1 (6)	0.23	0.07	0.11
Thyroid	**4.9 (4)**	6 (3)	5.8(3)	5.8 (3)	0.41	0.72	0.96
WDBC	6.8 (17)	7.3 (9)	6.7 (12)	**6.6 (22)**	**0.02**	**0.00**	**0.02**
Ljubljana	**25.8 (5)**	28.2 (4)	29.2(9)	25.8 (7)	**0.00**	**0.00**	0.55
Twonorm	4.2 (19)	6.1 (18)	5.2(20)	**4 (20)**	0.90	0.80	0.97
Heart	30.3 (5)	35.7 (4)	36.6(12)	**28 (10)**	**0.00**	**0.03**	**0.00**
DLBCL	**4.7 (120)**	6.8 (40)	5.3(90)	5.3 (200)	0.84	0.89	**0.01**
Lung cancer	**5.7 (80)**	8.4 (180)	7.3(200)	6.2 (300)	0.32	0.45	0.09
Prostate cancer	**7.5(15)**	18 (90)	17.8(40)	17.2 (30)	**0.00**	**0.00**	**0.00**
SRBC	**0(40)**	0 (140)	**0(40)**	**0 (40)**	0.43	0.81	0.27
					W/T/L= 3/6/1	**W/T/L= 4/6/0**	**W/T/L= 2/6/2**

Tab. 5.4. Optimal Testing Errors (%) and corresponding number of features on the Ten Data Sets with SVM. Last raw (W/T/L) summarizes Win/Tie/Loss in comparing MEMBAS with other approaches based on significance level 0.05.

Dataset	MEMBAS	SIMBA	RELIEF	I-RELIEF	P-value (MEMBAS-RELIEF)	P-value (MEMBAS-SIMBA)	P-value (MEMBAS-IRELIEF)
Diabetics	**23.8 (5)**	25.7 (8)	24.5 (7)	24.9 (7)	0.20	**0.00**	**0.05**
Thyroid	4 (4)	**4 (3)**	4.4 (4)	**4 (3)**	0.27	0.79	0.94
WDBC	**2.4(22)**	3.9 (20)	5.2 (30)	4.6 (30)	**0.00**	**0.01**	**0.00**
Ljubljana	**27 (4)**	28.5 (7)	29.4 (8)	27.6 (3)	**0.00**	**0.00**	0.10
Twonorm	**3.6 (20)**	5.4 (20)	4.7 (20)	**3.6 (20)**	0.88	0.81	0.99
Heart	**14.2 (9)**	18.5 (13)	17.3 (12)	15.9 (11)	**0.00**	**0.00**	0.14
DLBCL	1.3 (30)	3.8(90)	2.6 (50)	**1.2 (50)**	0.29	**0.05**	0.88
Lung cancer	**2.9 (40)**	5.4 (180)	7.4 (300)	4.9 (300)	0.07	0.26	0.15
Prostate cancer	**2.9(40)**	7.1 (300)	4.9 (200)	4.9 (140)	**0.00**	**0.00**	0.11
SRBC	0 (50)	**0 (25)**	0 (40)	0 (120)	0.62	0.98	0.30
					W/T/L= 4/6/0	**W/T/L= 6/4/0**	**W/T/L= 2/8/0**

A comparison between the obtained results leads mainly to the following observations:

1. Concerning the fuzzy-rule based classifier, we can observe from Figure 5.1 that, except for the Twonorm dataset, the proposed fuzzy weighting approach improves significantly the classifier performance on almost all datasets. A significant gain of performance is achieved by retaining only few features rather than the whole set of original features. For instance, almost 5% of performance gain is achieved using only the four top ranked features rather the nine original features of the Ljubljana dataset. Similarly for the Heart dataset, we gain almost 5% in term of classification performance with only seven features.
2. Although Relief, Simba and I-Relief are based on 1-NN principle, Membas performs similarly or best than I-Relief in nearly all datasets regardless of the used classifier (LAMDA, *k*-NN or SVM) and outperforms Relief and Simba. For more rigorous comparison between the three feature selection methods, a student's paired two-tailed t-test is also performed. The p-value of the t-test reported in each row in Tables 5.2 to 5.4 represents the probability that two sets of compared results come from distributions with equal means. The smaller the p-value, the more significant the difference of the two average values is. At the 0.05 p-value level, Membas wins against Relief and Simba on four cases out of six on UCI datasets with the fuzzy rule based classifier LAMDA (Diabetes, WDBC, Ljubljana, Heart), and in two cases against I-Relief method, and ties on the remaining cases. With *k*-NN, Membas wins on two cases against Relief, in three cases against Simba, and loss in two cases against I-

Relief. Whereas with SVM, Membas wins on four cases with Simba, three cases with Relief, two cases with I-Relief and ties on the remaining cases.

3. Moreover, Membas performs well on UCI datasets containing mixed-type data (ex: Heart, Ljubljana). Especially when the classifier handles appropriately mixed-type data as it can be observed with the fuzzy rule-based classifier LAMDA.

To further demonstrate these interesting properties of the proposed method, we focus on three datasets: Heart (6 qualitative and 7 quantitative features), Ljubljana (6 qualitative and 3 interval features) and Diabetes (8 quantitative features). Figures 5.4, 5.5 and 5.6 give, for the three datasets respectively, the obtained fuzzy weights for one run. For ease of comparison, a normalization of the maximum value of each weight vector is performed to be 1. For Heart dataset, we observe that I-Relief, Relief and Simba do not only assign zeros weights to irrelevant features (the last 50 features in Figure 5.4) but also to the first six qualitative features which are assumed useful ones. Whereas Membas does not only succeed to identify these qualitative features but its top ranked feature is qualitative (feature No.6). The obtained classification errors on this dataset, shown in Figures 5.1, 5.2 and 5.3, prove that Membas can significantly improve the performance of the three classifiers. One possible explanation is that the top ranked features obtained by Membas are more useful for the classification task. As expected, Membas leads to significant improvements of classification performance in the case of mixed feature-type data, due mainly to an appropriate and similar processing for each type of data with minimal loss of information.

Let us focus now on Ljubljana dataset which includes interval and qualitative type features, for which other feature selection methods cannot be applied directly. As the interval feature values of this dataset are regular (not overlapped), they are transformed into ordered numbers to enable I-Relief, Relief and Simba to handle them. It is worthwhile to note that Membas handles the interval features in their original form without any restriction on their relative positions (overlapped or regular); no arbitrary mapping is therefore required. Let us recall that I-Relief, Relief and Simba could not handle intervals if they were overlapped. We observe from Figure 5.5 that Membas identifies correctly the 9 presumably useful features (whatever their type) and assigns approximately zero weights to the 50 last added irrelevant features, whereas Simba and Relief identify mistakenly some irrelevant features as relevant ones. From Tables 5.1 to 5.3, we observe that the minimal classification error on this dataset is obtained with the fuzzy rule-based classifier LAMDA when only the first four top ranked features are

used. This result highlights also the attention to be devoted for choosing an adequate classifier when the data are of mixed type.

We finally focus on the Diabetes dataset (Figure 5.6), for which all the feature selection methods succeed to identify presumably useful features with at least the first common top ranked feature. The obtained classification errors illustrated in Figures 5.1, 5.2 and 5.3 prove the efficiency of Membas to process quantitative features as well as symbolic data. Whereas, we point out that I-Relief, Relief and Simba are typically well-conditionned for processing quantitative features, but are not proficient for handling a dataset of mixed-type data. As expected, SVM performs better than other classifiers on this dataset, especially when only the selected features by Membas are used (see Figure 5.3 and Table 5.3).

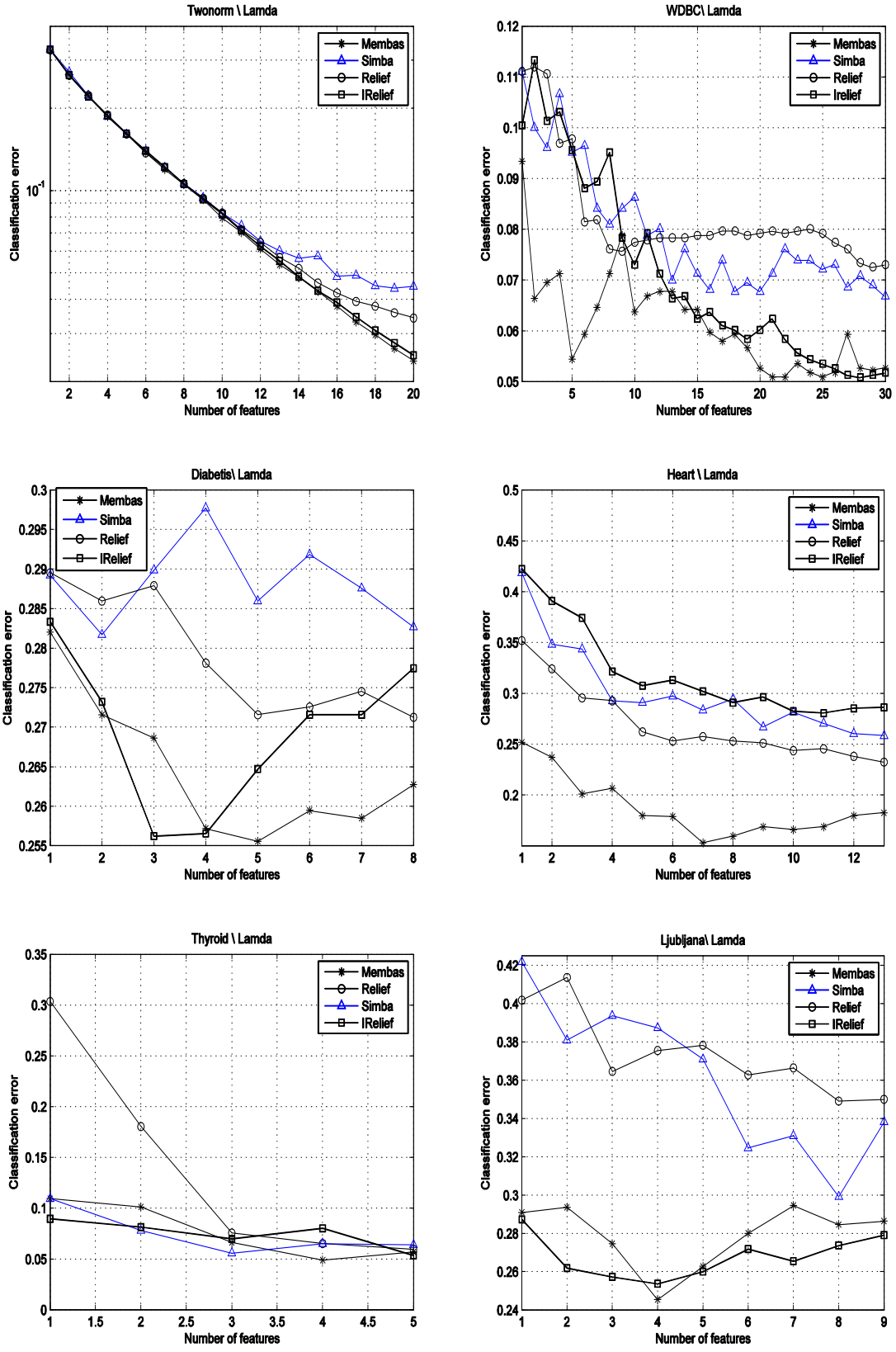


Fig. 5.1. Classification errors obtained by LAMDA on UCI datasets using Membas, I-Relief, Relief and Simba.

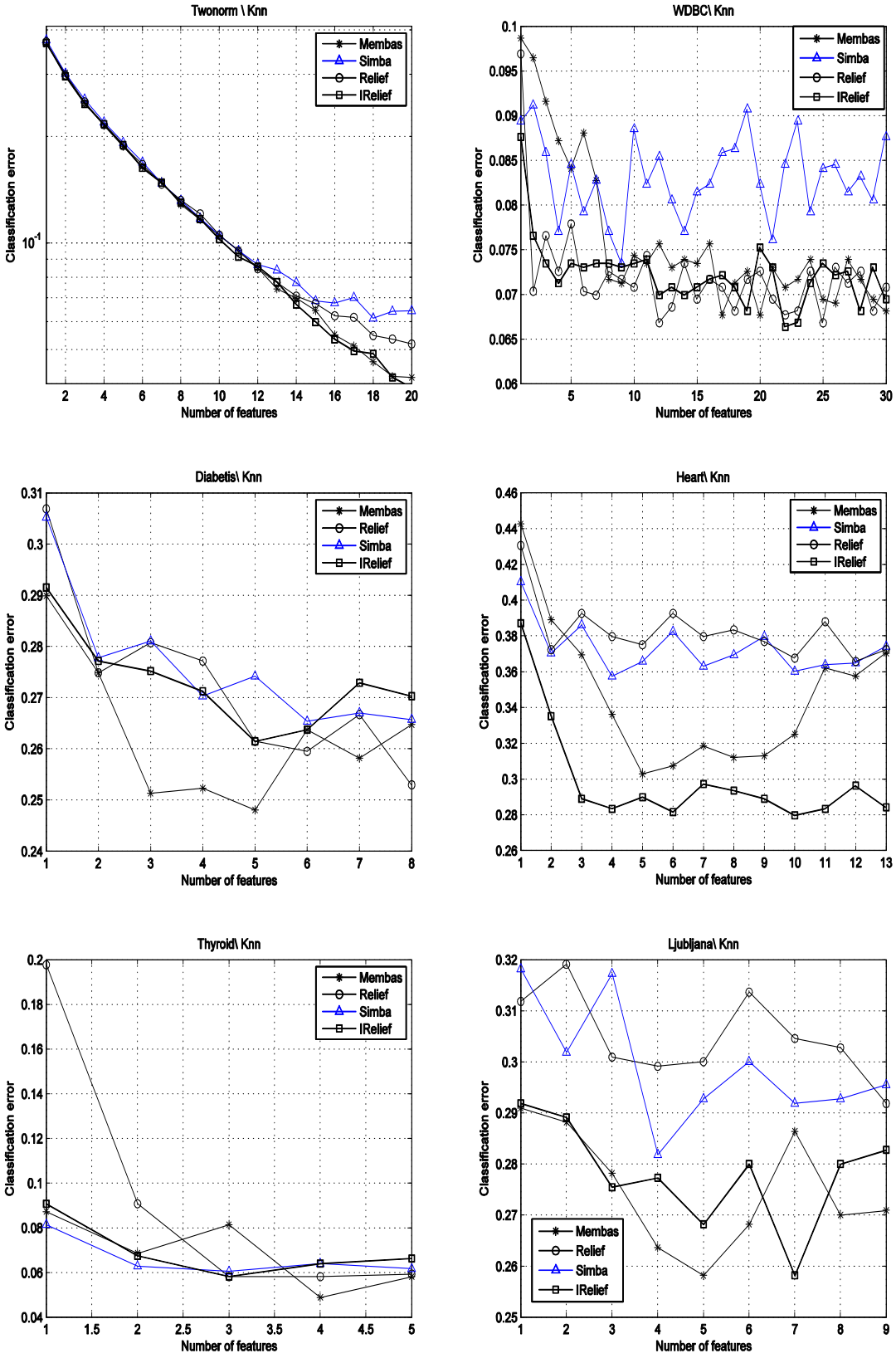


Fig. 5.2. Classification errors obtained by k-NN on UCI datasets using Membas, I-Relief, Relief and Simba

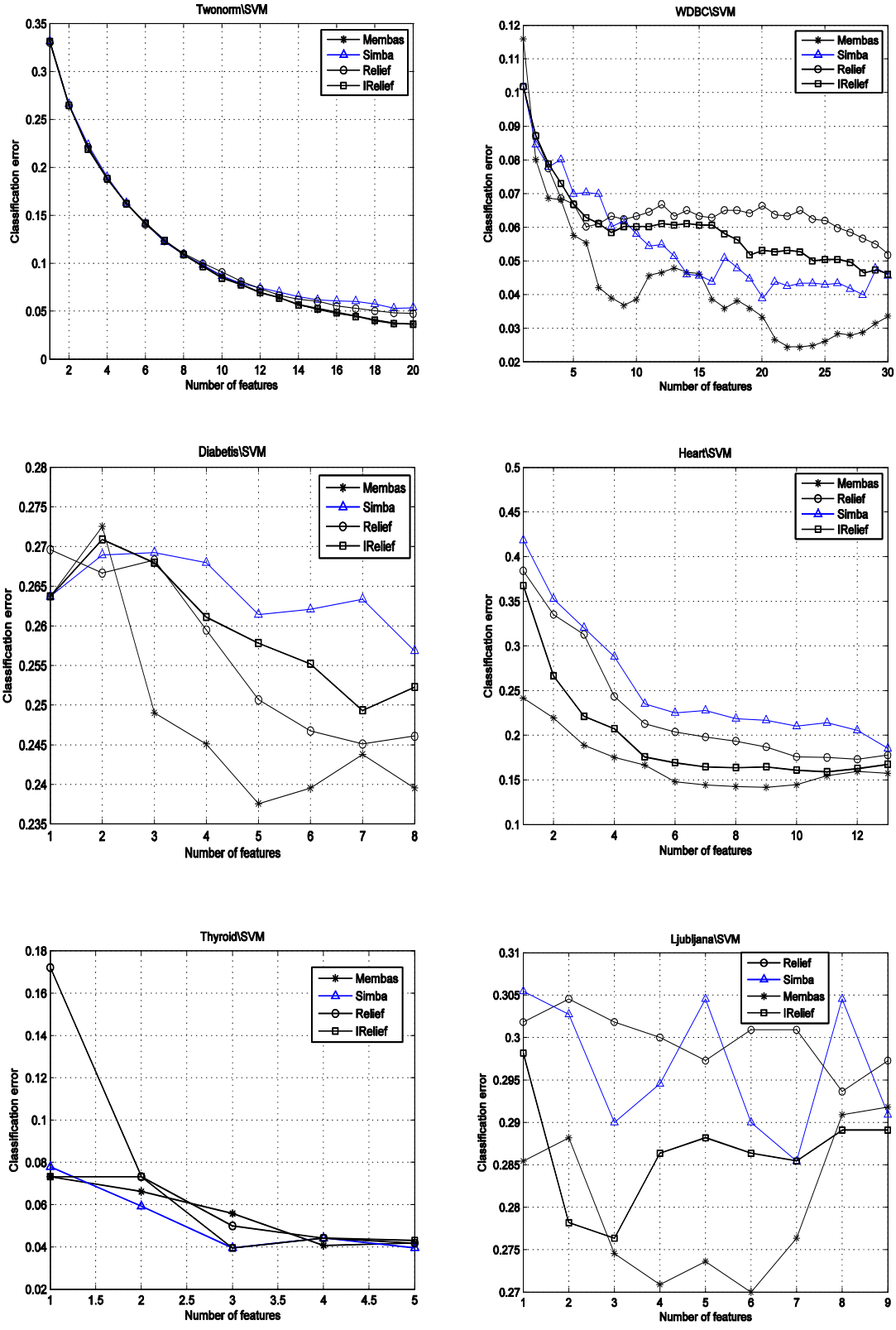


Fig. 5.3. Classification errors obtained by SVM on UCI datasets using Membas, I-Relief, Relief and Simba

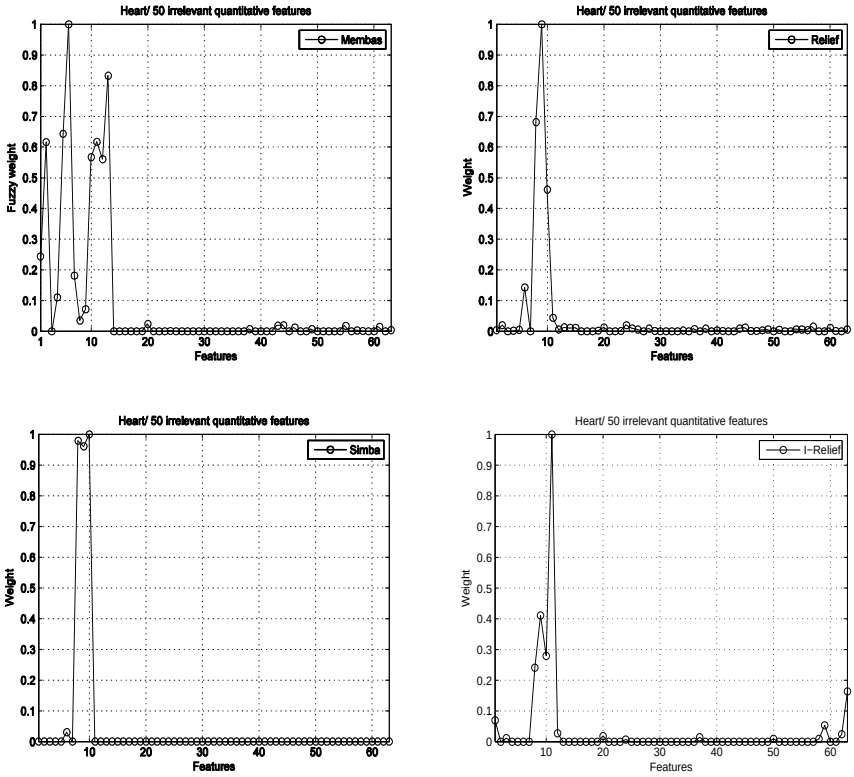


Fig. 5.4. Feature weights obtained by Membas, I-Relief, Relief and Simba on Heart dataset. The first 13 features are the original ones

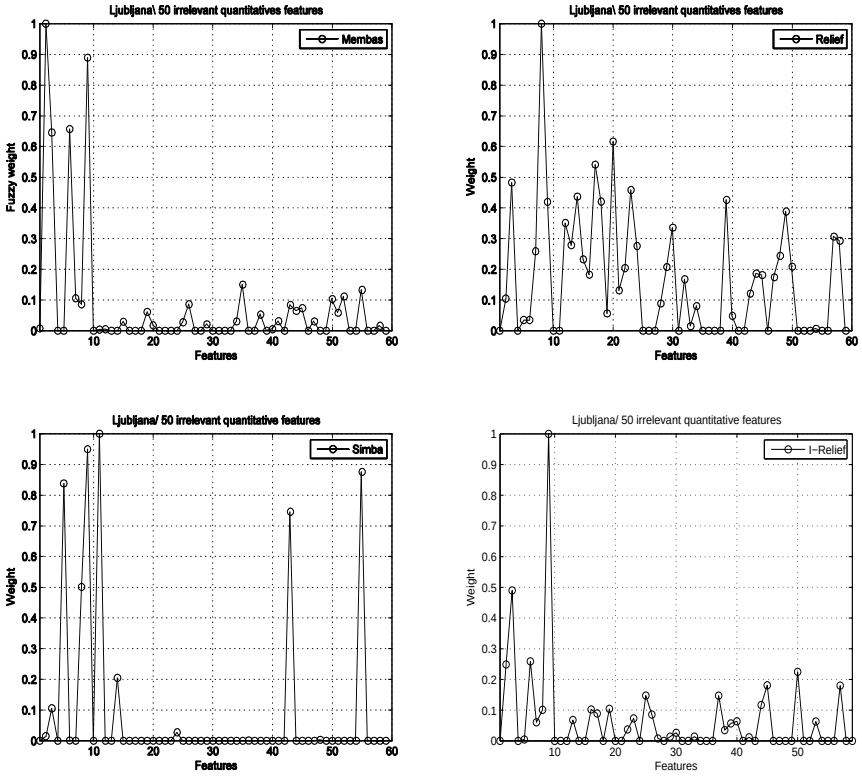


Fig. 5.5. Feature weights obtained by Membas, I-Relief, Relief and Simba on Ljubljana dataset. The first 9 features are the original ones

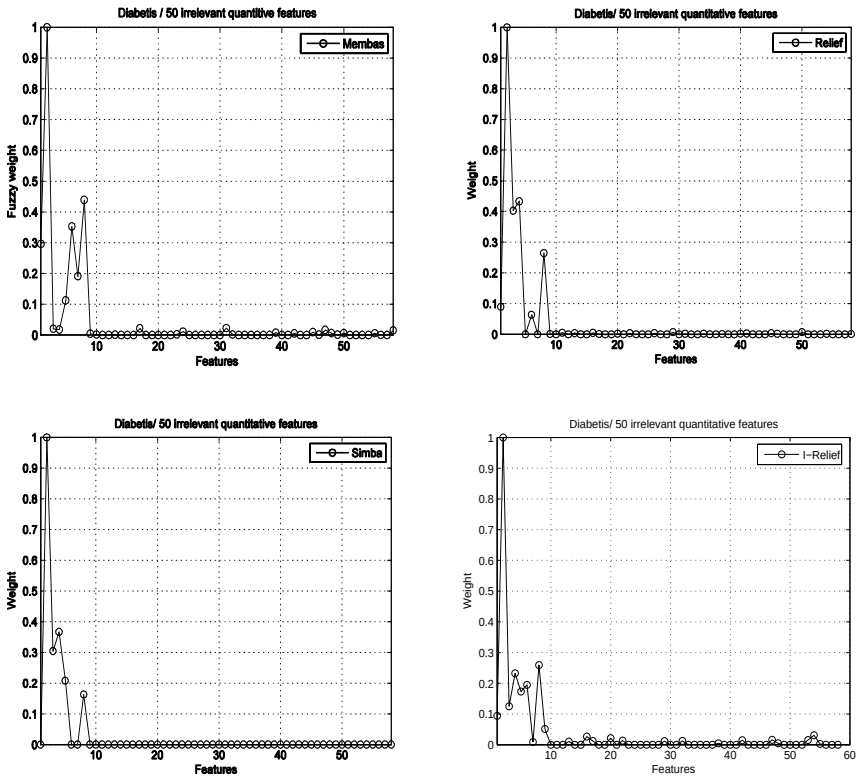


Fig. 5.6. Feature weights obtained by Membas, I-Relief, Relief and Simba on Diabetes dataset. The first 8 features are the original ones

5.5.4 Experiments on high-dimensional datasets

In this section, Membas is compared with I-Relief, Relief and Simba on four microarray datasets. Due to the limited number of samples, the leave-one-out cross validation has been performed to assess the performance of each algorithm. However, in this section we aim to illustrate how the proposed method performs in the presence of huge number of irrelevant features. We noticed in the previous section that Membas processes quantitative data as good as, or better, than other methods on small datasets.

The classification errors obtained using LAMDA, *k*-NN and SVM of the top 400 ranked features are plotted respectively in Figures 5.7 to 5.9 and the corresponding optimal classification performance are reported in Tables 5.2 to 5.4. It can be observed that Membas perform similarly or best than I-Relief, and outperforms Relief and Simba nearly in all datasets using the three classifiers. For prostate cancer dataset, Membas outperforms Relief and Simba over all ranges by 5 to 20 percent with respect to the three classifiers, whereas it yields a better optimal (Test error, No. genes) than I-Relief : a classification error of 5% for only 10 genes with Membas against 6.8% for 50 genes with I-Relief. For SRBC, with LAMDA classifier, Membas converges for only 20 genes whereas I-Relief converges for 60

gens, Simba for 100 genes and Relief for 160 genes. Nevertheless, with *k*-NN classifier, Membas, I-Relief and Relief converge together at 40 genes but it can be observed that Membas provides the minimal classification error before attaining the convergence. For example, in SRBC using LAMDA and *k*-NN classifiers, with 5 genes the error for Membas is 7 percent compared to more than 20 percent for Relief and I-Relief. One possible explanation is that Membas ranks the genes according to their real relevance to this problem so that the *k*-NN classifier performance is maximized. Note also that Membas in SRBC with *k*-NN and SVM reaches near zero percent for only 10 genes. For DLBCL dataset, Membas performs better than Relief and Simba with SVM and LAMDA classifier and yields nearly similar results to these two approaches with k-NN classifier. However, it achieves quite similar or slightly good results compared to I-Relief. For Lung-cancer, we observe that with both classifiers LAMDA and *k*-NN, the error obtained by Membas converges for 70 genes. Whereas with SVM, the classification error achieves its minimal value for only 40 genes.

However, one important issue in using feature selection algorithms in gene selection tasks is to determine a cut-off threshold in a ranked gene list. For some feature weighting approaches (e.g. Relief) a heuristic threshold is proposed for this purpose computed as a function of the number of features (Kira and Rendell, 1992b). One more commonly used method is through cross validation that uses a training data subset to estimate cut-off thresholds simultaneously with the classification parameters, and then using the estimated parameters to classify the held-out testing samples.

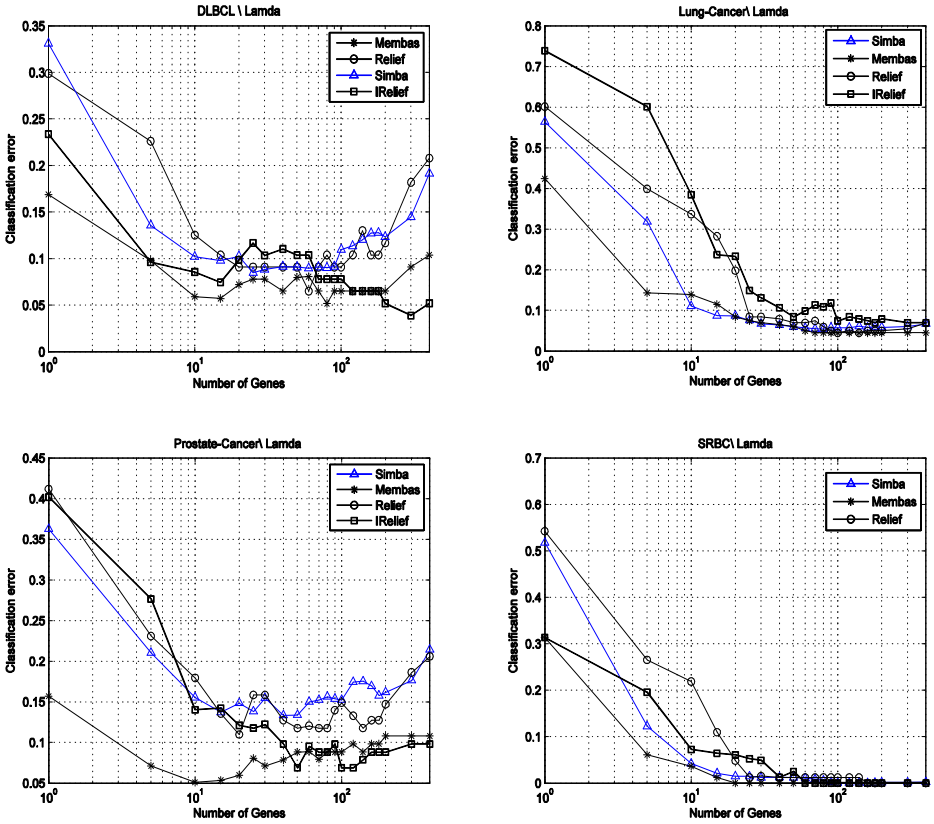


Fig. 5.7. Classification errors obtained by LAMDA on DNA microarray datasets using Membas, Relief and Simba.

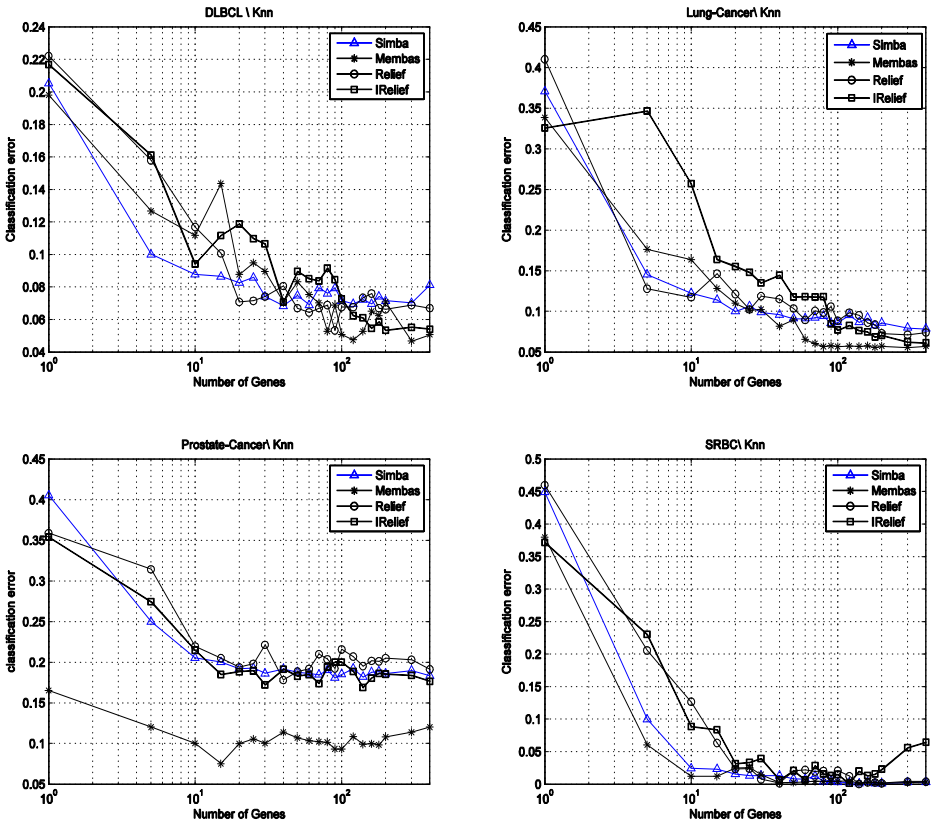


Fig. 5.8. Classification errors obtained by k-NN on DNA microarray datasets using Membas, Relief and Simba.

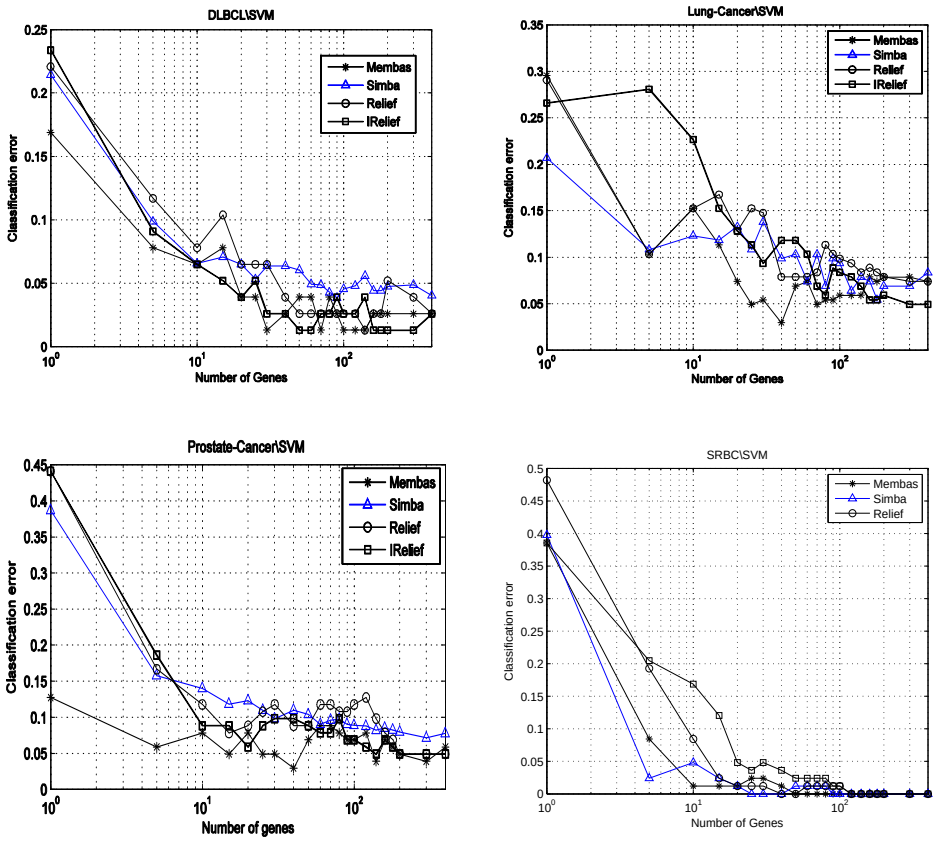


Fig. 5.9. Classification errors obtained by SVM on DNA microarray datasets using Membas, Relief and Simba.

5.6 Conclusion

In this chapter, we have proposed a new feature weighting method for mixed type data based on a membership margin to improve the performance of fuzzy-rule based classifiers. Thanks to the SMSP principle described in the previous chapter, a mapping of all the features from completely heterogeneous spaces to a common space represented by the membership space is performed. Then, the processing of issued data by the mapping step in unified way becomes straightforward for feature weighting. In this order, we introduced a new concept of weighted fuzzy rules such that each antecedent fuzzy set in the fuzzy if-then classification rule is weighted to characterize the importance of each proposition, and thereby the importance of the corresponding feature to the rule's consequence. This operation of fuzzy rule weighting is naturally justified by the estimation of weight in the membership space based on membership margin concept. To avoid any heuristic combinatorial search, these fuzzy weights are estimated by optimizing an objective function within the membership margin framework. An extension of the proposed method to multiclass problems has also been performed. The advantages of the proposed method were firstly illustrated and compared on low-dimensional

real-world datasets, characterized by the presence of mixed-type data, with some well-known feature weighting approaches. The experimental results show that this method leads to a significant improvement of classification performance using fuzzy rule based classifiers as well as other state-of-the art classifiers. The proposed method is however distinguished from other feature weighting methods by its ability to process symbolic intervals without any restrictions on their relative position (regular or overlapped intervals). Further experiments on high-dimensional datasets (DNA microarray dataset) have also proved the effectiveness of the proposed method to perform high-dimensional data.

Unlike the supervised case, feature weighting in the unsupervised case is revealed to be more challenging due to the absence of pattern labels. We try by next chapter to extend this weighting approach to the unsupervised case.

CHAPITRE 6- Résumé
Apprentissage non supervisé basé sur le principe «SMSP»

Le résultat de classification n'est pas toujours disponible au moment de la prise de décision dans de nombreux problèmes pratiques. Dans de telles situations le recours à l'apprentissage non supervisé est une pratique courante. Cependant, pour maintenir une interprétation facile et une grande transparence dans les applications médicales, l'utilisation d'approches non supervisée basées sur des règles peut être aussi d'un grand intérêt. L'apprentissage non supervisé flou notamment offre l'avantage de fournir une base pour la construction des modèles basés sur des règles floues fournissant une représentation simple et une bonne performance pour les problèmes non-linéaires (Yao et al., 2000).

D'un coté, selon la façon dont les données sont traitées, les approches de regroupement non-supervisés (ou «clustering» en anglais) peuvent être divisés en deux classes: «batch» et en ligne. Les algorithmes «batch» traitent à la fois toutes les données disponibles représentées sous la forme d'une table d'individus hors ligne. Alors que dans le cas des algorithmes en ligne, nous considérons que les individus sont reçues en ligne et les partitions de données doit être adaptée itérativement au cours du temps par les informations apportées par les nouveaux individus. Il est maintenant bien reconnu par la communauté d'apprentissage automatique qu'un algorithme de type en ligne est plus efficace qu'un algorithme de type «batch» (Cai et al., 2009). Une approche en ligne est adaptative dans la façon que chaque fois qu'un nouvel individu est reçu, soit un nouveau cluster est généré, dues par exemple à l'apparition d'un nouveau mode, ou seulement les clusters existants sont mis à jours. Le regroupement en ligne nécessite donc un apprentissage non supervisée et incrémental permettant d'incorporer de nouvelles informations dans l'évolution de la partition fur et à mesure qu'un nouvel individu est reçu. Les approches en ligne sont plus générales que les approches dites «batch» dans le sens où les premiers peuvent être utilisés également pour traiter une table d'individus de façon itérative.

D'un autre coté, à l'instar du contexte de classification supervisée, non pas toutes les variables sont utiles pour la tâche de classification non-supervisée et donc juste l'ensemble des variables qui aident à guider le processus de regroupement devrait être sélectionné. Cependant, le problème est plus complexe que lorsqu'une partition de référence est

disponible afin d'évaluer l'importance des variables. Par rapport à l'apprentissage supervisé, peu de travaux ont été consacrés pour aborder le problème de sélection de variables dans le contexte d'apprentissage non supervisé. La plupart des algorithmes de sélection de variables non supervisés sont basées sur des mesures d'information ou de consistance (Mitra et al, 2002; Dy et Brodley, 2004; Wei et Billings, 2007). Comme une approche non supervisée, l'ACP (Analyse en Composantes Principales), par exemple, permet de trouver le sous-ensemble des composants utiles pour la représentation des données. Néanmoins, ces composantes ne sont pas nécessairement utiles pour discriminer entre les groupes dans une tâche de classification non-supervisée (Duda et al., 2001). Si certains de cet ensemble de variables, indépendamment de leur pertinence et importance, sont de type mixte la tâche de classification non-supervisée devient beaucoup plus compliquée.

Dans ce chapitre, nous avons proposé une nouvelle approche basée sur la pondération en ligne de variable pour le regroupement de données hétérogènes. L'algorithme proposé est une extension de notre algorithme de pondération de variables développé précédemment pour la classification supervisée. Pour faire face au problème de l'hétérogénéité des données, le principe SMSP est étendu ici aussi pour traiter d'une façon unifiée les données hétérogènes dans un cadre non supervisé. Toutefois, il a été montré que l'étape de projection des données dans un espace commun doit être réalisée de façon incrémentale pour tenir compte du nouvel individu reçu à chaque itération du processus d'apprentissage. Pour cette raison, une version itérative de la fonction caractéristique introduit dans le cas supervisé a été fournie en fonction de chaque type de variable.

Tout d'abord, l'algorithme de la méthode de regroupement incrémental en ligne basé sur des règles floues a été décrit. Ensuite, nous avons étudié comme pour le contexte supervisé, l'intégration de la tâche de pondération de variables dans le processus du regroupement pour la conception de notre approche en se basant sur le concept de règles floues pondérées. Cette approche est basée aussi sur la maximisation itérative de la marge d'appartenance. Une étude extensive expérimentale a été ensuite effectuée sur des problèmes artificiels et réels pour prouver l'efficacité de l'approche proposée. Sur un exemple artificiel cette approche a permet d'identifier correctement l'ensemble des classes et aussi l'ensemble des variables non pertinentes. Alors que dans le cas des problèmes réels, cette approche a gagné contre la méthode C-Moyennes Floues (FCM) dans 12 cas sur 14. Cet algorithme ne parvient pas cependant à traiter des problèmes de haute dimension (par exemple données issues de biopuces). Cela est dû probablement au grand nombre de variables non pertinentes (des milliers) par rapport à celle pertinentes (des dizaines au maximum).

CHAPTER 6
Unsupervised learning based on SMPS principle

The pattern labels are not always available the time of decision making in many practical problems. In such situations a resort to unsupervised learning capabilities is a common practice. However, to maintain an easy interpretation and high transparency in medical applications, the use of rule-based unsupervised approaches can be also of big interest. Fuzzy unsupervised learning particularly offers the advantage to provide a basis for constructing rule-based fuzzy model that has simple representation and good performance for non-linear problems (Yao *et al.*, 2000).

From one hand, according to how the data is processed, clustering approaches can be divided into two classes: batch and online. Batch algorithms process at once all available data represented by a table of patterns offline. Whereas in online algorithms we consider that patterns are received online and data partitions should be adapted iteratively over the time by information brought by new patterns. It is now well-recognized by the machine learning community that an online algorithm is computationally more efficient than a batch one (Cai *et al.*, 2009). An online approach is adaptive in the way that each time a new pattern is received, it either generates a new cluster, due for instance to new mode apparition, or only updates the existing clusters. Online clustering requires therefore unsupervised and incremental learning rules that enable to incorporate new information in partition evolution over time. Online approaches are more general than batch approaches in the sense that they can be used also to process a table of patterns in an iterative manner.

From other hand, similarly to the supervised classification context, not all the features are important for clustering task and therefore only the set of features that help to guide the clustering process should be selected. However, the problem is more complex than when a reference partition of patterns is available to assess the importance of features. Compared to the supervised learning only few works have been devoted to address the feature selection problem for unsupervised learning. Most of unsupervised feature selection algorithms are based on information or consistency measures (Mitra *et al.*, 2002; Dy and Brodley, 2004; Wei and Billings, 2007). As an unsupervised approach, PCA (Principal Component Analysis) for

instance enables to find subset of components useful for data representation. Nevertheless, these components are not necessarily useful to discriminate between clusters in a clustering task (Duda *et al.*, 2001). If some of the original set of features, regardless of their relevance and importance, are of mixed type the clustering task becomes more challenging.

In this chapter we propose a novel approach based on online feature weighting for clustering of heterogeneous data. The proposed algorithm is built based on an extension of our previously developed supervised learning feature weighting algorithm. So, first, to cope with the problem of data heterogeneity, the SMSP principle presented in chapter 4 is extended here also to reason in a unified way about heterogeneous data in an unsupervised framework. However, the mapping step should be performed in an incremental fashion to take into account new pattern at each iteration of the learning process. In this order, iterative version of the mapping function introduced in chapter 4 is provided here according to each feature type. We describe first an online incremental clustering algorithm based on a fuzzy rule-based system. We investigate then, as for the supervised context, the integration of the feature weighting task in the clustering process to design our proposed approach based on fuzzy weighted rules concept. An extensive experimental study is then performed on artificial and real-world problems to prove its effectiveness. However, it is worthwhile to note that, even of its interesting properties, this approach has been found unable to fit with high-dimensional data.

6.1 Iterative membership functions updating

Unlike the supervised case, the mapping step in the unsupervised framework should be performed iteratively based on online learning. At each iteration the membership functions are updated by the information brought by a new pattern according to each feature type as follows:

6.1.1 Quantitative type features

Different possible membership functions used in the supervised case can be adapted for the unsupervised case to quantitative feature type such as:

a. Gaussian-like memebership function

$$\mu_k^i(x_i) = e^{-\frac{1}{2}(x_i - \varphi_k^i)^2 / \sigma_i^2} \tag{6.1}$$

b. Binomial membership function

$$\mu_k^i(x_i) = \varphi_k^{i\,1-x_i}\left(1-\varphi_k^i\right)^{x_i} \tag{6.2}$$

However, the major difference with the supervised case is that the parameters φ_k^i representing the i^{th} feature's mean of the N_k patterns clustered in class C_k are updated iteratively by online learning as follows:

$$\varphi_k^i(N_k+1) = \varphi_k^i(N_k) + \frac{1}{N_k+1}\left(x_i^j(N_k) - \varphi_k^i(N_k)\right) \tag{6.3}$$

σ_i represents an approximation of the standard deviation, which converges to the real one whenever a big number of samples is considered, and is updated iteratively by the following expression:

$$\sigma_k^{i\,2}(N_k+1) = \sigma_k^{i\,2}(N_k) + \frac{\left(x_i^j(N_k) - \varphi_k^i(N_k)\right)^2 - \sigma_k^{i\,2}(N_k)}{N_k+1} \tag{6.4}$$

6.1.2 Interval type features

The membership function for interval type features is also taken as the similarity described by the equations (5.6, 5.7 or 5.8) between the symbolic interval value of the i^{th} feature x_i and the interval $\rho_k^i = \left[\rho_k^{i-}, \rho_k^{i+}\right]$ representing cluster C_k as

$$\mu_k^i(x_i) = S\left(x_i, \rho_k^i\right) \tag{6.5}$$

For N_k patterns assigned to cluster C_k, the cluster prototype is a vector whose components are the intervals obtained by the mean bounds updated also iteratively by online learning as follows:

$$\begin{cases} \rho_k^{i-}(N_k+1) = \rho_k^{i-}(N_k) + \frac{1}{N_k+1}\left(x_i^{j-}(N_k) - \rho_k^{i-}(N_k)\right) \\ \rho_k^{i+}(N_k+1) = \rho_k^{i+}(N_k) + \frac{1}{N_k+1}\left(x_i^{j+}(N_k) - \rho_k^{i+}(N_k)\right) \end{cases} \tag{6.6}$$

Where x_i^{j-} is the i^{th} feature lower bound for the j^{th} sample and x_i^{j+} is its upper bound. Consequently, the resulted cluster prototype at each iteration for the r interval features is given by the vector of intervals:

$$\rho_k = \left[\rho_k^1, \rho_k^2, \ldots, \rho_k^r\right]^T$$

6.1.3 Qualitative type features

The membership function for the i^{th} qualitative feature is specified as:

$$\mu_k^i\left(\chi_i\right)=\left(\Phi_{k1}^i\right)^{q_{i1}}*\cdots*\left(\Phi_{kMi}^i\right)^{q_{iMi}} \tag{6.7}$$

Where Φ_{kj}^i is the frequency of modality Q_j^i in cluster C_k updated iteratively by online learning as:

$$\Phi_{kj}^i(N_k+1)=\Phi_{kj}^i(N_k)+\frac{1}{N_k+1}\left(q_j^i(N_k)-\Phi_{kj}^i(N_k)\right) \tag{6.8}$$

and $$q_j^i=\begin{cases}1 \;\; if \;\; x_i=Q_j^i\\ 0 \;\; if \;\; x_i\neq Q_j^i\end{cases}$$

Therefore, the cluster prototypes at each iteration are represented by $\Omega_k^i=\left[\Phi_{k1}^i,\ldots,\Phi_{kj}^i,\ldots,\Phi_{kMi}^i\right]$

Unlike the supervised case, the mapping step of the SMSP principle is performed here online for different types of features in the membership space. At each learning iteration (i.e. receive a new pattern) a Membership Degree Vector (MDV) of dimension *m* is associated for a given pattern $x^{(n)}$ to each cluster as follows:

$$U_{nc_k}=\left[\mu_k^1(x_1^{(n)}),\mu_k^2(x_2^{(n)}),\ldots,\mu_k^m(x_m^{(n)})\right]^T \;\; ; k=1,2,\ldots,l \tag{6.9}$$

where $\mu_k^i(x_i^{(n)})$ (i.e. $\mu_k^i(x_i=x_i^{(n)})$), is the membership function of cluster C_k at the current iteration evaluated for a given value x_{ni} of the i^{th} feature of pattern $x^{(n)}$.

As for the supervised learning, once all features are simultaneously mapped into a common space, they can be henceforth processed similarly for clustering.

6.2 Online fuzzy clustering for mixed-type data

We describe here separately the approach used to cluster online a set of patterns, possibly represented by mixed type of features, based on a simple fuzzy reasoning mechanism. Contrary to the supervised case, no predefined partition is available, for that, an *adaptive* fuzzy reasoning mechanism based on incremental online learning is adopted.

When a new pattern is received, the reasoning mechanism should place it in one of the already pre-established clusters corresponding to the highest degree of adequacy. To ensure that each pattern satisfies a minimal threshold of adequacy to each cluster, a Virtual Cluster (VC) is assumed to be always present in the space of clusters. This cluster receives the pattern for whom its adequacy degree is not sufficient to place it in any of the previously created clusters. Whenever a pattern appears to have a higher membership to VC, it means that a new cluster must be created to correspond to the new information brought by this pattern. Then,

the representation of the cluster which receives this pattern (VC or one of the pre-established clusters) is updated to take into account the information brought by this element.

As for the supervised context, the fuzzy inference mechanism proposed here for clustering of heterogeneous data is rule-based. In the beginning of a clustering task, the rule base is initialized by a single If-Then rule representing the virtual class VC given as:

R_{vc}: If x_1 is A_{1vc} and x_2 is A_{2vc} and x_m is A_{mvc} then x belongs to cluster VC

Where the antecedent fuzzy sets A_{ivc} are pre-defined membership functions specified here according to the i^{th} feature type as follow:

(i) Quantitative type feature

$$\mu_{VC}^{i}(x_i) = \varphi_{VC}^{i\ 1-x_i}\left(1-\varphi_{VC}^{i}\right) \quad \text{with} \quad \varphi_{VC}^{i} = 1/2$$

(ii) Interval type feature

$$\mu_{VC}^{i}(x_i) = S\left(x_i, \rho_{VC}^{i}\right) \quad \text{with} \quad \rho_{VC}^{i} = [0,1]$$

(iii) Qualitative type feature

$$\mu_{VC}^{i}(x_i) = \frac{1}{cardinal(D_i)}$$ with D_i the set of possible modalities of the i^{th} feature

Then, whenever the creation of new cluster C_k is deemed necessary, a single *adaptive* fuzzy If-Then rule is generated and associated to this cluster in the clustering rule base:

R_k: If x_1 is A_1 and x_2 is A_2...and x_m is A_m then x belongs to cluster C_k

When a new pattern should be allocated to a cluster, the following fuzzy *adaptive* inference engine can be used also but taking into account its membership to class VC as follows:

$$R^{*} = \underset{R_k}{Arg\ \max}\left\{\alpha\ \gamma\left[\mu_k^1(x_1),...,\mu_k^m(x_1)\right] + (1-\alpha)\ \beta\left[\mu_k^1(x_1),...,\mu_k^m(x_1)\right] / k = VC,1,...,l\right\}$$

where γ and β are dual fuzzy aggregation functions that combine membership (given by the components of the MDV U_{nc_k}) of features value of a pattern $x^{(n)} = [x_1^{(n)}, x_2^{(n)},...,x_m^{(n)}]$ to a cluster C_k. and α is the "*exigency*" parameter playing the same role as in the supervised context.

If the new pattern is placed in one of the existing clusters, the antecedent fuzzy sets in the rule corresponding to this class are updated by the information brought by this element as described in section (6.1). Otherwise, a new cluster including this unique element is created using VC and therefore its corresponding fuzzy If-Then rule must also be generated and added to the already established rule base.

Consequently, the word *adaptive* refers here to the fact that the reasoning mechanism is able to either update the membership functions corresponding to antecedent fuzzy sets of the winner rule or generate new rules, according to the fuzzy reasoning decision made about the current processed pattern. Without the unification of the space of features, this simple inference mechanism could not be applied, and the influence of the different types of features would not be equal.

This simple approach of online fuzzy clustering can be described by the following algorithm.

Algorithm

1. *Initiate the space of classes by the cluster VC*
2. *For t=1…T (T : number of input patterns)*
    a) *Input a new pattern* $x^{(n)}$
    b) *Obtain membership degree vectors MDVs* U_{nc} *and* $U_{n\tilde{c}}$ *for sample* $x^{(n)}$ *through antecedent fuzzy set of If-Then rules*
    c) *Perform the fuzzy inference and assign* $x^{(n)}$ *to a cluster based on maximum membership rule.*
    d) *Update the parameters of antecedent fuzzy sets of the winner rule by the information brought by* $x^{(n)}$.

As reported for the supervised case, not all of the features are important for clustering task and therefore there is a need to discard the irrelevant ones. We describe by next an online feature weighting approach for clustering of heterogeneous data based on the previously presented clustering approach.

6.3 Online fuzzy feature weighting for heterogeneous data clustering

Online learning was considered previously to describe clustering process. In the aim to improve the clustering performance, we investigate here an integration of a feature weighting task in the clustering process based on fuzzy weighted rule concept. In literature, a first attempt to use a similar concept, denoted as Weighted Fuzzy Production Rules WFPR, was performed by (Chen, 1994). In addition to the assignment of a weight to each proposition in the antecedent part, WFPR allows to contain some fuzzy quantifiers (such as “strong”, “ weak”,…) and introduces a certainty factor to characterize the belief on the rule (Ishibuchi and Yamamato, 2005; Ishibuchi and Tomoharu, 2001). In (Chen, 1994) weighted fuzzy rules were used to perform medical diagnosis but assuming that the rules and their corresponding weights were known or fixed a priori by the expert. An alternative approach is reported in

(Rasmani and Shen, 2004) which uses subsethood concept to promote certain linguistic terms as part of the antecedent of a fuzzy rule. However, in all previously stated works only a supervised learning problem (classification task) was considered. Moreover, although the weight assignment borrows the idea of fuzzy classification in these works, it is still actually not explicitly related to feature selection. We propose here to assign a weight to each proposition in the antecedent of the fuzzy clustering rule that represents, at each iteration, the degree of importance of its corresponding feature in the membership space.

Similarly to the supervised case, we propose here to perform the clustering by weighting the antecedents of all IF-THEN rules according to the current estimated importance (fuzzy weight) of each feature in the membership space. Thus the clustering rule associated to each cluster becomes:

R_k: If x_1 is A_1 and x_2 is A_2...and x_m is A_m then x belongs to cluster C_k, w

Which can be noted equivalently

R_k: If x_1 is (w_1) A_1 and x_2 is (w_2) A_2...and x_m is (w_m) A_m then x belongs to cluster C_k

Each of the antecedent fuzzy sets A_i is modeled by a membership function according to the i^{th} feature type.

Therefore, the remaining issue is how to evaluate appropriately the weights of each feature knowing that they must be used to modify the membership to antecedent fuzzy sets of clustering rules. Similarly to the supervised case, it would be natural to estimate these weights also in the membership space to justify such an operation. In this order, we take advantage of the SMSP principle here to define a membership-margin based objective function to evaluate the importance of each feature in the membership space. The main difference with the supervised framework is that the weights have to be estimated and updated iteratively to guide the clustering task. At each iteration the weights are computed such that the discriminative power between all existing clusters is maximized based on an optimization approach. Therefore, only a single weight vector is needed to weight fuzzy antecedents' sets of all clustering rules such that it reflects the relevance of each feature simultaneously to all rule's consequences. Furthermore, the integration of feature weighting into the clustering process becomes straightforward thanks to the unification of feature spaces described in chapter 4.

Weighted Fuzzy Rule-based Clustering Algorithm

Our final aim is to combine clustering and feature weighting by introducing the feature weighting method into the online clustering algorithm described previously. The fuzzy feature weighting approach proposed in chapter 5 is based on online learning. We show by next that this approach can be extended for the clustering task. For ease of presentation, let's consider for a moment that after some iterations of fuzzy weighted clustering the resulted partition of data is described by the dataset $D_t = \{x^{(n)}, C_k\}_{n=1}^{N_t} \in X \times C$, where $x^{(n)}$ is the n^{th} pattern (item) and N_t is the number of already clustered patterns. C_k is the class label assigned to each pattern among the clusters generated by the fuzzy weighted clustering task; *k=1,2,…,l*. Let us consider that w_t is the fuzzy weight vector computed during the fuzzy clustering task. The margin concept proposed for feature weighting in supervised case can be extended here to perform an iterative feature weighting task for clustering. When a new pattern $x^{(n)}$ should be processed, our clustering system based on the weights w_t , estimated in the previous iteration, assigns it either to one of existing clusters or to the VC cluster (i.e. creation of new cluster). For simplicity, let's consider that after the clustering of the pattern $x^{(n)}$ the data exhibits only two clusters, namely the cluster to whom $x^{(n)}$ has been affected noted $c = C_1$ and an alternative cluster noted $\tilde{c} = C_2$. We seek in a next step to take into account the information brought by $x^{(n)}$ for updating the feature weights to use them then for the clustering of future patterns. Once the pattern $x^{(n)}$ is clustered, its Membership Margin can be defined based on SMSP principle as:

$$\beta_n = \psi(U_{nc}) - \psi(U_{n\tilde{c}}) \tag{6.10}$$

where *c* is the cluster in which $x^{(n)}$ has just been clustered and $\tilde{c}$ is the alternative cluster. As for the supervised context, by scaling the features in the membership space a weighted version of the membership margin can be defined. A similar margin-based objective function can be therefore designed for the same purpose and the feature weighting problem can be solved using the same optimization approach. Consequently, the weight vector can be updated using the analytical solution $w_f^* = s^+ / \|s^+\|$ at each iteration, through the updating of the vector s resulted at the previous iteration by the information brought by $x^{(n)}$. The extension of this approach to the multiclass problems is also straightforward using the definition employed in the supervised framework.

The proposed approach can be summarized in two alternating steps:

1- Clusters a new pattern based on weighted fuzzy clustering rule.

In this step, the clustering is performed by weighting the antecedents of all IF-THEN rules according to the a priori estimated importance (fuzzy weight) of each feature in the membership space. For initialization step, the fuzzy weight should be set to one which means that initially all features, and thereby their associated propositions in the fuzzy IF-THEN rules, are considered of equal importance.

The clustering rule associated to each cluster is given by:

R_k: If x_1 is (w_1) A_1 and x_2 is (w_2) A_2...and x_m is (w_m) A_m then x belongs to cluster C_k ($k=1,\ldots,l$) where A_i ($i=1,\ldots m$) is the antecedent fuzzy set defined according to the type of feature x_i. When a new pattern $x^{(n)} = [x_1^{(n)}, x_2^{(n)}, ..., x_m^{(n)}]$ should be clustered, its membership degree vector for a cluster C_k is $U_{nc_k} = \left[\mu_k^1(x_1^{(n)}), \mu_k^2(x_2^{(n)}), ..., \mu_k^m(x_m^{(n)})\right]^T$, obtained by evaluating the antecedent fuzzy sets of the rule R_k. If we characterize the operation of weighting the antecedent fuzzy sets as a scalar multiplication, the weighted membership degree vector of $x^{(n)}$ can be computed as:

$$\hat{U}_{nc_k} = w_f .* U_{nc_k} = [\hat{\mu}_k^1(x_{n1}), \hat{\mu}_k^2(x_{n2}), ..., \hat{\mu}_k^m(x_{nm})]^T ; k = VC, 1, 2, ..., l \quad (6.11)$$

This is equivalent to write $\hat{\mu}_k^i(x_{ni}) = w_i . \mu_k^i(x_{ni})$.

Then, a fuzzy *adaptive* engine inference can be used to cluster $x^{(n)}$ using its previously computed weighted membership degree vector as follows:

$$R^* = \arg\max_{R_k} \left\{ \alpha\, \gamma[\hat{\mu}_k^1(x_1), ..., \hat{\mu}_k^m(x_m)] + (1-\alpha)\beta[\hat{\mu}_k^1(x_1), ..., \hat{\mu}_k^m(x_m)]; k = VC, 1, 2, ..., l \right\}$$

2- Updating the fuzzy weights by information brought by this new pattern.

The second step pertains the updating of fuzzy weights by the information brought by the new pattern $x^{(n)}$ which has been clustered in the step 1. To ensure that any of the features, at a given iteration, is not definitely excluded from clustering process, its weight w_i must be updated through the vector *s* calculated in the previous iteration. When a new pattern brings new information attesting an increasing importance of one feature which was deemed till previous iteration irrelevant, our fuzzy system should take into account this information online and update the confidence of clustering task on this feature and therefore on its associated proposition in the IF-THEN rule.

For that, we update firstly the vector s given by (5.9):

$$s = s + \min_{\{\tilde{c} \in C, \tilde{c} \neq c(x_n)\}} \{U_{nc} - U_{n\tilde{c}}\} \tag{6.12}$$

Thus, the fuzzy weight vector w_f at this iteration can be computed based on s by

$$w_f^* = \frac{s^+}{\|s^+\|} \tag{6.13}$$

where $s^+ = [\max(s_1,0), \ldots, \max(s_m,0)]^T$

Interpretation: The weights are estimated in the membership space so it is natural that we take advantage of this available information to change iteratively the importance of each fuzzy set in the If-Then rules to guide the clustering task.

Algorithm

The weighted fuzzy clustering algorithm can be described as follows:

1. *Initiate the fuzzy weight vector to one, initiate the space of clusters by the class VC*
2. *For t=1…T*
   a) *Input a new pattern* $x^{(n)}$
   b) *Calculate the membership degree vectors MDVs* U_{nc} *for sample* $x^{(n)}$ *to each cluster through antecedent fuzzy set of If-Then rules.*
   c) *Perform the weighted fuzzy inference and assign* $x^{(n)}$ *to the cluster corresponding to the winner rule*
   d) *Update the parameters of antecedent fuzzy sets of the winner rule by the information brought by* $x^{(n)}$.
   e) *Update vector s as*

$$s = s + \{U_{nc} - U_{n\tilde{c}}\}$$

   f) *Calculate the new optimal fuzzy weight vector at iteration t as*

$$w_f^* = \frac{s^+}{\|s^+\|}$$

   with $s^+ = [\max(s_1,0), \ldots, \max(s_m,0)]^T$

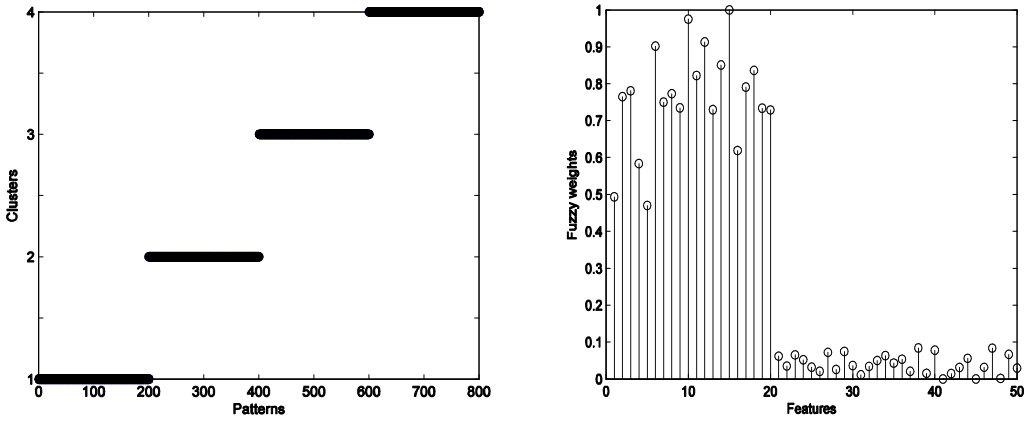
6.4 Experimental results

The performance of the proposed weighted fuzzy clustering approach, referred to as WFCA, was evaluated using artificial and real-world datasets described in Table 6.1. Patterns are already grouped into a priori known classes of unequal size corresponding to their class label. An independent clustering on all datasets using the weighted fuzzy reasoning tool is performed and the obtained cluster partitions are compared with the classes known a priori.

6.4.1 Synthetic Data

The used synthetic dataset here consists of 800 data points from a mixture of four equiprobable Gaussians (200 patterns for each class) described by 20 quantitative relevant features. Moreover, 30 independently normal distributed irrelevant features with zero mean and unit variance were appended to the 20 relevant features to assess the robustness of the newly proposed method against irrelevant features, yielding a set of 800 50-dimensional patterns. We ran the proposed algorithm 10 times using a Gaussian-like membership function, and feature weight vector initialized at 1. For better visualization, a normalization of the maximum value of each weight vector is performed to be 1. In all the 10 runs, the four clusters were correctly identified. Figure 6.1 (a) shows the obtained classification results and (b) the fuzzy weights of all the 50 features. It can be observed in this case that the algorithm successfully clusters all the patterns and correctly identifies the last 30 irrelevant features from the relevant ones by assigning them close to zero weights.

However, we have found out empirically that when the number of irrelevant features becomes very important (10 times multiple of the number of relevant features) this approach fails completely to locate the good clusters. That is may be the reason why it becomes unpractical on high dimensional data. This algorithm was applied without success on microarray data characterized by a huge number (thousands) of irrelevant features.



(a) (b)

Fig. 6.1. (a) Clustering results (b) Fuzzy features weights

6.4.2 Real data

We tested our algorithm on several datasets with different characteristics (Table 6.1). Since all datasets have been collected for supervised classification (i.e. a previous partition of the dataset was available), the class labels were only used to evaluate the clustering performance. An independent clustering was carried out on all datasets and the overall rate of accuracy is

shown in Table 6.2. Clustering accuracy is calculated by comparing the obtained clusters with the real partition provided in the dataset.

Tab. 6.1. Summary of used Datasets

Dataset	No. Feature	Quant.	Qual.	Interv.	Class	Nb. patterns
Iris	4	4	0	0	3	150
Ljubljana*	9	0	6	3	2	286
Thyroid	5	5	0	0	3	215
WDBC	30	30	0	0	2	699
Liver	6	6	0	0	2	345
Australian credit card*	15	6	9	0	2	690
Hepatitis*	17	4	13	0	2	155
Diabetics	8	8	0	0	2	768
Heart	13	7	6	0	2	270
Wine	13	13	0	0	2	178
Car data	8	0	0	8	4	33
Fish data	13	0	0	13	4	12
Barcelona water	48	0	0	48	5	316
Temperature cities	11	0	0	11	4	37

(*) Missing data excluded

To further evaluate the performance of the proposed methodology for clustering, we compared it with the well known Fuzzy c-means clustering method (FCM) (Bezdek, 1981) using all features. We simply set the number of clusters equal to the number of original labels provided in the dataset. However, various clustering validity indices (Wang and Zhang, 2007) can be used here to select the optimal number of clusters whenever the number of original clusters is unavailable. It must be noted that FCM does not handle qualitative and interval data. However, qualitative and regular intervals (interval features take their values from a countable set of interval values), are transformed into quantitative values to enable handling them by FCM. We notice also that three of the fourteen used datasets (“Car”, “temperature cities”, “Fish” and “Barcelona water”) are characterized by overlapped interval features and therefore FCM could not be applied on them. Results obtained with FCM are shown also in Table 6.2. It can be observed that the proposed approach outperforms the FCM on almost all datasets (12 among 14). One possible explanation is the incorporation of the importance of each feature to guide the clustering process. Moreover, thanks to the SMSP principle, the proposed approach allows handling appropriately the qualitative and interval data, unlike FCM which requires a transformation of qualitative and interval spaces into quantitative space. Figure 6.2 shows the fuzzy weights obtained at the end of clustering task for each dataset. FWCA approach can provide online precious information about the importance of

each feature for the ongoing clustering task. Moreover, the user can fix a weight threshold in some specific cases to discard the deemed irrelevant features during the clustering task.

Tab. 6.2. Clustering error of the proposed and FCM approaches

Dataset	**FWCA***	**FCM**
Iris	**1.33**	3.74
Ljubljana	**26.71**	28.88
Thyroid	**3.72**	51.16
WDBC	15.47	**7.21**
Liver	**11.01**	51.30
Credit card	**13.63**	17.15
Hepatitis	**10.85**	33.33
Heart	**16.67**	26.76
Diabetics	**28.26**	33.33
Wine	28.65	**5.06**
Car data	**36.36**	-
Temperature cities	**16.22**	-
Fish data	**41.67**	-
Barcelona Water	**35.76**	-
14	**12**	**2**

(*): Fuzzy Weighted Clustering Approach

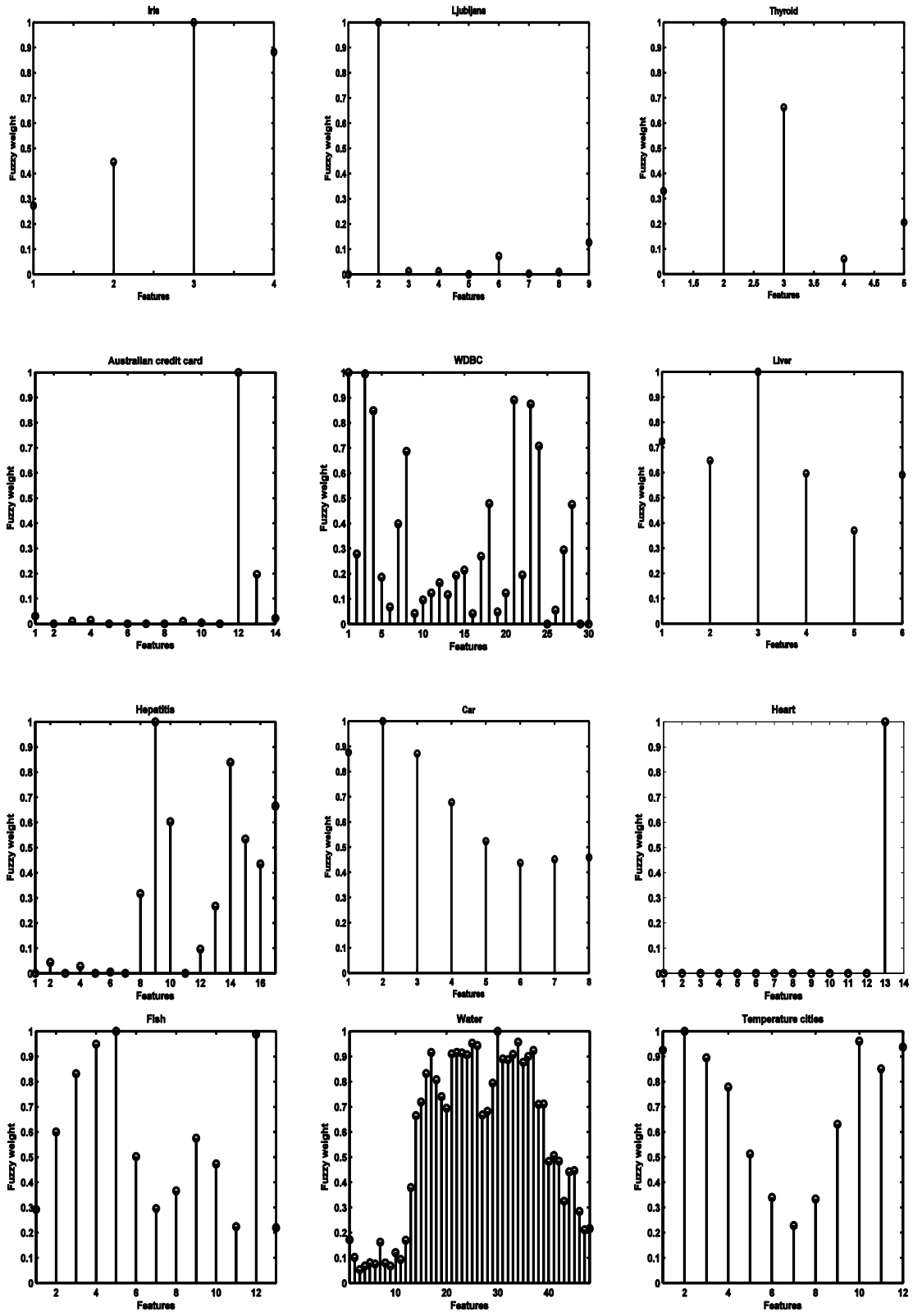


Fig. 6.2. Fuzzy feature weights resulted by WFCA

6.5 Conclusion

In this chapter we proposed a novel approach based on online feature weighting for clustering of heterogeneous data. The proposed algorithm is an extension of our previously developed supervised feature weighting algorithm. To cope with the problem of data heterogeneity, the SMSP principle is extended here also to reason in a unified way about heterogeneous data in an unsupervised framework. However, we have shown that the mapping step should be performed in an incremental fashion to take into account a new pattern at each iteration of the learning process. In this order, an iterative version of the mapping function introduced in the supervised case has been provided according to each feature type.

First, we described separately the online incremental clustering algorithm based on a fuzzy rule-based system. Then, we investigated as for the supervised context, the integration of the feature weighting task in the clustering process to design our proposed approach based on fuzzy weighted rule concept. An extensive experimental study has been then performed on artificial and real-world problems to prove the effectiveness of the proposed approach. This algorithm fails however to scale with high dimensional data (e.g. microarray data) characterized by a huge number of irrelevant features.

CHAPITRE 7- Résumé
Application au cancer du sein

Avant l'ère des biopuces, la gestion du cancer a été guidée uniquement par les connaissances cliniques et histo-pathologiques acquises durant plusieurs décennies de recherche sur le cancer. Cependant, la forte mortalité par le cancer du sein a poussé les chercheurs à rechercher de nouveaux outils de pronostic du cancer plus précis aidant les médecins à prendre les décisions de traitement nécessaire et réduire ainsi les frais médicaux. Pendant la dernière décennie, l'analyse par biopuces a eu un grand intérêt dans la gestion du cancer tels que le diagnostic (Ramaswamy et al., 2001), le pronostic (Van't Veer et al., 2002), et la prédiction de la réponse au traitement (Straver et al., 2009). Cependant, l'introduction de cette technologie a apporté avec elle de nouveaux défis tels que la dimension élevée en termes de nombre de marqueurs et un ratio bruit/signal élevé. Dans ce chapitre, quelques applications au problème du cancer du sein en utilisant les approches proposées dans les chapitres précédents ont été présentées. Nous nous sommes concentrés surtout sur le pronostic du cancer du sein et la prédiction de la réponse au traitement comme des tâches primordiales pour l'amélioration de la vie des patientes atteintes du cancer, en se basant sur les données cliniques et/ou données de biopuces. Les applications sont appuyées par des analyses statistiques diverses et des interprétations biologiques sur la base des connaissances actuelles.

D'abord une application sur le pronostic du cancer basée uniquement sur des données hétérogènes cliniques a été effectuée. Grâce à cette application, nous avons montré que l'approche de pondération des variables floue sélectionne des facteurs cliniques significatifs. Deux autres approches de sélection de variables ont été testées sur le même problème, afin de comparer les performances de la méthode que nous avons développée.

Dans la deuxième application, le pronostic du cancer est basé uniquement sur des données issues de biopuces pour extraire une signature de pronostique constituée de 20 gènes. Les résultats obtenus en utilisant plusieurs critères de comparaison montrent que la valeur prédictive de cette signature de pronostique peut être supérieure à celle d'autres signatures de pronostique existantes et les facteurs cliniques classiques. En particulier, la signature de 20 gènes améliore significativement la spécificité de l'une des approches génétiques bien connues (signature des 70 gènes dite « d'Amsterdam »).

La troisième application a été consacrée à étudier l'intégration des données cliniques aux données de biopuce. Dans de telles applications, les problèmes d'hétérogénéité des données et la dimensionnalité élevée doivent être confrontés conjointement. Nous avons profité de la propriété intéressante de l'approche proposée qui permet de gérer simultanément les deux problèmes pour extraire une signature pronostique hybride. Nous avons montré ensuite à travers quelques analyses que l'intégration des approches peut améliorer le pronostic du cancer du sein. En particulier, la signature hybride améliore la sensibilité de la signature des 20 gènes, tout en maintenant une spécificité comparable.

Pour défier le problème du faible rapport signal/bruit dans les données de biopuces pour le pronostic du cancer, une approche symbolique a été considérée pour extraire une signature de pronostique plus robuste, dénommé ici GenSym. Nous avons décrit d'abord la génération de la base de données intervallaires par le remplacement de l'expression de chaque gène par un intervalle en y incorporant un bruit blanc gaussien avec un ratio signal/bruit spécifique. Nous avons montré à travers quelques expériences et analyses statistiques que la signature GenSym peut surpasser les autres approches existantes. En particulier, elle permet de conserver la bonne sensibilité apportée par la signature hybride tout en améliorant la bonne spécificité de la signature des 20 gènes. Par ailleurs, la liste des gènes de cette signature comporte des gènes significatifs liés à l'invasion, le cycle cellulaire et la prolifération. Nous croyons que cette première tentative dans cette direction a également ouvert la porte à la communauté d'apprentissage automatique pour développer d'autres approches afin de résoudre ce problème.

La dernière application concerne le problème de la prédiction de la réponse à un traitement néoadjuvant pour des patientes atteintes du cancer du sein avec HER2 surexprimé. Grâce à l'approche proposée, une signature constituée de quatre marqueurs (PTEN, HER2, eI4E, EGFR) a été extraite, qui améliore significativement le pouvoir discriminant entre les deux groupes des répondeurs positifs et négatifs comparé à celui obtenu avec la signature de 2 marqueurs utilisés habituellement (PTEN, HER2). En particulier, la combinaison de 4 marqueurs améliore significativement la spécificité de la combinaison 2-marqueurs. Ceci souligne l'importance de deux nouveaux facteurs prédictifs (eI4E, EGFR) pour améliorer la précision de la prédiction de la réponse à un traitement néoadjuvant pour des patientes atteintes de cancer du sein avec HER2 surexprimé.

CHAPTER 7
Breast Cancer Applications

During the pre-microarray era cancer management was guided only by the clinical and histo-pathological knowledge gained from many decades of cancer research. However, the high mortality from breast cancer has pushed researcher to seek for accurate cancer prognosis tools that help physicians to take the necessary treatment decisions and thereby reduce its related expensive medical costs. In the past decade microarray analysis has had a great interest in cancer management such as diagnosis (Ramaswamy *et al.*, 2001), prognosis (Van't Veer *et al.*, 2002), and treatment response prediction (Straver *et al.*, 2009). However, the introduction of this technology has brought with it new challenges such as high feature-to-sample and noise-to-signal ratios. In this chapter we present some breast cancer applications based on our proposed approaches in previous chapters. We focus particularly on breast cancer prognosis and treatment benefit prediction based on clinical and/or microarray data. The applications are supported by various statistical analysis and biological interpretations based on the current knowledge.

7.1 Cancer prognosis based on clinical and/or microarray data

7.1.1 Cancer prognosis application based on clinical information

a- Ljubljana Prognosis Dataset

The dataset used here concerns the Ljubljana breast cancer prognosis dataset; it contains a total of 286 patients where 201 have not relapsed after five years and 85 have relapsed (Blake and Merz, 1998). Patients with missing data were excluded from this study (9 patients). All patients are described by 9 features (6 qualitative and 3 interval type):

(a) Menopause: >40, <40, pre-menopause.
(b) Ablation Ganglia: yes, no.
(c) Malignancy Degree (Grade): I, II, III
(d) Breast: right, left
(e) Quadrant: sup. left, inf. left sup. right, inf. right, center.
(f) Irradiation: yes, no
(g) Age: 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99

(h) Tumor Size: 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59.

(i) Involved Nodes: 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39.

b- Experimental setup and results

1. Factor selection for cancer prognosis within supervised context

Choosing accurately the powerful prognostic factors among the nine features is of big interest as it can help the physician to predict, based only on those factors, whether a patient will relapse. In this aim, the proposed reasoning tool in chapter 5, referred to as Membas, is used to identify the set of important factors for this problem. We compared then the proposed approach with some existing feature selection approaches: Neighborhood Rough Set (NRS) (Hu *et al.*, 2008) and Simba (Gilad-Bachrach *et al.*, 2004). Indeed, NRS is a heterogeneous feature subset selection based on neighborhood rough set concept and Simba is based on 1-NN rule. To assess the robustness of each method against irrelevant features, 50 random quantitative features were also added. To analyze the importance of each selected feature, the weights obtained respectively by Membas and Simba in a random realization have been plotted for respectively each feature in Figure 7.1. NRS ranks the features based on a dependency measure shown also in Figure 7.1. It can be observed for Membas that only four of the nine features have a significant weights and the others seems to be weakly relevant. The order of the most relevant features by Membas appears to be:

1- “Involved Nodes” (interval feature type),
2- “Ablation ganglia” (qualitative feature type),
3- “Grade” (qualitative feature type),
4- “Irradiation" (qualitative feature type).

In addition, the proposed mechanism succeeds to identify the 50 added irrelevant features by assigning them approximately zero weights (they correspond to the last 50 features in Figure 7.1 (left)). Furthermore, the two features selected by Membas (“Involved Nodes” and “Grade”) are still considered as important prognostic factors in day-to-day clinical practice (Deepa *et al.*, 2005). Obviously, the selection of the two additional factors (“Ablation ganglia” and “Irradiation") suggests that these treatments have influenced the breast cancer evolution and therefore the prognosis outcome. On the other hand the optimal set of feature selected by Simba contains many irrelevant features (first top ranked is irrelevant). Moreover, Figure 7.1 (center) shows that only two among the five top ranked features are relevant which

are “Involved nodes” and “Quadrant”. However, only one feature among the nine presumably useful features has been deemed important by NRS as shown by Figure 7.1 (right).

In order to assess the relevance of selected factors to improve the cancer prognosis task, we compared the three feature selection methods on two classifiers: the fuzzy reasoning tool LAMDA and *k*-NN. The same procedure of cross-validation and statistical variation elimination as in section 5 of the fifth chapter, is adopted here. Figure 7.2 and 7.3 show the obtained classification error with respectively LAMDA and *k*-NN approaches as a function of the top ranked features by MEMBAS, NRS and Simba.

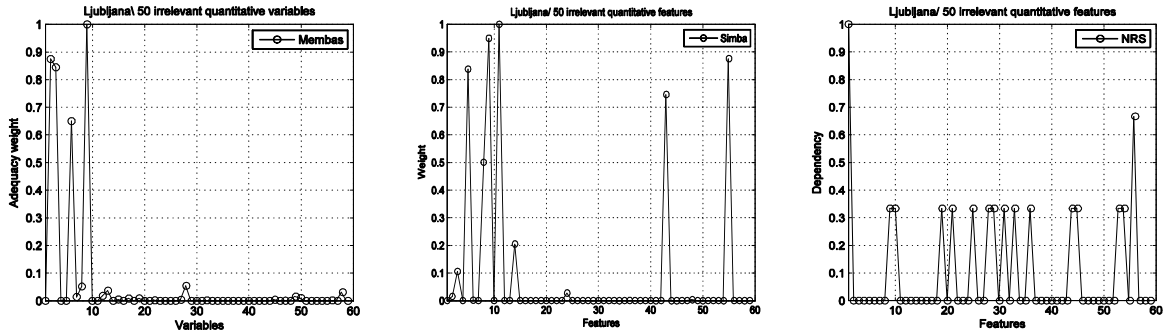


Fig. 7.1. (left) Feature weights by Membas; (center) Feature weights by Simba; (right) Dependency by NRS

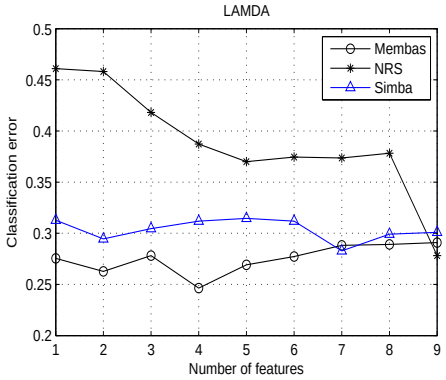


Fig. 7.2. Classification error by LAMDA as function of top ranked features using Membas, NRS and Simba

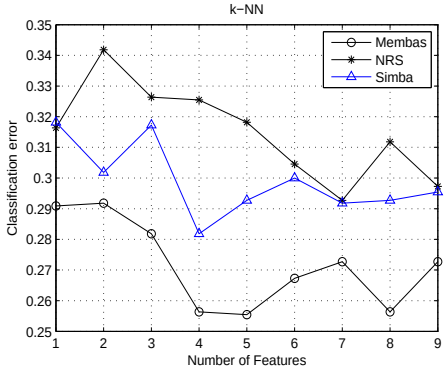


Fig. 7.3. Classification error by *k*-NN as function of top ranked features using Membas, NRS and Simba

It can be observed that MEMBAS performs best on this dataset regardless of the used classifier (LAMDA or k-NN), Simba the second and NRS the worst. Interestingly, the

minimal classification error on both classifiers corresponds to the resulted four top ranked features by MEMBAS (Figure 7.1), provided that the error difference on *k*-NN with five or four features is insignificant (Figure 7.2 and 7.3). Table 7.1 summarizes the optimal obtained averaged classification errors with the two classifiers and the corresponding optimal number of selected features by each feature selection method. It can be observed that the best couple (classification performance, number of selected features) is obtained by Membas on both classifiers. Furthermore, a student t-test was performed to assess if the classification error comes from the same distribution. At a level of p-value= 0.05, Membas wins against NRS and Simba whatever the used classifier (Table 7.1). It must be noted also that Figure 7.2 and 7.3 show that LAMDA outperforms *k*-NN on this heterogeneous dataset almost over all the rang of feature subsets.

Table 7.1. Classification error (%) and corresponding optimal number of factors on Ljubljana dataset

Method	**Membas**	**NRS**	**Simba**	**p-value (Membas-NRS)**	**p-value (Membas -Simba)**
LAMDA	**24.64 (4)**	27.82 (9)	28.27 (7)	**1.79e-005**	**1.97e-004**
k-NN	**25.55 (5)**	29.27 (7)	28.18 (4)	**1.34e-005**	**4.43e-004**

2. Unsupervised learning

An independent clustering using the fuzzy reasoning tool proposed in section 6.2 of the sixth chapter has been also performed on this dataset and the obtained 2-cluster partitions are compared with the 2-clusters known a priori. The obtained clustering error is given in Table 7.2 with nine features. To further demonstrate the performance of the proposed methodology for clustering, we compared it with the well known Fuzzy c-means clustering (FCM) method (Bezdek, 1981). It must be noted that FCM does not handle neither qualitative nor interval data. However, as interval features in this dataset are regulars (interval features take their values from an accountable set of interval values), their transformation into quantitative values is straightforward to enable handling them by FCM. A similar procedure is adopted for the transformation of qualitative data. Results obtained with FCM are reported also in Table 7.2. It can be observed that the proposed approach outperforms the FCM on this specific problem of cancer prognosis. One possible explanation is that the transformation of qualitative and interval spaces into a quantitative space required for FCM leads probably to information loss whereas the proposed approach based on SMSP principle allows handling appropriately the qualitative and interval data.

Table 7.2. Clustering error for Ljubljana dataset

Method	Number of clusters	Accuracy
LAMDA-cluster	2	**22.74%**
FCM	2	28.88%

Furthermore, to show the effectiveness of the proposed approach, we analyzed the obtained prototypes of each class which correspond to the parameters of the fuzzy features partition resulted at the end of the clustering task. The obtained class parameters for the three interval features are shown in Figure 7.4. It can be observed that the interval feature "Involved Nodes" has the most discriminatory power between the two classes which may be considered as a confirmation of its selection as an important predictive factor in the previous section (top ranked in Figure 7.1). Nonetheless, that does not mean that the other two interval features are not useful but their relevance is weaker for this specific problem as it can be seen in Figure 7.1. The prototypes of the other three qualitative features: Ablation ganglia, Malignancy Degree, Irradiation are also shown in Figures 7.5, 7.6 and 7.7 respectively. Interestingly, the three features exhibit also an important discrimination power between classes. These results are in complete agreement with the selection of the three qualitative features by the fuzzy mechanism proposed in chapter 5 (see Figure 7.1).

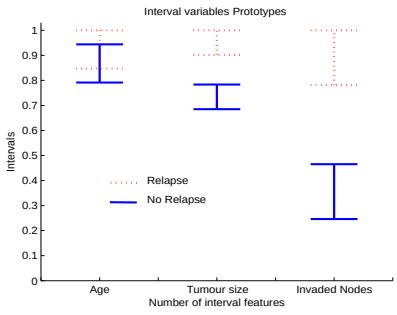


Fig. 7.4. Class prototypes obtained by clustering for interv. features "Age, Tumour size, Invaded Nodes"

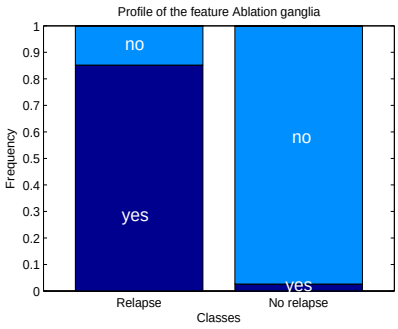


Fig. 7.5. Class prototypes obtained by clustering for qual. feature "Ablation ganglia"

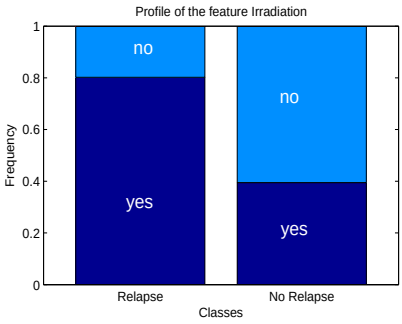


Fig. 7.6. Class prototypes obtained by clustering for qual. feature "Irradiation"

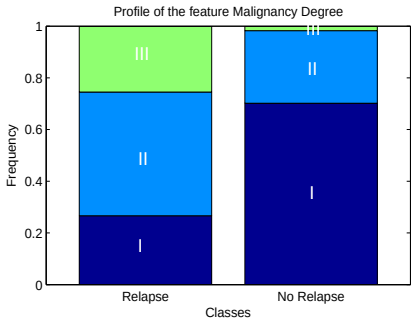


Fig. 7.7. Class prototypes obtained by clustering for qual. feature "Malignancy degree"

Let's now compare these class prototypes with those obtained when a supervised learning is considered. Using the fuzzy rule based classifier LAMDAwe obtained the prototypes shown in figures (7.8 to 7.11) for respectively all features. It can be observed that the profiles in both cases are quite similar.

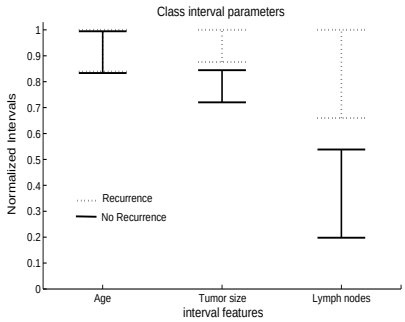


Fig 7.8. Class prototypes obtained by classification for for interv. features "Age, Tumor size, Involved Nodes"

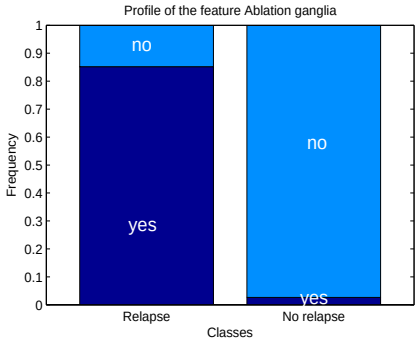


Fig. 7.9. Class prototypes obtained by classification for qual. feature "Ablation ganglia"

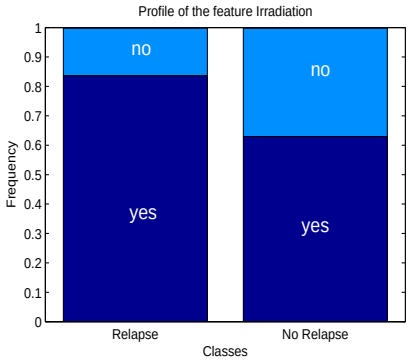


Fig. 7.10. Class prototypes by classification for qual. feature "Irradiation"

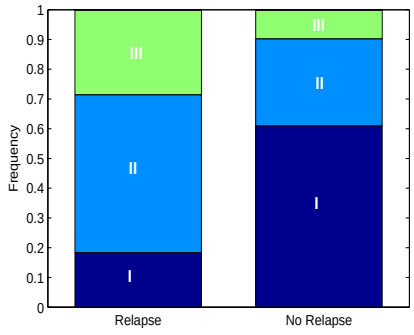


Fig. 7.11. Class prototypes obtained by classification qual. feature "Malignancy degree"

7.1.2 Cancer prognosis application based on microarray data

Recent studies have demonstrated the potential value of gene expression signatures in assessing the risk of post-surgical disease. In this study we focus on the use of our proposed approaches for gene signature derivation for cancer prognosis.

a- Dataset and experimental setup

The study is performed using the well-known Van't Veer dataset (Van't Veer *et al.*, 2002). Van't Veer and colleagues used a dataset containing 78 sporadic lymph-node-negative patients younger than 55 years of age and less than 5 cm in tumour size, to derive a prognostic signature in their gene expression profiles. Forty-four patients remained disease-free after their initial diagnosis for an interval of at least 5 years (good prognosis group), and 34 patients had developed distant metastases within 5 years (poor prognosis group). We use the same group of patients in the aim to derive a gene prognostic signature. Patients with missing data

(1 poor prognosis patient) were excluded in our study. We use our feature selection approach described in chapter 5, referred to as MEMBAS, to build a computational model that accurately predicts the risk of distant recurrence after 5-years period of breast cancer diagnosis. Due to the small sample size we performed a LOOCV (Leave One-Out Cross Validation) to estimate the optimal classification parameters. At each iteration of this procedure one sample is held-out for testing and the remaining samples are used for training. The training data are used to estimate the optimal parameters of the classifier and to perform the feature selection task. The resulting model is employed then to classify the held-out sample. This experiment is repeated until each sample has been used for testing. In this study we used LAMDA classifier for which only one parameter needed to be specified in the training phase (exigency index). It is worthwhile to note here that in the study performed by Van't Veer and its colleagues, a 70-gene signature has been derived from the same dataset using a feature selection method based on correlation coefficient.

We demonstrate the predictive values of the gene signature derived using Membas on this microarray dataset by comparing its performance with those of the clinical markers, 70-gene signature, St Gallen and NIH criterions. The performances are also estimated through a LOOCV procedure.

b- Results

A 20-gene signature was derived based on Membas approach corresponding to the optimal classification performance using LAMDA classifier based on the Guassian-like membership function. Classification performance obtained based on this signature with LAMDA are reported in Table 7.3. For comparison, classification performance using the 70-gene signature, the clinical markers, the St-Gallen consensus and the NIH criterion using LAMDA classifier are also reported in Table 7.3. We observe that the 20-gene signature outperforms the 70-gene, clinical and classical clinical criterions (St-Gallen, NIH). Particularly, the 20-gene signature improves significantly the specificity of 70-gene signature while assuring a comparable sensitivity.

Tab. 7.3. Classification performance using 20-gene signature, 70-gene signature, all clinical markers, St Gallen consensus and NIH criteria

Method	TP	FP	FN	TN	sensitivity	Specificity	Accuracy
20-gene	28/33	5/44	6/33	38/44	**82.35**	**88.37**	**85.71**
70-gene	27/33	9/44	6/33	35/44	81.82	79.55	80.52
Clinical	26/33	14/44	7/33	30/44	78.79	68.18	72.73
St-Gallen	33/33	39/44	0/33	6/44	100	6.49	50.65
NIH	33/33	44/44	0/33	0/33	100	0	42.86

Classification performance is not always a sufficient criterion for comparing predictive value of marker signatures. Performance measurement can also depend strongly on a decision threshold when only a limited number of patients are available. Varying this decision threshold enables to visualize the performance of a given classifier over all sensitivity and specificity levels through a Receiver Operating Characteristic (ROC) curve (See Appendix 4). To further demonstrate the superiority of the 20-gene signature, we decided to plot in Figure 7.12 also the ROC curve of the three models based respectively on the 20-gene signature, 70-gene signature and clinical markers. The obtained ROC curve confirms the outperformance of the 20-gene signature over other signatures.

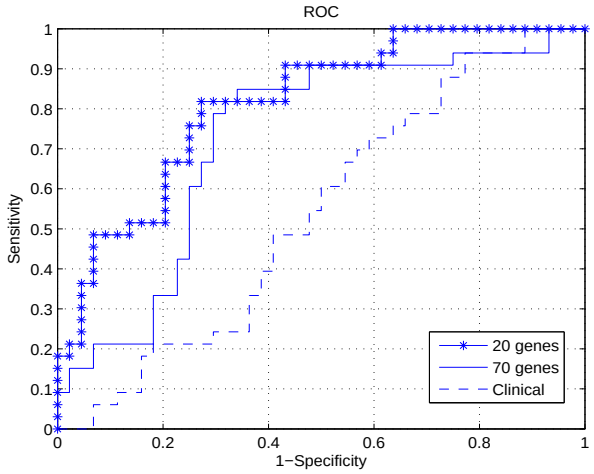


Fig. 7.12. ROC curve of clinical, 20-gene and 70-gene signatures.

We perform also survival data analysis of the four approaches, 20-gene signature, 70-gene signature, clinical markers and St-Gallen criterion, to further demonstrate the prognostic value of the 20-gene signature. The St-Gallen and NIH criteria are not shown here since the good prognosis group contains very few patients. The Kaplan-Meier curves with 95% confidence intervals of respectively the four approaches are shown in Figure 7.13. Particularly the 20-gene signature induces a significant difference in the probability of remaining metastases-free in patients with a good signature and the patients with a poor prognostic signature (P-value<0.001). Hazard ratio estimated by Mantel-Cox approach of distant metastases within five years for the 20-gene signature is 7.6 (95% CI: 3.86- 15.06), which is superior to either the 70-gene, St Gallen consensus or clinical markers.

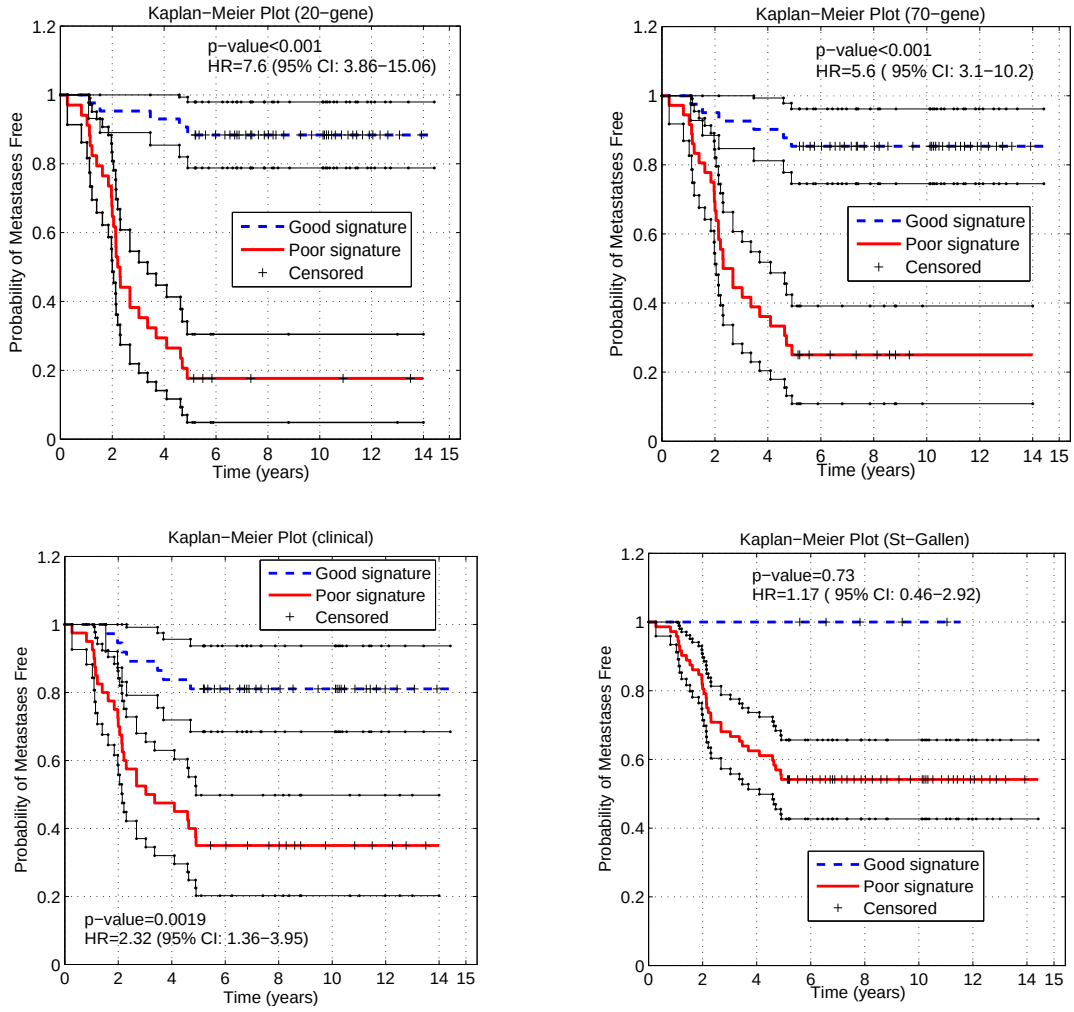


Fig. 7.13. Kaplan-Meier estimation of the probabilities of remaining metastases-free for the good and poor prognosis groups. The p-value is computed by using log-rank test.

c- Analysis of the twenty-gene signature

Among the 20-gene signature, given in Table 7.4, eight genes are listed in the 70-gene signature and both gene signatures share the first gene (AL080059). Note that the number of genes derived is significantly short compared to the number required to perform the caner prognosis task using Amsterdam 70-gene signature. A brief description of the biological implication of gene is provided in Table 7.4 according to the National Center for Biotechnology Information (NCBI) databases.

Tab. 7.4: Notation and description of 20-gene signature.

Rank	Gene ID	70-gene	Notation	Description
1	AL080059	■	TSPYL5	A subsequent analysis has revealed a significant homology with human protein factors, including NAPs, which play a role in DNA replication and thereby proliferation (Schnieders et al., 1996). It is thought that NAPs act as histone chaperones shuttline histone proteins involved in regulating chromatin structure and accessibility and therefore can impact gene expression.

2	NM_003748	■	ALDH4A1	This protein belongs to the aldehyde dehydrogenase family of proteins. This enzyme is a mitochondrial matrix NAD-dependent dehydrogenase which catalyzes the second step of the proline degradation pathway, converting pyrroline-5-carboxylate to glutamate. Deficiency of this enzyme is associated with type II hyperprolinemia, an autosomal recessive disorder characterized by accumulation of delta-1-pyrroline-5-carboxylate (P5C) and proline. Alternatively spliced transcript variants encoding different isoforms have been identified for this gene.
3	NM_020974	■	SCUBE2	SCUBE2 signal peptide, CUB domain, EGF-like 2 [Homo sapiens]. The SCUBE2 (known also as CEPG1) is located on human chromosome 11p15 and has homology to the achaetescute complex (ASC)of genes in the basic helix-loop-helix (bHLH) family of transcription factors.
4	D42044	□	KIAA0090	Protein binding
5	NM_006681	■	NMU	NMU neuromedin U [*Homo sapiens*]
6	NM_006544	□	EXOC5	Exocyst complex component 5 [Homo sapiens].The protein encoded by this gene is a component of the exocyst complex, a multiple protein complex essential for targeting exocytic vesicles to specific docking sites on the plasma membrane. Though best characterized in yeast, the component proteins and functions of exocyst complex have been demonstrated to be highly conserved in higher eukaryotes. At least eight components of the exocyst complex, including this protein, are found to interact with the actin cytoskeletal remodeling and vesicle transport machinery. The complex is also essential for the biogenesis of epithelial cell surface polarity.
7	Contig14882_RC	□	N\A	N\A
8	Contig20217_RC	■	N\A	N\A
9	Contig37063_RC	□	N\A	N\A
10	NM_019028	□	ZDHHC13	Zinc finger, DHHC-type containing 13 [Homo sapiens]
11	NM_003450	□	ZNF174	Zinc finger protein 174 [Homo sapiens]
12	Contig54742_RC	□	N\A	N\A
13	Contig63649_RC	■	N\A	N\A
14	Contig42933_RC	□	N\A	N\A
15	NM_004994	■	MMP9	Matrix metallopeptidase 9 (gelatinase B, 92kDa gelatinase, 92kDa type IV collagenase) [Homo sapiens]. Proteins of the matrix metalloproteinase (MMP) family are involved in the breakdown of extracellular matrix in normal physiological processes, such as embryonic development, reproduction, and tissue remodeling, as well as in disease processes, such as arthritis and metastasis. Most MMP's are secreted as inactive proproteins which are activated when cleaved by extracellular proteinases. The enzyme encoded by this gene degrades type IV and V collagens. Studies in rhesus monkeys suggest that the enzyme is involved in IL-8-induced mobilization of hematopoietic progenitor cells from bone marrow, and murine studies suggest a role in tumor-associated tissue remodeling.
16	NM_000286	□	PEX12	Peroxisomal biogenesis factor 12 [Homo sapiens]. This

				gene belongs to the peroxin-12 family. Peroxins (PEXs) are proteins that are essential for the assembly of functional peroxisomes. The peroxisome biogenesis disorders (PBDs) are a group of genetically heterogeneous autosomal recessive, lethal diseases characterized by multiple defects in peroxisome function. The peroxisomal biogenesis disorders are a heterogeneous group with at least 14 complementation groups and with more than 1 phenotype being observed in cases falling into particular complementation groups. Although the clinical features of PBD patients vary, cells from all PBD patients exhibit a defect in the import of one or more classes of peroxisomal matrix proteins into the organelle. Defects in this gene are a cause of Zellweger syndrome (ZWS).
17	Contig6238_RC	□	N\A	N\A
18	NM_014489	□	PGAP2	Post-GPI attachment to proteins 2 [Homo sapiens] .
19	NM_002779	□	PSD	Pleckstrin and Sec7 domain containing [Homo sapiens]
20	Contig32185_RC	■	N\A	N\A

■: Listed in 70-gene signature, □: Not listed in 70-gene signature

The functional annotation for the genes should provide insight into the underlying biological mechanism leading to rapid metastases. Among the 20-gene signature, genes involved in proliferation, invasion and metastasis are significantly unregulated in the metastasis group. For instance we find SCUB2 which has been revealed to play important roles in development, inflammation and perhaps carcinogenesis (Yang *et al.*, 2002). The expression of SCUBE2 gene has been found to be associated with ER status in a recent SAGE-based study of breast cancer specimens (Abba *et al.*, 2005). It has been reported recently that SCUBE2 suppresses breast tumor cell proliferation and confers a favorable prognosis in invasive breast cancer (Cheng *et al.*, 2009). TSPYL5 is involved in modulation of cell growth and cellular response probably via regulation of the akt signaling pathway. It is reported that TSPYL5 is a poor prognosis marker and reduces the p53 protein levels and inhibits activation of p53-target genes. It is known that EXOC5 gene is related to cell mobility and invasion. MMP-9 are related to tumor invasion and metastasis by their capacity for tissue remodeling via extracellular matrix as well as basement membrane degradation and induction of angiogenesis. Evaluation of MMP-9 expression seems to add valuable information on breast cancer prognosis. The KIAA0090 is one of the breast cancer markers identified in (Dettling *et al.*, 2005).

7.1.3 Hybrid Signature derivation by integrating clinical and microarray data for cancer prognosis

In the past decade microarray analysis has had a great interest in cancer management. Meanwhile, clinical and histo-pathological factors are still considered as valuable tool to make day-to-day cancer management decisions. It has been however established recently that the integration of both information may improve cancer management (Sun *et al.*, 2007a; Gevaert *et al.*, 2006). In (Sun *et al.*, 2007a) a feature selection method (I-Relief) was used to perform markers selection. However, the used method works under the assumption that all the data are of quantitative type and therefore a transformation of symbolic data to quantitative one was performed to cope with data heterogeneity. This transformation can be a source of distortion and information loss as it introduces a distance which was not present in the original data. In (Gevaert *et al.*, 2006), a Bayesian network was used to perform breast cancer prognosis. The obtained results show only that their approach performs similarly to the 70-gene signature established by Van't Veer and colleagues (Van't Veer *et al.*, 2002) and claim that a feature selection is implicitly performed based on their (in) dependency through the Markov Blanket concept. These results do not mean necessarily that the clinical data contains no additional information to the genetic data; it only tells us that their approach does not fit well (Sun *et al.*, 2007a). In the present study, we use our hybrid feature selection method, referred to as MEMBAS, to assess the usefulness of the integration of both types of data by addressing both challenges simultaneously: high-dimensionality and heterogeneity of data (Hedjazi *et al.*, 2011d).

a- Dataset and experimental setup

We use here also the Van't Veer data set of 78 patients (Van't Veer *et al.*, 2002) to derive a hybrid signature by integrating clinical and microarray data. The clinical data contains eight features:

a) Age (quantitative)

b) Tumour grade (interval:[3,5]~ Grade I; [6,7] ~ Grade II; [8,9] ~ Grade III)

c) Tumour size (quantitative: mm)

d) Estrogen Receptor expression (quantitative: intensity)

e) Progesterone Receptor expression (quantitative: intensity)

f) Angioinvasion (qualitative: 'yes' or 'no')

g) Lymphocytic Infiltrate (qualitative: 'yes' or 'no')

h) BRCA1 mutation (qualitative: ‘yes’ or ‘no’)

The same LOOCV procedure employed in the previous study was adopted here to perform feature selection and learn classifier parameters, and then testing the performance on a hold-out sample not used for training. The classification task was performed by using the fuzzy classifier LAMDA. MEMBAS based on the binomial membership function is used here to derive a hybrid prognostic marker without resorting to any data transformation. To demonstrate the predictive power of the hybrid prognostic signature derived from the genetic and clinical markers, its performance was compared also with those of clinical markers and the well known Amsterdam 70-genes signature (Van’t Veer *et al.*, 2002). Another comparison with purely clinical indices (NIH, St Gallen) was also performed.

b- Results

Table 7.5 shows the obtained comparative results between the hybrid markers approach and other approaches. It can be observed that the best overall prediction accuracy is obtained by the proposed approach which achieves more than 87%. Particularly, the hybrid signature provides an improved specificity compared to the 70-gene signature while maintaining a relatively high sensitivity (~88%). If we compare to the 20-gene signature, the hybrid signature maintains an improved overall accuracy while gaining in sensitivity (2 more poor-prognosis patients have been correctly identified). A comparison with clinical conventional prognostic factors (St. Gallen’s and NIH) is also reported in Table 7.5. Both indices have a very high sensitivity, but an intolerable low specificity which would lead to give unnecessary adjuvant systematic treatment to almost all patients. Thus the obtained hybrid markers outperforms also the pure clinically indices.

It must be noticed here that MEMBAS selects only 15 hybrid markers, among them three are mixed-type clinical markers (Angioinvasion “qualitative” , Grade “interval” and Age “quantitative), added to them 12 genes as listed in Table 7.6.

Tab. 7.5: Comparatives results between hybrid, clinical and genetic signatures.

Method	TP	FP	FN	TN	sensitivity	Specificity	Accuracy
Hybrid	29/33	6/44	4/33	38/44	**87.88**	86.36	**87.01**
70-gene	27/33	9/44	6/33	35/44	81.82	79.55	80.52
20-gene	28/33	5/44	6/33	38/44	82.35	**88.37**	85.71
Clinical	26/33	14/44	7/33	30/44	78.79	68.18	72.73
St-Gallen	33/33	39/44	0/33	5/44	100	6.49	50.65
NIH	33/33	44/44	0/33	0/33	100	0	42.86

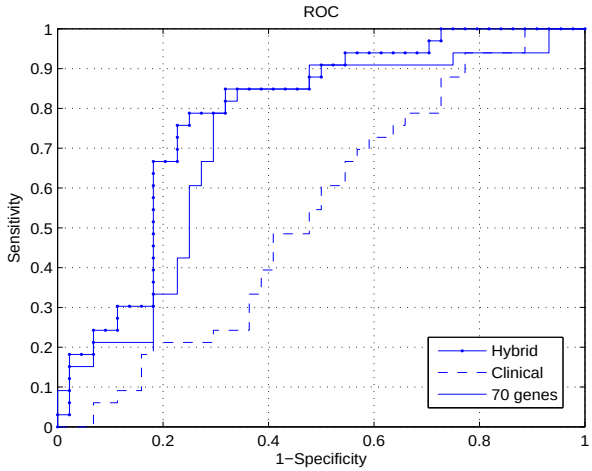


Fig. 7.14. ROC curve of hybrid, clinical and 70-gene signature.

For further comparison, we plotted the ROC curves for Hybrid, 70-gene signature and clinical markers. Figure 7.14 shows that the hybrid signature outperforms both the 70-gene signature and clinical markers.

To further demonstrate the prognostic value of the hybrid signature, we performed survival analysis using four approaches (hybrid signature, 70-gene signature, clinical markers and St-Gallen criterion). The Kaplan-Meier curves with 95% confidence intervals for respectively the four approaches are shown in Figure 7.15. Particularly, we can see that the hybrid signature induces a significant difference in the probability of remaining metastases-free in patients with a good signature and the patients with a bad prognostic signature (P-value<0.001). Hazard ratio estimated by Mantel-Cox approach of distant metastases within five years for the hybrid signature is 6.1 (95% CI: 3.22- 11.48), which is superior to either 70-gene and clinical markers.

a- Analysis of the Hybrid signature

Among the 12 genes of the hybrid signature, reported in Table 7.6, 4 genes are listed in the 70-gene signature and 4 in the 20-gene signature (with 2 in common). Note that the number of derived genetic markers is also significantly short compared to the number required to perform the cancer prognosis task using the 70-gene Amsterdam signature (12 Versus 70 genes). A brief description is provided about each marker in Table 7.6 according to the National Center for Biotechnology Information (NCBI) databases.

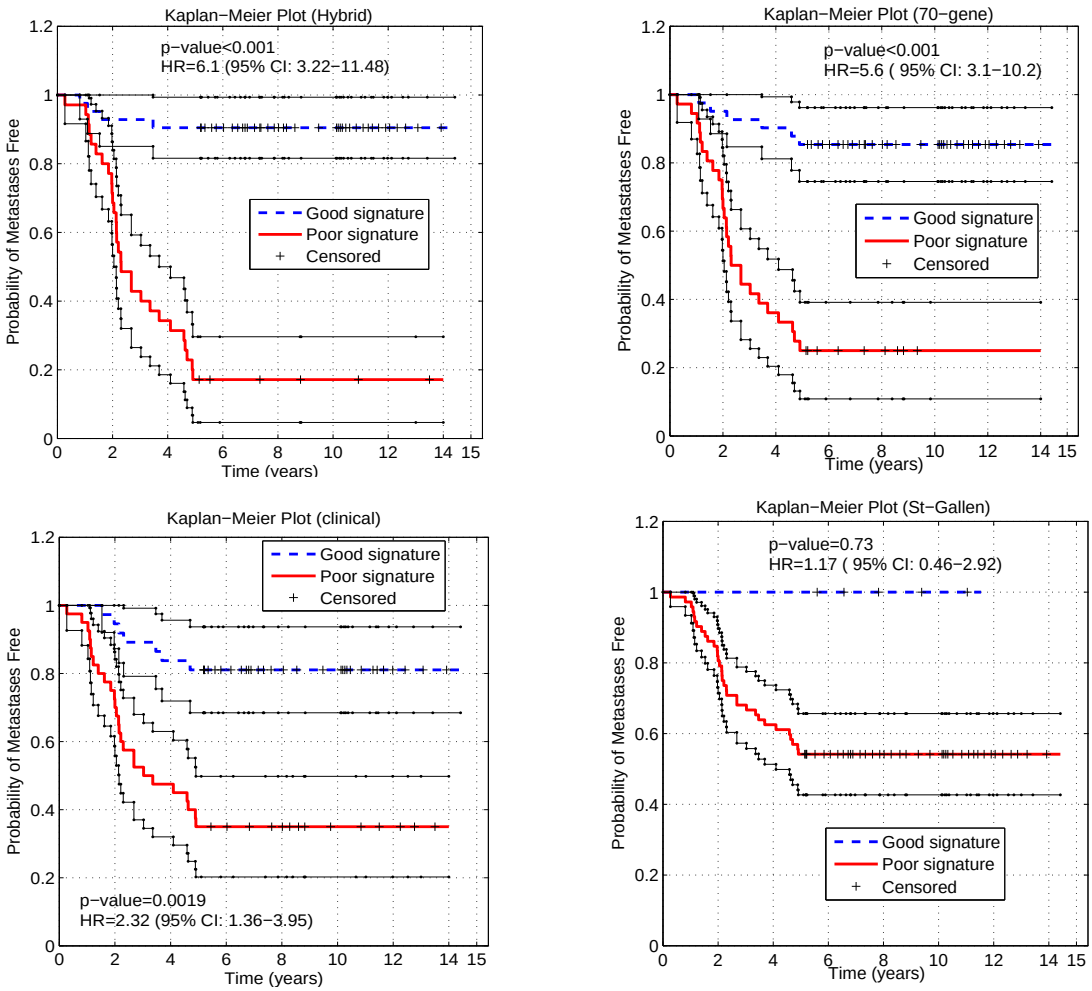


Fig. 7.15. Kaplan-Meier estimation of the probabilities of remaining metastases-free for the good and poor prognosis groups. The p-value is computed by using log-rank test.

Tab. 7.6: Notation and description of hybrid signature.

Rank	Gene ID	70-gene	20-gene	Notation	Description
1	Angioinvasion	-		N\A	N\A
2	Grade	-		N\A	N\A
3	Contig63649_RC	■		N\A	N\A
4	AL080059	■	x	TSPYL5	See Table 7.4
5	NM_006544	□	x	EXOC5	See Table 7.4
6	Contig55725_RC	■		N\A	N\A
7	NM_020974	■	x	SCUBE2	See Table 7.4
8	Age	-		N\A	N\A
9	NM_019028	□	x	ZDHHC13	See Table 7.4
10	NM_001787	□		LOC87720	Protein coding
11	AJ011306	□		EIF2B4	Eukaryotic translation initiation factor 2B, subunit 4 delta, 67kDa [Homo sapiens]. Eukaryotic initiation factor 2B (EIF2B), which is necessary for protein synthesis, is a GTP exchange factor composed of five different subunits. The protein encoded by this gene is the fourth, or delta, subunit. Defects in this gene are a cause of leukoencephalopathy with vanishing white matter (VWM) and ovarioleukodystrophy. Multiple

				transcript variants encoding different isoforms have been found for this gene.
12	NM_012429	□	SEC14L2	SEC14-like 2 (S. cerevisiae) [Homo sapiens]. This gene encodes a cytosolic protein which belongs to a family of lipid-binding proteins including Sec14p, alpha-tocopherol transfer protein, and cellular retinol-binding protein. The encoded protein stimulates squalene monooxygenase which is a downstream enzyme in the cholesterol biosynthetic pathway. Alternatively spliced transcript variants encoding different isoforms have been identified for this gene.
13	Contig14882_RC	□	N\A	N\A
14	Contig47042	□	N\A	N\A
15	NM_005176	□	ATP5G2	ATP synthase, H+ transporting, mitochondrial Fo complex, subunit C2 (subunit 9) [*Homo sapiens*]. This gene encodes a subunit of mitochondrial ATP synthase. Mitochondrial ATP synthase catalyzes ATP synthesis, utilizing an electrochemical gradient of protons across the inner membrane during oxidative phosphorylation. ATP synthase is composed of two linked multi-subunit complexes: the soluble catalytic core, F1, and the membrane-spanning component, Fo, comprising the proton channel. The catalytic portion of mitochondrial ATP synthase consists of 5 different subunits (alpha, beta, gamma, delta, and epsilon) assembled with a stoichiometry of 3 alpha, 3 beta, and single representatives of the gamma, delta, and epsilon subunits. The proton channel likely has nine subunits (a, b, c, d, e, f, g, F6 and 8). There are three separate genes which encode subunit c of the proton channel and they specify precursors with different import sequences but identical mature proteins. The protein encoded by this gene is one of three precursors of subunit c. Alternatively spliced transcript variants encoding different isoforms have been identified. This gene has multiple pseudogenes.

■: Listed in 70-gene signature, □: Not listed in 70-gene signature, x: listed in 20-gene signature, -: Clinical markers.

Clinical markers included in the previously derived hybrid signature are “Angioinvasion”, “Grade” and “Age”. Interestingly, the two first markers have been also identified as important factors by similar studies (Sun *et al.*, 2007a; Gevaert *et al.*, 2006). The “Age” has also been identified by (Gevaert *et al.*, 2006) as a supplementary clinical marker which is still used in day-to-day clinical practices. Regarding genetic markers, we can find some of the genes included in the previously reported 20-gene signature (SCUBE2, TSPYL5, EXO5, ZDHHC13) and other new genes such as the Eukaryotic translation factor 2B EIF2B4 which is necessary for protein synthesis, ATP5G2 and cytosolic protein SEC14L2.

7.1.4 Symbolic gene selection to defy low signal-to-noise ratio for cancer prognosis

It has been reported recently that the major difficulties in deciphering high throughput gene expression experiments comes from the noisy nature of the data (Stolovitzky *et al.*, 2002). Data issued from high throughput technology indeed is not only characterized by the dimensionality problem but present also another challenging aspect related to its low signal-to-noise ratio. The noise in such type of data is multisource: biological and noise measurement, slide manufacturing errors, hybridization errors, scanning errors of hybridized slide (see section 2.4.3, chapter 2, for more details).

All existing feature and classification approaches assume that microarray data is perfect without wondering about its reliability. The lack of appropriate methods does not mean that machine learning approaches are unable to tackle such problems. An interesting approach for instance would be to use symbolic data analysis (SDA) (Bock and Diady, 2000) to model usually uncertainty and noise inherent to gene expression measurements by an interval representation (Billard, 2008). Symbolic interval features are extensions of pure real data types, in the way that each feature may take an interval of values instead of a single value (Gowda and Diady, 1992). In this framework, the value of a quantity x (e.g. gene expression value) is expressed as a closed interval $[x^-,x^+]$ whenever x is noisy or uncertain; representing the information that $x^- \leq x \leq x^+$. Therefore, what is really needed is an approach that enables to process efficiently high dimensional interval datasets. We take advantage here of our proposed approaches that support such requirements to derive a more robust gene signature for cancer prognosis from microarray datasets.

a- Dataset and experimental setup

We use here also the Van't Veer data set of 78 patients (Van't Veer *et al.*, 2002) to derive a signature for cancer prognosis. In order to take into account the uncertainty in gene expression measurements under the form of symbolic intervals, an appropriate setup should be followed. The m gene expression levels are initially represented in a matrix $X=[x_1,x_2,...,x_m]$ where m is the number of genes. The microarray interval dataset generation is performed by adding a white Gaussian noise with a specific Signal-to-Noise Ratio (SNR=3). Let's consider that the added white Gaussian noise has an absolute value b, then the j^{th} interval feature $y_j=[y_j^-, y_j^+]$ corresponding to the j^{th} gene having an expression x_j is generated as follows:

$$y_j^- = x_j - b$$
$$y_j^+ = x_j + b$$

It results that

$$y_j = [y_j^-, y_j^+] = [x_j - b, x_j + b].$$

At the end of this step the m gene expression levels are represented in a matrix $Y=[y_1,y_2,...,y_m]$ where y_j is an interval vector. Once the microarray interval dataset is obtained, our proposed approaches can be used to derive a genetic signature. To do so, we adopted similar a LOOCV procedure as previously to assess the predictive value of this symbolic gene signature, referred to here as GenSym.

a- Results

GenSym signature was derived based on the Membas approach corresponding to the optimal classification performance using the LAMDA classifier. We note that both of Membas and LAMDA enable to handle appropriately interval data for classification and feature selection (see previous chapters for more details). Table 7.7 shows the classification performance obtained with LAMDA using GenSym signature. For comparison, classification performance using 70-gene signature, clinical markers, St-Gallen consensus and NIH criterion are also reported in Table 7.7. We observe that the GenSym signature significantly outperforms the 70-gene, clinical and classical clinical criterions (St-Gellen, NIH). GenSym achieves indeed a high accuracy (~90%) while significantly improving specificity and sensitivity of the 70-gene signature (by more than 5 % and 10% respectively). Moreover, GenSym improves the sensitivity of the previously derived 20-gene signature and improves the specificity of the hybrid signature while maintaining the high sensitivity of the latter one.

Tab. 7.7: Comparatives results between GenSym, clinical and genetic signatures.

Method	TP	FP	FN	TN	sensitivity	Specificity	Accuracy
GenSym	29/33	4/44	4/33	40/44	**87.88**	**90.91**	**89.61**
70-gene	27/33	9/44	6/33	35/44	81.82	79.55	80.52
20-gene	28/33	5/44	6/33	38/44	82.35	88.37	85.71
Clinical	26/33	14/44	7/33	30/44	78.79	68.18	72.73
St-Gallen	33/33	39/44	0/33	5/44	100	6.49	50.65
NIH	33/33	44/44	0/33	0/33	100	0	42.86

For further comparison of the different approaches, we plotted in Figure 7.16 the ROC curves for GenSym, 20-gene, 70-gene and clinical approaches. It can be observed that the GenSym signature significantly outperforms the 20-gene and 70-gene signatures as well as clinical markers.

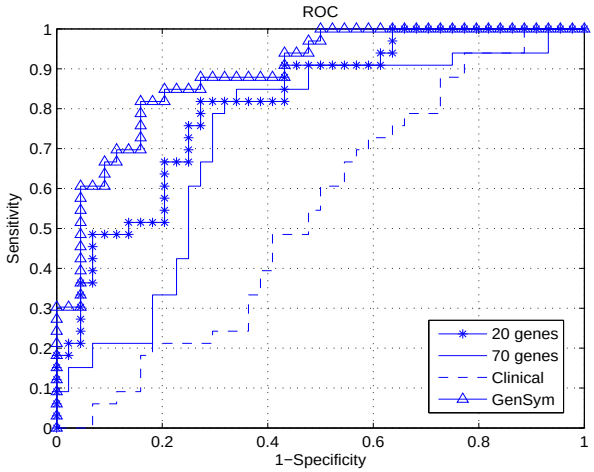


Fig. 7.16. ROC curve of GenSym, 20-gene, 70-gene, clinical approaches.

For more rigorous comparison, survival data analysis for the four approaches is also performed to further demonstrate the predictive value of the GenSym signature. The Kaplan-Meier curve with 95% confidence intervals of the GenSym signature, plotted in figure 7.17, exhibits a significant difference in the probability of remaining free of distant metastases in patients with a good signature and the patients with a poor prognostic signature (P-value<0.001). Hazard ratio estimated by Mantel-Cox approach of distant metastases within five years for the GenSym-23 signature is 8.20 (95% CI: 4.16- 16.2), which is superior to either 70-gene and clinical markers.

a- Analysis of GenSym signature

The GenSym signature is composed from 23 genes, given in Table 7.8, among them 12 genes are listed in the 70-gene signature. A brief description is provided about each gene in Table 7.8 according to the National Center for Biotechnology Information (NCBI) databases.

Additionally to the few genes identified in the previous signatures (TSPYL5, MMP9, NMU), GenSym signature holds many new meaningful genes (such as FBP1, IGFBP1, FGF18, SSX1, NUSAP1, C1GALT1, BTG2, PEX12). The importance of both (FBP1, IGFBP1) can be highlighted by the actually suspected relation between the insulin and tumor growth. But neither FBP1 nor IGFBP1 have been evaluated independently in human cancers. However, FBP1 have been also found strongly associated with disease outcome among the 231 top ranked genes in (Van't Veer *et al.*, 2002). FGF18 have been revealed clearly involved in the carcinogenesis of ~10% breast cancer. NUSAP1 has also been found to be related to proliferation and cells division. SSX1 is involved in certain sarcomas; it controls the cell cycle

and is considered as an important transcription factor. C1GALT1 is a protein that plays an important role in cell adhesion whereas BTG2 is considered as a tumor suppressor.

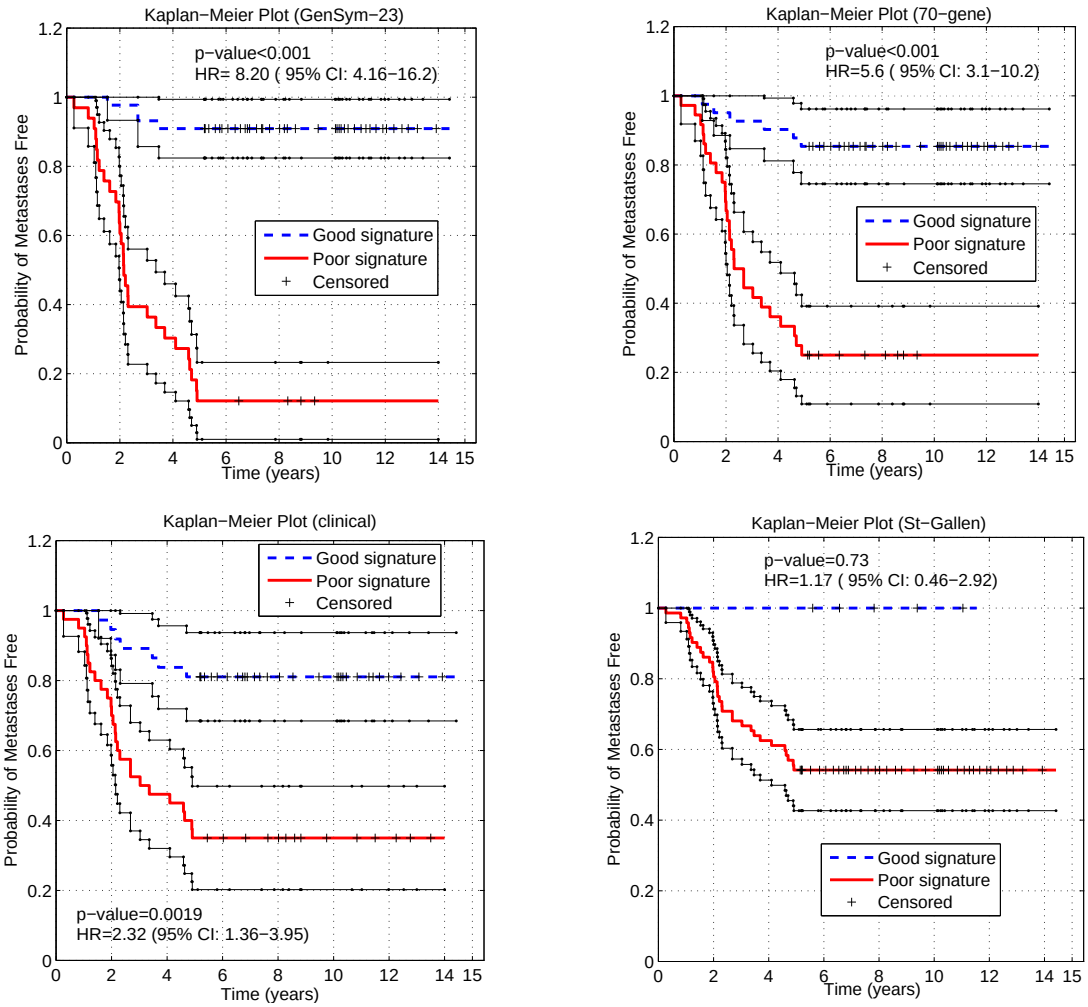


Fig. 7.17. Kaplan-Meier estimation of the probabilities of remaining metastases-free for the good and poor prognosis groups. The p-value is computed by using log-rank test.

Tab. 7.8: Notation and description of GenSym signature.

Rank	Gene ID	70-gene	Notation	Function
1	Contig37063_RC	□	N\A	N\A
2	Contig26388_RC	□	N\A	N\A
3	NM_003748	■	ALDH4A1	See Table 7.4
4	NM_006681	■	NMU	See Table 7.4
5	NM_000507	□	FBP1	Fructose-1,6-bisphosphatase 1 [Homo sapiens]. The protein encoded by this gene is a gluconeogenesis regulatory enzyme, catalyzes the hydrolysis of fructose 1,6-bisphosphate to fructose 6-phosphate and inorganic phosphate. Fructose-1,6-diphosphatase deficiency is associated with hypoglycemia and metabolic acidosis.
6	AF055033	■	IGFBP5	Insulin-like growth factor binding protein 5 [Homo sapiens]
7	NM_000286	□	PEX12	See Table 7.4
8	AL080059	■	TSPYL5	See Table 7.4

9	Contig33814_RC	□	N\A	N\A
10	NM_012429	□	SEC14L2	See Table 7.6
11	NM_000599	■	IGFBP5	Insulin-like growth factor binding protein 5 [Homo sapiens]
12	NM_003862	■	FGF18	Fibroblast growth factor 18 [Homo sapiens] . The protein encoded by this gene is a member of the fibroblast growth factor (FGF) family. FGF family members possess broad mitogenic and cell survival activities, and are involved in a variety of biological processes, including embryonic development, cell growth, morphogenesis, tissue repair, tumor growth, and invasion. It has been shown in vitro that this protein is able to induce neurite outgrowth in PC12 cells. Studies of the similar proteins in mouse and chick suggested that this protein is a pleiotropic growth factor that stimulates proliferation in a number of tissues, most notably the liver and small intestine. Knockout studies of the similar gene in mice implied the role of this protein in regulating proliferation and differentiation of midline cerebellar structures.
13	Contig63649_RC	■	N\A	N\A
14	NM_004994	■	MMP9	See Table 7.4
15	Contig11065_RC	■	N\A	N\A
16	Contig32185_RC	■	N\A	N\A
17	NM_016359	■	NUSAP1	Nucleolar and spindle associated protein 1 is a nucleolar-spindle-associated protein that plays a role in spindle microtubule organization (Raemaekers et al., 2003).
18	Contig15954_RC	□	N\A	N\A
19	NM_005635	□	SSX1	Synovial sarcoma, X breakpoint 1. The product of this gene belongs to the family of highly homologous synovial sarcoma X (SSX) breakpoint proteins. These proteins may function as transcriptional repressors. They are also capable of eliciting spontaneously humoral and cellular immune responses in cancer patients, and are potentially useful targets in cancer vaccine-based immunotherapy. SSX1, SSX2 and SSX4 genes have been involved in the t(X;18) translocation characteristically found in all synovial sarcomas. This translocation results in the fusion of the synovial sarcoma translocation gene on chromosome 18 to one of the SSX genes on chromosome X. The encoded hybrid proteins are probably responsible for transforming activity.
20	Contig49388_RC	■	N\A	N\A
21	Contig52554_RC	□	N\A	N\A
22	NM_020156	□	C1GALT1	Core 1 synthase, glycoprotein-N-acetylgalactosamine 3-beta-galactosyltransferase, 1 [Homo sapiens]. The protein encoded by this gene generates the common core 1 O-glycan structure, Gal-beta-1-3GalNAc-R, by the transfer of Gal from UDP-Gal to GalNAc-alpha-1-R. Core 1 is a precursor for many extended mucin-type O-glycans on cell surface and secreted glycoproteins. Studies in mice suggest that this gene plays a key role in thrombopoiesis and kidney homeostasis.

23	NM_006763	□	BTG2	BTG family, member 2 [Homo sapiens]. The protein encoded by this gene is a member of the BTG/Tob family. This family has structurally related proteins that appear to have antiproliferative properties. This encoded protein is involved in the regulation of the G1/S transition of the cell cycle

■: Listed in 70-gene signature, □: Not listed in 70-gene signature, N/A: No Available

7.2 Systemic responsiveness prediction to neoadjuvant treatment in breast cancer patients

Accurate prediction of treatment response in breast cancer can decrease significantly the number of patients receiving unnecessary systematic treatment and reduce its expensive medical costs. Currently, the selection of patient eligible for a treatment is generally based on classical factors such as tumor grade, age, lymph nodes status. However, the high heterogeneity of breast cancer highlights the need to design treatment regimens tailored specifically for each sub-molecular type cancer. HER2 overexpressed breast cancer for instance has an aggressive biological behavior and poor prognosis, requiring the design of specific treatment regimens. Although trastuzumab (Herceptin) has been shown to be a valuable remarkable therapeutic in certain HER2 overexpressing breast cancer patients, its overall response rate is still limited and its function mechanism is not yet very well understood. Less than 35% of patients with HER2 overexpressing metastatic breast cancer indeed respond to trastuzumab as a single therapy, whereas ~5% of patients suffer from severe side effects (e.g. cardiac dysfunction) and 40% of patients experience other adverse effects (Fujita et al. 2006). Therefore, the identification of new trastuzumab's predictive markers is urgently required to reduce the number of patients undergoing the side effects and unnecessary cost. The present study aims to identify new predictors of therapeutic responsiveness, among both available proteomic and clinical marker information, in HER2-overexpressing invasive breast cancer receiving neoadjuvant treatment. We used our proposed approach for feature selection that performs mixed-type data to derive a set of predictive factors.

a- Material and methods

Fifty-three patients with HER2-overexpressing invasive breast carcinoma received trastuzumba based neoadjuvant treatment from the cancer institute of Toulouse (ICR). The pathological response was evaluated on surgical specimens and categorized as complete response (pCR) (no residual or minimal invasive carcinoma) or incomplete response (pIR) (residual invasive carcinoma) according to sataloff criteria (Zindy *et al.*, 2011). In the present

study, among the 53 HER2-positive invasive breast cancer patients that received specific neoadjuvant treatment, 20 (37,73%) had achieved a pCR and 33 (62,26%) did not. Each patient is characterized by 14 features (proteomics and clinicopathological factors, see Table 7.9 for more details) and its outcome (pCR) listed below:

1. ER
2. PgR
3. Involved lymph Nodes
4. HER3
5. PTEN
6. CMYC
7. 4-EBP1
8. eI4E
9. HER2
10. HER4
11. PAX2
12. EGFR
13. Age (qualitative: <40, 40-50, >50)
14. Grade (interval: [3-5];[6-7],[8-9])
15. pCR : outcome (positive *vs* negative)

b- Results and discussion

In order to select important predictive factors, we applied Membas to this dataset and we obtained the weights shown in Figure 7.18 for respectively each factor. The marker's ranking is reported in Table 7.10 with brief description of its biological role in breast cancer evolution. It can be observed that, among the 14 markers, only 6 have a relatively significant weights and remaining factors seem to be weakly relevant to prediction task.

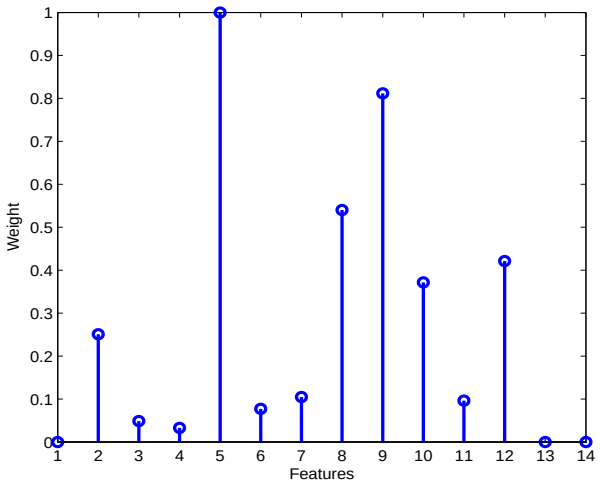


Fig. 7.18. Marker weights obtained by Membas

Tab. 7.9. List of Ranked predictive factors obtained by Membas.

Rank	Marker	Notation	Function
1	Phosphatase and tensin hom-ologue deleted on chromoso-me 10.	PTEN	A protein that helps control many cell functions, including cell division and cell death. Mutations (changes) in the gene that makes PTEN are found in many types of cancer and other diseases. It is a type of tumor suppressor protein. Also called PTEN tyrosine phosphatase.
2	Human Epidermal gro-wth factor receptor-2 status	HER2	See Glossary
3	Eukaryotic transaltion initia-tion factor	eI4E	Eukaryotic initiation factor
4	Epidermal Growth Factor Receptor	EGFR/HER1	The protein found on the surface of some cells and to which epidermal growth factor binds, causing the cells to divide. It is found at abnormally high levels on the surface of many types of cancer cells, so these cells may divide excessively in the presence of epidermal growth factor. Also called epidermal growth factor receptor, ErbB1, and HER1.
5	Human Epidermal gro-wth factor receptor-4 status	HER4	Tumor gene repressor
6	Progesterone receptor	PgR	See Glossary
7	4-EBP1	4-EBP1	Translation repressor
8	Paired box gene 2	Pax2	PAX2 encodes paired box gene 2, one of many human homologues of the Drosophila melanogaster gene prd. The central feature of this transcription factor gene family is the conserved DNA-binding paired box domain. PAX2 is believed to be a target of transcriptional suppression by the tumor suppressor gene WT1.
9	C-MYC	CMYC	Protein codes for a transcription factor that is located on chromosome 8 in humans and is believed to regulate expression of 15% of all genes through binding on Enhancer Box sequences (E-boxes) and recruiting histone acetyltransferases (HATs). This means that in addition to its role as a classical transcription factor, Myc also functions to regulate global chromatin structure by regulating histone acetylation both in gene-rich regions and at sites far from any known gene
10	Involved lymph Nodes	Invaded lymph Nodes	See Glossary
11	Human Epidermal gro-wth factor receptor-3 status	HER3	Tumor gene repressor
12	Estrogen receptor	ER	See Glossary
13	Age	/	See Glossary
14	Grade	/	See Glossary

LAMDA classifier has been used then to assess the importance of selected factors by retaining only a set of markers optimizing its classification performance. In this order, it has been found that the four top ranked markers by MEMBAS (PTEN, HER2, eI4E, EGFR) provide the optimal classification performance. Interestingly, the relation of both PTEN and HER2 with the response to trastuzumab is well established in cancer research literature (Fujita *et al.*, 2006; Vogel *et al.*, 2002) and are recognized as powerful predictive factors. In the very

close past, patients with metastatic are selected for trastuzumab-therapy if the primary tumor overexpresses the HER2 protein or HER2 gene amplification. However, in spite the importance of HER2 marker, less than 30% of patients respond to trastuzumab. This highlighted the fact that HER2 gene amplification is a necessary biomarker but not sufficient to predict the efficacy of trastuzumab (Fujita *et al.*, 2006). Recently, PTEN has been found to be one of the most common targets of mutation in human cancer and that a decreased PTEN expression is associated with invasive breast cancer and poor prognosis (Fujita *et al.*, 2006). It has been reported therein also that PTEN is a powerful predictive marker for the efficacy of trastuzumab in drug-resistant and parental HER2 over-expressing breast cancer cells. Eukaryotic transaltion initiation factor eIF4E is one of the most prominent downstream effector of mTOR (mammalian Target Of Rapamycin) signaling. It has been reported that a high level of eIF4E is often associated with poor prognosis (Byrnes *et al.*, 2006; Zhou *et al.*, 2004). In a recent study using the same group of patients, it has been found out that etopic expression of eIF4E in breast cancer tumors led to a loss in the trastuzumab-dependent decrease in both eIF4F formation and cell proliferation (Zindy *et al.*, 2011). This highlights the possible association between the expression of eIF4E and the pathological response to trastuzumab. A validation of such finding is underway on an independent multicenter cohort of patients.

The epidermal growth factor receptor (EGFR) is observed in 19-67% of malignant breast tumors and also appears to correlate with an adverse prognosis (Hudelist *et al.*, 2005). Both receptors EGFR and HER2 from the EGF family are linked to each other in an interdependent signaling network of considerable complexity (Hudelist *et al.*, 2005). It has been reported that EGFR Kinase activity largely depends upon the integrity of the HER2 kinase domain. Likewise, it has been found that the inhibition of EGFR kinase activity may be attenuated by HER2 overexpression. Conversely, HER2 activation is also strongly influenced by the presence and activation of EGFR (Hudelist *et al.*, 2005). It is therefore not surprising to consider EGFR marker in predicting the course of disease in patients receiving trastuzumab-based therapy.

Classification performance using those four markers is reported in Table 7.11. To show the effectiveness of the four markers combination, we compared this result with the performance with two different predictors: 1) when only two classical markers (PTEN and HER2) are used; and 2) when all available data (proteomics and clinical) are considered. It can be observed that the 4-markers combination outperforms both the 2-markers approach

(HER2+PTEN) and all the data. Particularly, the 4-markers combination (HER2+PTEN+eI4E+EGFR) improves significantly the specificity (more than 80% of positive responders are detected) compared to 2-marker combination. Figure 7.19 shows the obtained class profile for each marker.

Tab. 7.10. Comparatives results between 4-markers, 2-markers and all data approaches.

Method	TP	FP	FN	TN	Sensitivity (%)	Specificity (%)	Accuracy
4-markers*	15/20	6/33	5/23	27/33	**75**	**81.82**	**79.25**
HER2+ PTEN	15/20	11/33	5/20	22/33	75	66.67	69.81
All data	9/20	7/33	11/20	26/33	45	46.48	66.04

(*): PTEN+HER2+eI4E+EGFR

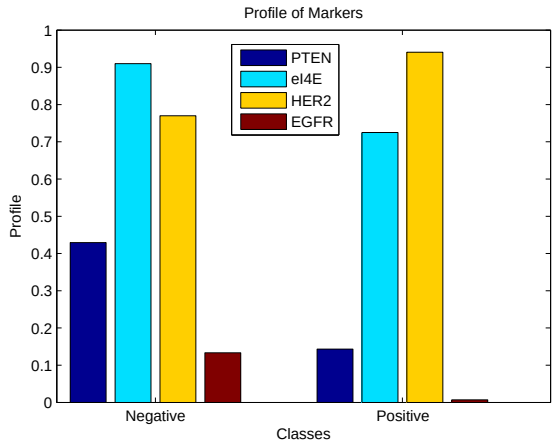


Fig. 7.19 Profile of negative and positive classes

To further demonstrate the predictive value of the 4-markers combination, we plotted in Figure 7.20 the ROC curve for the three predictors. It can be observed that the 4-markers combination outperforms significantly other approaches

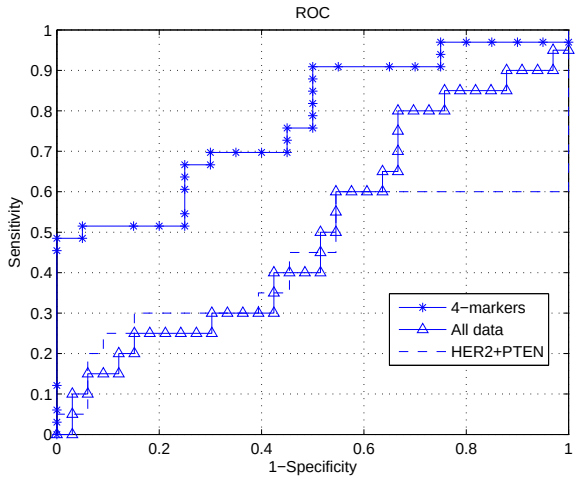


Fig. 7.20 ROC curve of three approaches.

7.3 Conclusion

In this chapter we presented some applications of our proposed approaches in breast cancer. We focused throughout this chapter on two main breast cancer management tasks: Prognosis and treatment responsiveness prediction.

First an application of cancer prognosis based only on heterogeneous clinical data was shown. Through this application we have shown that the feature weighting approach selects meaningful clinical factors. Two other feature selection approaches were tested on the same problem in order to compare the performance of our proposal.

In the second application cancer prognosis is based only on microarray data by deriving a prognostic signature. Obtained results using several criterions have shown that the predictive value 20-gene prognostic signature can be superior to other existing prognostic signatures and classical clinical guidelines. Particularly, the 20-gene signature improves significantly the specificity of one of the well known genetic approaches (70-gene signature).

The third application was devoted to investigate the integration of both clinical and microarray data. In such applications both problems of data heterogeneity and high dimensionality should be faced jointly. We have taken advantage of the interesting property of the proposed approach that enable to handle simultaneously both problems to derive a hybrid prognostic signature. We have shown then through some analysis that the integration of both approach may improve the breast cancer prognosis. The hybrid signature improves the sensitivity of 20-gene signature while maintaining comparable specificity.

To defy low signal-to-noise ratio in microarray data for cancer prognosis, a symbolic approach has been considered to derive a more robust prognostic signature, referred to as GenSym. We described first the microarray interval dataset generation by incorporating a white Gaussian noise with a specific Signal-to-Noise Ratio. We have shown through some experiments and analysis that the GenSym signature can outperform other existing approaches. Particularly, it enables to keep the good sensitivity raised with the hybrid signature while improving further the good specificity of the 20-gene signature. Moreover, the gene list of this signature holds meaningful genes related to invasion, cell cycle and proliferation. We believe that this first attempt in that direction open also the door to the machine learning community to investigate other approaches for addressing this problem.

The last application concerns the problem of responsiveness prediction to neoadjuvant treatment in HER2 over-expressed breast cancer patients. Using our proposed approach we derived a signature constituted of four markers (PTEN, HER2, eI4E, EGFR), that improves significantly the discriminative power among positive and negative responders compared to the usually used 2-marker approach (PTEN, HER2). Particularly, the 4-markers combination improves significantly the specificity of the 2-marker combination. This highlights the importance of two new predictive factors (eI4E, EGFR) to predict accurately the responsiveness of HER2 over-expressed breast cancer patients to neoadjuvant treatment.

Conclusion et perspectives- Résumé

Notre objectif dans ce travail était de développer de nouveaux outils pour une gestion plus précise du cancer de sein. Nous présentons ici une tentative pour proposer des approches adaptées dans le cadre de l'apprentissage automatique, permettant de surmonter les principaux défis récents rencontrés dans le domaine du cancer tels que la dimension élevée des informations à traiter, les bruits de mesure, les incertitudes sur l'appartenance du patient aux différents sous-types de cancer et l'hétérogénéité des données (quantitatif ou symbolique).

Dans un premier travail, une approche intégrée de sélection de variables basée sur l'apprentissage ℓ_1 capable de traiter des données de haute dimension a été proposée. En particulier, cette approche propose un nouvel algorithme pour résoudre le problème ℓ_1SVM dans le domaine primal. Cependant, avec la récente tendance vers une bioinformatique intégrative qui vise à intégrer différentes sources de données, l'occurrence conjointe de trois défis est possible dans certaines applications. Pour faire face simultanément à ces trois défis, une deuxième approche a été proposée. Tout d'abord, un principe unifié pour faire face au problème de l'hétérogénéité des données a été établi. Ensuite, une approche floue de pondération de variables supervisée a été proposée en se basant sur ce principe. Le processus de pondération est basé principalement sur l'optimisation d'une fonction objective intégrant la notion de marge d'appartenance. En se basant sur le même principe, la méthode de pondération a été ensuite étendue au cas non supervisé afin de développer un nouvel algorithme de pondération à base de règles floues pour effectuer la tâche de regroupement. L'efficacité de toutes ces approches a été validée dans une étude expérimentale extensive et comparées avec celles de méthodes bien connues dans la littérature. Enfin, certaines applications dans le domaine du cancer du sein ont été effectuées en utilisant les approches proposées. Ces applications ont concerné essentiellement le développement de modèles pronostiques et prédictifs à partir de l'analyse de données de puces à ADN et/ou de données cliniques. Nous avons montré à travers une étude comparative l'efficacité de ces modèles en termes de précision et de détermination de la survie. Nous avons examiné aussi l'interprétation de ces signatures d'un point de vue biologique. Enfin des perspectives de ce travail ont été présentées que ce soit de nature méthodologique ou applicative.

Conclusion and future works

Our aim in this work was to develop new tools for breast cancer management to help the physicians in their decision-making practices. In this order an attempt to propose suitable approaches has been performed within machine learning framework, to enable handling the main recent challenges encountered in breast cancer management field. Some challenges are due to the intrinsic complexity of data issued from high throughput technologies introduced recently in cancer management such as microarrays. The gene expression profiling, through microarray technology, has indeed brought the hope to gain new insights into cancer biology but requires meanwhile smart approaches capable to fit with high dimensional data and uncertainties. Uncertainties can be in the form of either measurement noise or membership uncertainty of a patient to different cancer subtype groups. Another challenge is related to the use of traditional clinical factors characterized by its heterogeneity; the data can be of quantitative or symbolic type.

In a first work an embedded feature selection approach based on ℓ_1 learning able to deal with high dimensional data has been proposed. This approach proposes a new algorithm to solve the ℓ_1SVM problem in the primal domain. The basic idea is the transformation of the initial convex optimization problem into unconstrained non-convex one, upon which, via gradient descent method, reaching a globally optimum solution is guaranteed. The non differentiable property of the hinge loss function has been overcome by using its approximated Huber loss function. It has been shown through large-scale numerical experiments that the proposed approach is computationally more efficient than the few existing methods solving the same problem.

However, with the recent trends towards an integrative bioinformatics that aims to integrate different data sources, the occurrence of three challenges simultaneously is possible in some cancer applications. To deal simultaneously with these three challenges; data dimensionality, heterogeneity and uncertainties, a second approach has been proposed. First of all, a unified principle to deal with data heterogeneity problem has been established. To take into account membership uncertainty and increase model interpretability, this principle has been proposed within a fuzzy logic framework. Besides, in order to alleviate the problem of high level noise, a symbolic approach has been developed suggesting the use of interval representation to model the noisy measurements. This principle is based on the mapping of different type of

data from initially heterogeneous spaces into a common space through an adequacy measure. This allows then to reason in unified way about the data in the new space whatever its initial type for different data analysis purposes.

In particular, a supervised fuzzy feature weighting approach has been proposed based on this principle. This approach has been integrated based on a fuzzy weighted rule concept into a fuzzy rule-based classifier in the aim to improve its performance. In addition to its ability to handle the problems of data heterogeneity and uncertainties, the proposed approach is capable to fit with high data dimensionality. The weighting process is mainly based on the definition of a membership margin for each sample. It optimizes then a membership-margin objective function using classical optimization approach to avoid combinatorial search. The effectiveness of this approach has been assessed through an extensive experimental study and compared with well-know feature selection methods. Based on the same principle, the weighting approach has been then extended to the unsupervised case in order to develop a new weighted fuzzy rule-based clustering algorithm. An extensive study has been also performed to compare this algorithm with one of the state-of-the-art clustering algorithm.

Finally some breast cancer applications have been presented. These applications have concerned mainly cancer prognosis and prediction of treatment benefit. Predictive and prognostic models were derived based on microarray and/or clinical data. We have shown through a comparison the effectiveness of these models in term of accuracy and survivability. Since the aim of developing new predictive tools for breast cancer prognostication from microarray data is twofold, i.e. to yield good prediction performance and to gain new biological insights into cancer biology, we have shown also that the derived models were interpretable from a biological point of view. In particular, the applications have concerned

1) Cancer prognosis based only on clinical data
2) Derivation of 20 genes signature for cancer prognosis based on microarray data
3) Derivation of hybrid signature for cancer prognosis based on the integration of clinical and microarray data
4) Derivation of more robust prognostic signature (referred to as GenSym) based on a symbolic approach by modeling the noisy microarray measurments as symbolic intervals
5) Derivation of 4-markers signature for the prediction of neoadjuvant treatment benefit in HER2 over-expressed breast cancer patients.

Different future works are twofold in the framework of the current work such as:

a) Methodological

- The implementation of the proposed approach in one package with suitable interface destined for medical applications is underway to facilitate its use by clinicians. This package should enable also the comparison with existing predictive models.
- In the present work we consider only three types of data (quantitative, qualitative and interval data). Although these are the most used type of data, there are other types of data which could be faced in many real-world applications (such as histograms, fuzzy numbers, …). So it will be interesting to investigate the extension of this approach to further type of data.
- Another interesting direction would be to propose other shapes of membership function for different types of data and investigate their use.

b) Cancer applications

- With the recent trends towards an integrative bioinformatics, it will be interesting to use the proposed approach to integrate other type sources of data instead of only microarray and clinical data. High dimensional data will be generated by new high throughput technologies, e.g. single nucleotide polymorphism (SNP) or comparative genomic hybridization (CGH), at a continuously growing rate. Therefore, with this huge quantity of data, an increase need of more effective tools by physicians is expected, enabling to extract useful biological knowledge and gain insights into cancer biology.
- We considered here only the breast cancer. However, this approach can be applied for the derivation of molecular signatures for other type of cancers.
- It has been reported recently that breast cancer is very heterogeneous disease and can be divided to several molecular subtypes. It is possible to investigate the application of this approach to derive molecular signature for each subtype.
- The factor of time is still to date neglected in the design of predictive and prognostic tools. Cancer progressiveness is strongly related to time and we believe that taking into account this factor, jointly with molecular cancer subtyping, can play a central role in improving cancer management tools. This direction can also be investigated using the proposed approach.

Moreover, it must be noted here also that the proposed approaches have been applied with success on other fields related to dynamical system diagnosis such as the diagnosis of chemical reactors (Hedjazi *et al.*, 2010c; Hedjazi *et al.*, 2011e, Hedjazi *et al.*, 2011f).

Finally, it is worthwhile to note that the results presented in this work are now subject of use in a recent ANR project (INNODIAG: Innovation in molecular diagnostic in health using the latest development in Nanotechnology: Application to breast cancer prognosis). This project is concerned first of all by the selection of a set of genetic and clinical markers for breast cancer treatment derived from the obtained signatures during the present work. A big number of public datasets issued from different medical centers and using different technologies is being used for the signature extraction and validation. This set of biomarkers will be tested then and compared to other signatures on a pool of patients issued from the Institut Claudius Regaud. In parallel a new bioship generation relied on soft lithography and optical detection will be developed. A first prototype developed with the optimal signature derived in the first part will be then designed. This new type of bioship will enable to direct toward a personalized medicine and help clinicians and oncologists to select the optimal cancer treatment.

Glossary of Cancer Terms

Estrogen receptor A protein found inside the cells of the female reproductive tissue, some other types of tissue, and some cancer cells. The hormone estrogen will bind to the receptors inside the cells and may cause the cells to grow. Also called ER.

Estrogen receptor negative Describes cells that do not have a protein to which the hormone estrogen will bind. Cancer cells that are estrogen receptor negative do not need estrogen to grow, and usually do not stop growing when treated with hormones that block estrogen from binding. Also called ER-.

Estrogen receptor positive Describes cells that have a receptor protein that binds the hormone estrogen. Cancer cells that are estrogen receptor positive may need estrogen to grow, and may stop growing or die when treated with substances that block the binding and actions of estrogen. Also called ER+.

Estrogen receptor test A lab test to find out if cancer cells have estrogen receptors (proteins to which estrogen will bind). If the cells have estrogen receptors, they may need estrogen to grow, and this may affect how the cancer is treated.

Progesterone receptor A protein found inside the cells of the female reproductive tissue, some other types of tissue, and some cancer cells. The hormone progesterone will bind to the receptors inside the cells and may cause the cells to grow. Also called PR or PgR.

Progesterone receptor negative Describes cells that do not have a protein to which the hormone progesterone will bind. Cancer cells that are progesterone receptor negative do not need progesterone to grow, and usually do not stop growing when treated with hormones that block progesterone from binding. Also called PR-.

Progesterone receptor positive Describes cells that have a protein to which the hormone progesterone will bind. Cancer cells that are progesterone receptor positive need progesterone to grow and will usually stop growing when treated with hormones that block progesterone from binding. Also called PR+.

Progesterone receptor test A lab test to find out if cancer cells have progesterone receptors (proteins to which the hormone progesterone will bind). If the cells have progesterone receptors, they may need progesterone to grow, and this can affect how the cancer is treated.

HER2/neu A protein involved in normal cell growth. It is found on some types of cancer cells, including breast , ovarian and other cancer type. Cancer cells removed from the body may be tested for the presence of HER2/neu to help decide the best type of treatment. HER2/neu is a type of receptor tyrosine kinase. Also called c-erbB-2, human EGF receptor 2, and human epidermal growth factor receptor 2.

uPA An enzyme that is made in the kidney and found in the urine. A form of this enzyme is made in the laboratory and used to dissolve blood clots or to prevent them from forming. Also called u-plasminogen activator, urokinase, and urokinase-plasminogen activator.

Biopsy The removal of cells or tissues for examination by a pathologist. The pathologist may study the tissue under a microscope or perform other tests on the cells or tissue. There are many different types of biopsy procedures. The most common types include: (1) incisional biopsy, in which only a sample of tissue is removed; (2) excisional biopsy, in which an entire lump or suspicious area is removed; and (3) needle biopsy, in which a sample of tissue or fluid is removed with a needle. When a wide needle is used, the procedure is called a core biopsy. When a thin needle is used, the procedure is called a fine-needle aspiration biopsy.

Adjuvant therapy Additional cancer treatment given after the primary treatment to lower the risk that the cancer will come back. Adjuvant therapy may include chemotherapy, radiation therapy, hormone therapy, targeted therapy.

Neoadjuvant therapy Treatment given as a first step to shrink a tumor before the main treatment, which is usually surgery, is given. Examples of neoadjuvant therapy include chemotherapy, radiation therapy, and hormone therapy. It is a type of induction therapy.

Mastectomy Surgery to remove the breast (or as much of the breast tissue as possible)

Docetaxel A drug used together with other drugs to treat certain types of breast cancer, stomach cancer, prostate cancer, and certain types of head and neck cancer. It is also being studied in the treatment of other types of cancer. Docetaxel is a type of mitotic inhibitor. Also called Taxotere.

Tamoxifen A drug used to treat certain types of breast cancer in women and men. It is also used to prevent breast cancer in women who have had ductal carcinoma in situ (abnormal cells in the ducts of the breast) and in women who are at a high risk of developing breast cancer (in US). Tamoxifen is also being studied in the treatment of other types of cancer. It

blocks the effects of the hormone estrogen in the breast. Tamoxifen is a type of antiestrogen. Also called tamoxifen citrate.

Immunohistochemistry A technique used to identify specific molecules in different kinds of tissue. The tissue is treated with antibodies that bind the specific molecule. These are made visible under a microscope by using a color reaction, a radioisotope, colloidal gold, or a fluorescent dye. Immunohistochemistry is used to help diagnose diseases, such as cancer, and to detect the presence of microorganisms. It is also used in basic research to understand how cells grow and differentiate (become more specialized).

Chemotherapy Treatment with drugs that kill cancer cells. It is usually followed by docetaxel and anthracyclin.

Adjuvant! Online It is a tool that helps health professionals make estimates of the risk of poor outcome (cancer related mortality or relapse) without systemic adjuvant therapy, estimates of the reduction of these risks afforded by therapy, and risks of side effects of the therapy. These estimates are based on information entered about individual patients and their tumors (e.g. patient age, tumor size, nodal involvement,or histological grade). These estimates are then provided on printed sheets in simple graphical and text formats to be used in consultations.

Radiation therapy The use of high-energy radiation from x-rays, gamma rays, neutrons, protons, and other sources to kill cancer cells and shrink tumors. Radiation may come from a machine outside the body (external-beam radiation therapy), or it may come from radioactive material placed in the body near cancer cells (internal radiation therapy). Systemic radiation therapy uses a radioactive substance, such as a radiolabeled monoclonal antibody, that travels in the blood to tissues throughout the body. Also called irradiation and radiotherapy.

Aromatase inhibitor A drug that prevents the formation of estradiol, a female hormone, by interfering with an aromatase enzyme. Aromatase inhibitors are used as a type of hormone therapy for postmenopausal women who have hormone-dependent breast cancer.

Hormone therapy Treatment that adds, blocks, or removes hormones. For certain conditions (such as diabetes or menopause), hormones are given to adjust low hormone levels. To slow or stop the growth of certain cancers (such as prostate and breast cancer), synthetic hormones or other drugs may be given to block the body’s natural hormones. Sometimes surgery is

needed to remove the gland that makes a certain hormone. Also called endocrine therapy, hormonal therapy, and hormone treatment. It is usually followed by Tamoxifen and anti-aromatase or aromatase inhibitors.

Axillary lymph node A lymph node in the armpit region that drains lymph from the breast and nearby areas.

Overall survival rate The percentage of people in a study or treatment group who are alive for a certain period of time after they were diagnosed with or treated for a disease, such as cancer. The overall survival rate is often stated as a five-year survival rate, which is the percentage of people in a study or treatment group who are alive five years after diagnosis or treatment. Also called survival rate.

Disease-free survival The length of time after treatment for a specific disease during which a patient survives with no sign of the disease. Disease-free survival may be used in a clinical study or trial to help measure how well a new treatment works.

Fine-needle aspiration biopsy The removal of tissue or fluid with a thin needle for examination under a microscope. Also called FNA biopsy.

TNM staging system A system developed by the American Joint Committee on Cancer (AJCC) that uses TNM to describe the extent of cancer in a patient's body. T describes the size of the tumor and whether it has invaded nearby tissue. N describes whether cancer has spread to nearby lymph nodes, and M describes whether cancer has metastasized (spread to distant parts of the body). The TNM staging system is used to describe most types of cancer. Also called AJCC staging system.

Level Of Evidence (LOE) Three levels of evidence can be distinguished:

- LOE III: Low level of evidence
- LOE II: Intermediate level of evidence
- LOE I: High level of evidence

Appendixes

Appendix 1

A.1 Prognostic and predictive factors in breast cancer

It is crucial to have a clear understanding of the definitions of prognostic factors and predictive factors and their roles in guiding patient care before embarking on a discussion of their utility in breast cancer.

A.1.1 Prognostic factors

Prognostic factors determine the natural history of disease progression, in the absence of systemic therapies. Such factors often reflect the intrinsic biologic characteristics of tumors, such as their ability to proliferate and metastasize. Putative tumor markers or factors are ideally evaluated for their prognostic ability prospectively in the systemically untreated patient in order to eliminate the confounding effects of treatment. Unfortunately, much of the data about prognostic factors is obtained from retrospective analysis of banked tumor samples. As a result, published studies often include only small sample sizes, have different lengths of follow-up, lack complete data on conventional prognostic factors, do not control for confounding variables, and report a variety of different endpoints, including overall survival (OS) and disease-free survival (DFS). All of the stated factors render it difficult to compare results from different studies, and diminish the strength of the evidence obtained.

A.1.2 Predictive factors

Predictive factors are cues that a particular tumor might respond (or not) to a specific therapy. A purely predictive factor separates treated patients into good and poor outcome groups, but does not predict outcome in untreated patients. Usually, factors are often both prognostic and predictive, rather than purely prognostic or purely predictive. A classic example is estrogen-receptor (ER) status. Not only does ER-negativity give a less favorable prognosis, but more significantly, it predicts the category of patients who do not derive benefit from anti-estrogen therapy.

We review by next the important prognostic and predictive factors and their role in breast cancer care:

a- Classical prognostic factors

- **Axillary lymph node status:** Characterized by N-positive or N-negative according whether the patient present invaded nodes or not. It has been reported that 20 to 30%

of N-negative patients likely present a recurrence within 10 years compared to 70% for N-positive patients. The number of invaded nodes is also an important prognostic factor, patients with 4 invaded lymph nodes or more will likely have a poor prognosis than patients with less than 4 invaded lymph nodes (Carter *et al.*, 1989). To date, the single most powerful prognostic factor in primary breast cancer remains the status of the auxiliary lymph nodes.

- **Tumor size:** It has been reported that patients with tumor less than 1 cm had a 5-year relative OS of close to 99%, compared to 89% for those with tumors 1 to 3 cm, and 86% for those with tumors 3 to 5 cm (Carter *et al.*, 1989).
- **Histologic subtype:** The most common histological subtypes of breast cancer are infiltrating ductal and lobular carcinomas. Infrequent histologies such as pure tabular, mucinos, or modullary subtypes are associated with a particularly favorable prognosis with long-term recurrence rates of less than 10% (Diab *et al.*, 1999). Wong *et al.* (2002) examined the rate of axillary lymph node involvement in more than 3300 women with breast cancer. axillary lymph node were identified in 35% of women with infiltrating ductal carcinoma, but in only 11% of those with favorable subtypes. In addition, women with inflammatory breast cancer have an extremely poor prognosis.
- **Hormone receptor status:** It concerns Estrogen Receptor status (ER) and Progestrone Receptor status (PgR). While ER status has been reported a relatively weak prognostic factor, it strongly predicts for response to adjuvant hormonal therapy (Smith *et al.*, 2003). More specifically, ER-negative status appears to predict lack of responsiveness to hormonal therapy. Thus, ER status should be used primarily in making recommendations regarding the use of hormonal therapy in the adjuvant setting. Whereas the impact of progesterone receptor (PR) status as a prognostic and predictive marker was recently analyzed in a retrospective study of a large dataset of early-stage breast cancer patients who were randomized to either no adjuvant systemic therapy or adjuvant tamoxifen alone. Progesterone receptor status was found to add little further prognostic information over and above ER status. However, it appeared to further predict responsiveness to tamoxifen. Patients with ER+/PR+ tumors treated with tamoxifen had a 53% reduction in their risk of recurrence compared to a 25% reduction in risk noted in those with ER+/PR- tumors, relative to the risk of recurrence in ER-/PR- tumors (Bardou *et al.*, 2003).
- **Tumor grade:** The most widely accepted grading system is the semiquantitative Elston and Ellis modification of the Scarff-Bloom-Richardson (SBR) classification

(Bloom *et al.*, 1957). Investigators using the SBR classification observed a statistically significant correlation between histological grade and 5-year DFS for both node-negative and node-positive patients. Women with tumors with an SBR score of grade 3 had a relative risk of recurrence of 4.4 when compared with those with an SBR of grade 1. However, tumor grade as a prognostic factor are limited by the high degree of inter-observer variability and the lack of consistent methodology of objective and quantitative grading. Comparisons between studies are difficult because of varying grading systems. Moreover, studies examining the prognostic significance of tumor grade are inconsistent in the groupings of tumor grades. Whereas studies typically compare grade 1 versus grade 3, the position of grade 2 is variable. In some studies, grade 2 is clustered with grade 1, and in others, with grade 3. As a result of the above inconsistencies, the most recent revision of the American Joint Committee on Cancer Staging (AJCC) chose not to include histological grade in the TNM-staging criteria for breast cancer (Singletary *et al.*, 2002).

- **Human epidermal growth factor receptor-2 status (HER-2)**: HER-2 amplification (and the overexpression of receptor by the tumors) is associated to a poor prognosis and maybe predictive to certain treatments response. Studies suggest that tumors with HER-2 overexpression or amplification may have differential sensitivities to chemotherapeutic agents and to hormonal agents. The knowledge of HER2 status is required in all clinical situations. Human epidermal growth factor receptor-2 levels can be measured in several ways, including IHC utilizing a variety of antibodies to determine protein expression, and fluorescence in situ hybridization (FISH) or chromogenic in situ hybridization (CISH) to determine gene amplification.

b- Newer prognostic factors

- **Urokinase Plasminogen Activator system:** Research in the past decade has provided increasingly compelling evidence to suggest that the urokinase plasminogen activator (uPA) system plays a critical role in cancer invasion and metastasis. Urokinase plasminogen activator proteolytically converts plasminogen to plasmin. Plasmin activates matrix metalloproteases that degrade the extracellular matrix and modulate cellular adhesion, proliferation and migration. Both uPA and its physiologic inhibitor, plasminogen activator inhibitor-1 (PAI-1) have been shown to be upregulated in multiple cancer types, especially breast cancer (Duffy *et al.*, 2004). Based on large, well-controlled, retrospective studies and data from a prospective randomized trial, high levels of uPA/PAI-1 have been demonstrated to provide independent prognostic

value. In addition, data are mounting to suggest that these factors may also predict for tumor response to chemotherapy. The determination of uPA/PAI-1 levels must be performed by enzyme-linked immunosorbent assay (ELISA), which requires fresh frozen tissue. This issue limits its routine integration in clinical practice. Current studies are underway to develop reproducible assays from smaller amounts of tissue obtained from core needle biopsy material.

- **Markers of proliferation: S-phase fraction, thymidine-labeling index, Ki-6**

  The role of markers of proliferation as prognostic factors has been extensively investigated. Different methodologies exist to assess the rate of proliferation including thymidine labeling index (TLI), DNA-flow cytometry, S-phase fraction (SPF), mitotic index, bromodeoxyuridine (BrDu) incorporation, and IHC techniques with antibodies directed at antigens present during cell proliferation, such as Ki-67 (MIB-1) and PCNA. There is abundant literature on this topic, with over 200 publications examining the role of SPF as a prognostic marker alone. This literature is complex to interpret because of the variability of methodologies and assay systems and different cut-offs for high versus low rates of proliferation. Nonetheless, the majority of the studies that included large numbers of women with long follow-up, that controlled for the classical prognostic factors, suggest that proliferative rate is an independent predictor of patient outcome.

- **Gene expression profile by cDNA microarray:** Recently microarray technology has made it possible to measure simultaneously thousands of gene expressions. By analyzing the expression differentiation, genetic markers can be derived either for prognosis or prediction purposes that has been shown able to outperform classical factors in many prospective studies (Van't Veer *et al.*, 2002). However, this field is still presenting various challenges related to the incoherence between the obtained results (variability observed according to the used platform, data samples,…) and the lack of prospective studies to validate its use in the clinical routine. We review below the practical aspects related to the microarray technology and different existing platforms.

A.2 Microarray technology:

Microarray technology is based on the central dogma of molecular biology, namely the production of proteins from DNA as illustrated in Figure A.1. Briefly, this operation is based mainly on two steps (Figure A.1): Transcription and Translation which consists in the first

step of DNA (gene) translation into pre-mRNA and once this pre-mRNA is processed the resulting mRNA message is in the second step translated by ribosome in order to produce proteins (Translation). The detailed biological operation can be summarized as follow:

A.2.1 Transcription

Transcription is the process by which the information contained in a section of DNA is transferred to a newly assembled piece of messenger RNA (mRNA). It is facilitated by RNA polymerase and transcription factors. In eukaryote cells the primary transcript (pre-mRNA) is often processed further via alternative splicing. In this process, blocks of mRNA are cut out and rearranged, to produce different arrangements of the original sequence.

A.2.2 Translation

Eventually, this mature mRNA finds its way to a ribosome, where it is translated. In prokaryotic cells, which have no nuclear compartment, the process of transcription and translation may be linked together. In eukaryotic cells, the site of transcription (the cell nucleus) is usually separated from the site of translation (the cytoplasm), so the mRNA must be transported out of the nucleus into the cytoplasm, where it can be bound by ribosomes. The mRNA is read by the ribosome as triplet codons, usually beginning with an AUG, or initiator methionine codon downstream of the ribosome binding site. Complexes of initiation factors and elongation factors bring aminoacylated transfer RNAs (tRNAs) into the ribosome-mRNA complex, matching the codon in the mRNA to the anti-codon in the tRNA, thereby adding the correct amino acid in the sequence encoding the gene. As the amino acids are linked into the growing peptide chain, they begin folding into the correct conformation. Translation ends with a UAA, UGA, or UAG stop codon. The nascent polypeptide chain is then released from the ribosome as a mature protein. In some cases the new polypeptide chain requires additional processing to make a mature protein. The correct folding process is quite complex and may require other proteins, called chaperone proteins. Occasionally, proteins themselves can be further spliced; when this happens, the inside "discarded" section is known as an intein.

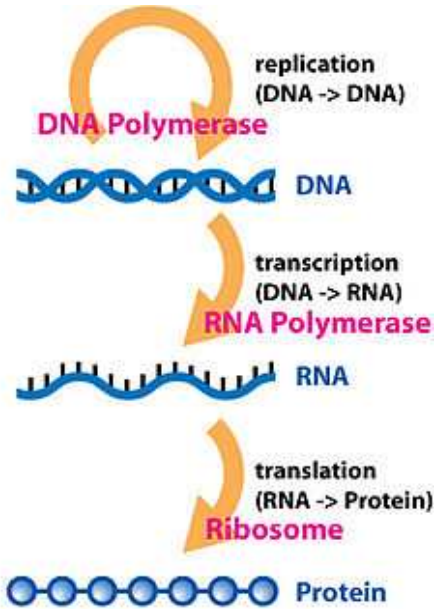


Fig. A.1 Biology dogma, from the DNA (gene) to the protein. Image from Wikipedia.

The concept behind DNA chip or microarray technology relies on the accurate binding, or hybridization, of strands of DNA with their precise complementary copies in experimental conditions where one sequence is also bound onto a solid-state substrate (glass). RNA is extracted from frozen breast tumour samples collected either at surgery or before treatment, labeled with a detectable marker (fluorescent dye), and hybridized to the array containing individual gene-specific probes. Gene-expression levels are estimated by measuring the fluorescent intensity for each gene probe (Figure A.2). A gene-expression vector is then collected by summarizing the expression levels of each gene in the sample. To facilitate the comparison between the different experiments and compensate for difference in labeling, hybridizations and detection methods, a normalization step is usually performed (Figure A.2). Gene-expression prognostic classifiers are usually built by correlating gene-expression patterns, generated from tumour surgical specimens, with clinical outcome (development of metastases during follow-up). Gene-expression predictive classifiers of response to treatment are generated by correlating gene-expression data, derived from biopsies taken before pre-operative systemic therapy, with clinical and/or pathological response to the given treatment.

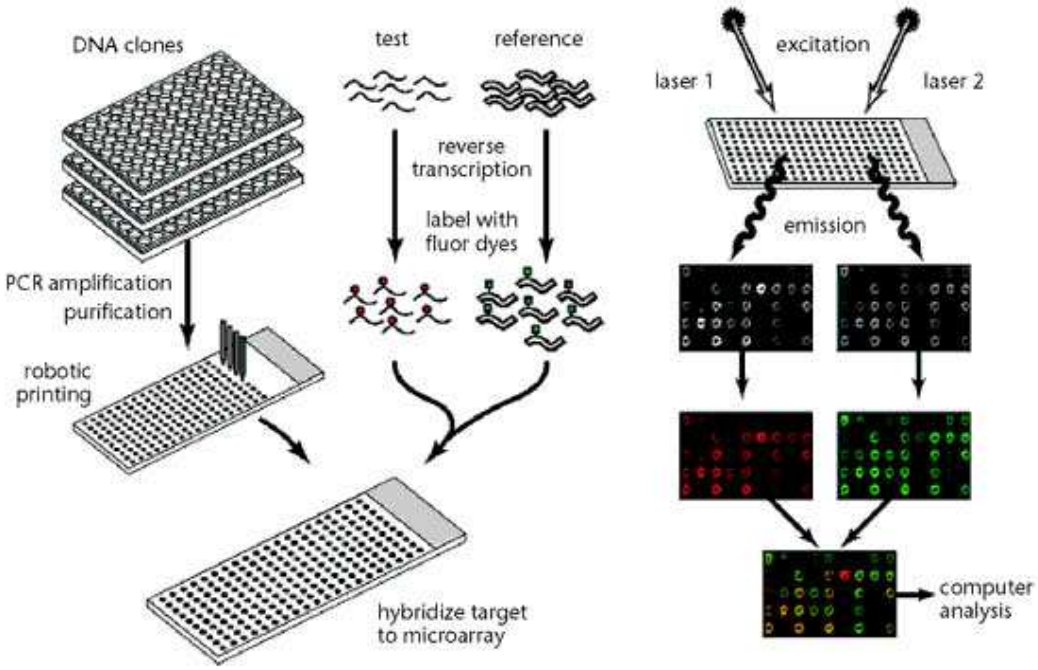


Fig. A.2 Microarray experiment schema. Image from (Duggan *et al.*, 1999).

Appendix 2

Proof Lemma 3.1:

For notational convenience lets denote *[N]* = [1, · · · ,*N*]. We first prove that the minimum of the objective functions of both problems are identical given the same data set. By the triangle inequality, $\left\|\overline{\mathrm{w}}^{*(1)}-\overline{\mathrm{w}}^{*(2)}\right\|_1 \le \left\|\overline{\mathrm{w}}^{*(1)}\right\|_1+\left\|\overline{\mathrm{w}}^{*(2)}\right\|_1=\left\|\overline{\mathrm{w}}^{*}\right\|_1$. Also by construction, $\forall n\in[N]$, $L\left(y_n,\mathrm{w}^{*T}\mathrm{x}^{(n)}+\mathrm{b}^{*}\right)=L\left(y_n,\overline{\mathrm{w}}^{*T}\overline{\mathrm{x}}^{(n)}+\mathrm{b}^{*}\right)$. It follows that if $(\overline{\mathrm{w}}^{*},\mathrm{b}^{*})$ is an optimal solution to (3.2), we have $\min f_1 \le f_1(\overline{\mathrm{w}}^{*(1)}-\overline{\mathrm{w}}^{*(2)},\mathrm{b}^{*}) \le f_2(\overline{\mathrm{w}}^{*},\mathrm{b}^{*})=\min f_2$. On the other hand, let $(\mathrm{w}^{*},\mathrm{b}^{*})$ be an optimal solution to (3.1). We construct two vectors $\overline{\mathrm{w}}^{o(1)}$ and $\overline{\mathrm{w}}^{o(2)}$:

$$\overline{w}_j^{o(1)}=\begin{cases} w_j^{*} & w_j^{*}\ge 0 \\ 0 & w_j^{*}<0, \end{cases} \text{ and } \overline{w}_j^{o(2)}=\begin{cases} 0 & w_j^{*}\ge 0 \\ w_j^{*} & w_j^{*}<0, \end{cases} \tag{A2.1}$$

By construction, we thus have $\min f_2 \le f_2([(\overline{\mathrm{w}}^{o(1)})^T,(\overline{\mathrm{w}}^{o(2)})^T]^T,\mathrm{b}^{*})=f_1(\mathrm{w}^{*},\mathrm{b}^{*})=\min f_1$. Hence, $\min f_1=\min f_2$, and $(\overline{\mathrm{w}}^{*(1)}-\overline{\mathrm{w}}^{*(2)},\mathrm{b}^{*})$ or $(\mathrm{w}^{*},\mathrm{b}^{*}))$ and $(\overline{\mathrm{w}}^{*},\mathrm{b}^{*})$ (or $(\overline{\mathrm{w}}^{o},\mathrm{b}^{*}))$ are the optimal solutions to (3.1) and (3.2), respectively.

Proof Lemma 3.2:

Let $\mathrm{w}^{*}=\overline{\mathrm{w}}^{*(1)}-\overline{\mathrm{w}}^{*(2)}$. By Lemma 3.1, $(\mathrm{w}^{*},\mathrm{b}^{*})$ is an optimal solution to (3.1). Using Eq. (A3.1), we construct a vector $\overline{\mathrm{w}}^{o}$. Suppose that there exists an element *j* so that $\overline{w}_j^{*(1)}\neq 0$ and $\overline{w}_j^{*(2)}\neq 0$. Then, by the triangle inequality, it follows that $\left\|\overline{\mathrm{w}}^{*}\right\|_1>\left\|\overline{\mathrm{w}}^{o}\right\|_1$. Hence, $f_2(\overline{\mathrm{w}}^{*},\mathrm{b}^{*})>f_2(\overline{\mathrm{w}}^{o},\mathrm{b})$, which contradicts the fact that $(\overline{\mathrm{w}}^{*},\mathrm{b}^{*})$ is an optimal solution. Therefore, $\forall j\in[\Im]$, either $\overline{w}_j^{*(1)}$ or $\overline{w}_j^{*(2)}$ or both equal to zero.

Proof Theorem 3.1:

For simplicity, we use $\partial G/\partial\overline{\mathrm{v}}^{*}$ to denote $\partial G/\partial\overline{\mathrm{v}}\big|_{\overline{\mathrm{v}}=\overline{\mathrm{v}}^{*}}$. Also, we use $\mathrm{A}>0$ and $\mathrm{A}\ge 0$ to denote that matrix A is positive definite or semi-definite, respectively.

We examine the properties of the Hessian matrix of $G(\overline{\mathrm{v}})$, denoted as H. Let $\overline{\mathrm{v}}^{+}$ be a stationary point of $G(\overline{\mathrm{v}})$ satisfying:

$$\frac{\partial G}{\partial\overline{\mathrm{v}}^{+}}=\left[\frac{\partial f}{\partial w_1^{+}}2v_1^{+},\ldots,\frac{\partial f}{\partial w_J^{+}}2v_J^{+},\frac{\partial f}{\partial \mathrm{b}^{+}}\right]^T=0 \tag{A2.2}$$

where $\mathrm{w}^{+}=[w_1^{+},\ldots,w_J^{+}]^T=[(v_1^{+})^2,\ldots,(v_J^{+})^2]^T$ and $\mathrm{b}^{+}=v_{J+1}^{+}$.

Note that some elements of $\bar{\mathrm{v}}^+$ may be equal to zero. For simplicity and without loss of generality, assume that the first *M* elements of $\bar{\mathrm{v}}^+$ belong to $S_0 = \{v_j^+ : v_j^+ = 0, 1 \le j \le J\}$, while the rest $J - M$ elements belong to $S_{\neq 0} = \{v_j^+ : v_j^+ \neq 0, 1 \le j \le J\}$. From Eq. (A3.2), we have $\partial f / \partial w_j^+ = 0$ for $v_j^+ \in S_{\neq 0}$. Then, the Hessian matrix of $G(\bar{\mathrm{v}})$, evaluated at $\bar{\mathrm{v}}^+$, is given by

$$\mathrm{H}(\bar{\mathrm{v}}^+) = \begin{bmatrix} \mathrm{A}_1 & 0 \\ 0 & \mathrm{A}_2 \end{bmatrix} \tag{A2.3}$$

where

$$\mathrm{A}_1 = \begin{bmatrix} 2\dfrac{\partial f}{\partial w_1^+} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 2\dfrac{\partial f}{\partial w_M^+} \end{bmatrix}$$

$$\mathrm{A}_2 = \begin{bmatrix} \mathrm{A}_3 & z \\ z^T & \dfrac{\partial^2 f}{\partial^2 \mathrm{b}^+} \end{bmatrix}$$

$$\mathrm{A}_3 = \begin{bmatrix} 4v_{M+1}^{+2}\dfrac{\partial^2 f}{\partial^2 w_{M+1}^+} & \cdots & 4v_{M+1}^+ v_J^+ \dfrac{\partial^2 f}{\partial w_{M+1}^+ \partial w_J^+} \\ \vdots & \ddots & \vdots \\ 4v_{M+1}^+ v_J^+ \dfrac{\partial^2 f}{\partial w_{M+1}^+ \partial w_J^+} & \cdots & 4v_J^{+2}\dfrac{\partial^2 f}{\partial^2 w_J^+} \end{bmatrix}$$

$$z = \left[2v_{M+1}^+ \frac{\partial^2 f}{\partial w_{M+1}^+ \partial \mathrm{b}^+}, \ldots, 2v_J^+ \frac{\partial^2 f}{\partial w_J^+ \partial \mathrm{b}^+} \right]^T$$

Here we have used the fact that $\partial f / \partial w_j^+ = 0$ for $v_j^+ \in S_{\neq 0}$. Since $f(\mathrm{w}, \mathrm{b})$ is a convex function of w and b, we have

$$\mathrm{B} = \begin{bmatrix} \dfrac{\partial^2 f}{\partial w_{M+1}^2} & \cdots & \dfrac{\partial^2 f}{\partial w_{M+1} \partial w_J} & \dfrac{\partial^2 f}{\partial w_{M+1} \partial \mathrm{b}} \\ \vdots & \ddots & \vdots & \vdots \\ \dfrac{\partial^2 f}{\partial w_{M+1} \partial w_J} & \cdots & \dfrac{\partial^2 f}{\partial w_J^2} & \dfrac{\partial^2 f}{\partial w_J \partial \mathrm{b}} \\ \dfrac{\partial^2 f}{\partial w_{M+1} \partial \mathrm{b}} & \cdots & \dfrac{\partial^2 f}{\partial w_J \partial \mathrm{b}} & \dfrac{\partial^2 f}{\partial \mathrm{b}^2} \end{bmatrix} \ge 0,$$

It is easy to prove that

$$\mathrm{C}=\begin{bmatrix} 4v_{M+1}^2 & \cdots & 4v_{M+1}v_J & 2v_{M+1} \\ \vdots & \ddots & \vdots & \vdots \\ 4v_{M+1}v_J & \cdots & 4v_J^{+2} & 2v_J \\ 2v_{M+1} & \cdots & 2v_J & 1 \end{bmatrix} \geq 0,$$

Hence, by Schur product theorem (Horn *et al.*, 1985), $\mathrm{A}_2=(\mathrm{B}\otimes\mathrm{C})\big|_{\bar{\mathrm{v}}=\bar{\mathrm{v}}^+}$ is a positive semi-definite matrix. It follows that $\mathrm{H}(\bar{\mathrm{v}}^+)\geq 0$ if and only if $\mathrm{A}_1\geq 0$.

If $\mathrm{H}(\bar{\mathrm{v}}^+)$ is not positive semi-definite, then $\bar{\mathrm{v}}^+$ is a saddle point (Note that it cannot be a maximizer because $G(\bar{\mathrm{v}})$ is convex with respect to v_{J+1} and $\mathrm{H}(\bar{\mathrm{v}}^+)$ cannot be negative semi-definite). If $\mathrm{H}(\bar{\mathrm{v}}^+)\geq 0$, $\bar{\mathrm{v}}^+$ can be either a saddle point, or a local or global minimizer. We now prove that if $\mathrm{H}(\bar{\mathrm{v}}^+)\geq 0$, $\bar{\mathrm{v}}^+$ must be a global minimizer.

Let us first consider the following optimization problem:

$$\min f(\mathrm{w},\mathrm{b}),\ \text{subject to}\ \mathrm{w}\geq 0 \qquad \text{(A2.4)}$$

Since both the objective function and constraints are convex, the Karush–Kuhn–Tucker (KKT) conditions are the sufficient conditions of a global optimal solution. It can be shown that $(\mathrm{w}^+,\mathrm{b}^+)$ is a global minimizer of $f(\mathrm{w},\mathrm{b})$, if for all $\forall j\in[\Im]$ the following KKT conditions hold simultaneously:

1) $\partial f/\partial \mathrm{b}^+=0,$
2) $\partial f/\partial w_j^+=0,$ or $w_j^+=0$ and $\partial f/\partial w_j^+\geq 0,$

Since $\bar{\mathrm{v}}^+$ is a stationary point, by Eq. (A3.2),

$\forall i\in\left\{i:v_i^+\in S_{\neq 0}\right\}, \partial f/\partial w_j^+=0$ and $\partial f/\partial \mathrm{b}\big|_{b=v_{J+1}^+}=0$

Moreover, $\mathrm{H}(\bar{\mathrm{v}}^+)\geq 0$ implies that $\mathrm{A}_1\geq 0$. Since A_1 is a diagonal matrix, it holds that

$$\partial f/\partial w_j^+\geq 0, \forall i\in\left\{i:v_i^+\in S_{\neq 0}\right\}$$

Hence by the KKT conditions, (w^+,v_{J+1}^+) and $\bar{\mathrm{v}}^+$ is a global minimizer of $f(\mathrm{w},\mathrm{b})$ and $G(\bar{\mathrm{v}})$, respectively.

Proof Theorem 3.2

Suppose that $\partial G/\partial\bar{\mathrm{v}}^*=0$ and $\bar{\mathrm{v}}^*$ is a saddle point. Again, we assume that the first *M* elements of $\bar{\mathrm{v}}^*$ belong to S_0, while the following $J-M$ elements belong to $S_{\neq 0}$. There exists an element $j\in S_0$ so that $\partial f/\partial w_j^*<0$ (otherwise $\mathrm{H}(\bar{\mathrm{v}}^*)\geq 0$ and $\bar{\mathrm{v}}^*$ is a global minimizer). Due to the

continuity, there exists $\xi > 0$, such that $\partial f / \partial w_j < 0$ for every $w_j \in \Omega = \{w : |w - w_j^*| < \xi\}$. It follows that $\partial G / \partial v_j = 2v_j(\partial f / \partial w_j) < 0$ for $v_j = \sqrt{w_j}$, and $\partial G / \partial v_j > 0$ for $v_j = -\sqrt{w_j}$. That is, a gradient descent method given by $v_j^{(k+1)} \leftarrow v_j^{(k)} - \eta(\partial G / \partial v_j^{(k)})$, drives the solution out of the neighborhood of a saddle point except when (1) the component $v_j^{(k)}$ is set to *exactly* zero, or (2) $v_j^{(k)}$ is outside Ω and a line search hits $\bar{\mathrm{v}}^*$ exactly. The latter event cannot happen since gradient $\mathrm{g}(\bar{\mathrm{v}}^*)$ equals to zero at $\bar{\mathrm{v}}^*$, and thus the descending condition does not hold (see Definition 1). Instead, a line search will find a solution around $\bar{\mathrm{v}}^*$, and in the subsequent steps the solution will move away from $\bar{\mathrm{v}}^*$. On the contrary, if $\bar{\mathrm{v}}^*$ is a global optimal, $\bar{\mathrm{v}}^{(k+1)}$ will approach it continuously with improved solution quality. We go on to prove that if $v_j^{(0)} \neq 0$, $v_j^{(k)}$ will be set to exactly zero at a non-stationary point with a zero probability. Let $\bar{\mathrm{v}}^{(k)}$ be the solution obtained in the *k*-th iteration, $-\mathrm{d}^{(k)}$ be the descending direction, $\bar{\mathrm{v}}^{(k)-} = \bar{\mathrm{v}}^{(k)} - \eta\,\mathrm{d}^{(k)}$ be a point in the line-search path at which some elements are zeros, and $\mathrm{g}^{(k)-}$ be the gradient at $\bar{\mathrm{v}}^{(k)-}$. Since $\bar{\mathrm{v}}^{(k)-}$ is not a stationary point, $\|\mathrm{g}^{(k)-}\| > 0$.

1. If $\mathrm{d}^{(k)T}\mathrm{g}^{(k)-} \leq 0$, the line search will not reach $\bar{\mathrm{v}}^{(k)-}$.

2. On the other hand, If $\mathrm{d}^{(k)T}\mathrm{g}^{(k)-} > 0$, $-\mathrm{d}^{(k)}$ is also a descending direction at $\bar{\mathrm{v}}^{(k)-}$. Without loss of generality, we assume $\|\mathrm{d}^{(k)}\| = 1||$. Due to the continuity, there exists a $\zeta_1 > 0$ such that for all $\bar{\mathrm{v}} \in \Omega_2 = \{\bar{\mathrm{v}} : \|\bar{\mathrm{v}} - \bar{\mathrm{v}}^{(k)-}\| < \xi_1\}$, $\mathrm{d}^{(k)T}\mathrm{g}(\bar{\mathrm{v}}) > 0$. This means that for any $\alpha \in (0,1)$, $G\left(\bar{\mathrm{v}}^{(k)-} - \alpha\xi_1 d^{(k)}\right) < G\left(\bar{\mathrm{v}}^{(k)-}\right)$.

a) If the chosen interval length $\varepsilon^{(k)} < \xi_1$, the line search has at least one candidate solutions which is correspond with an $\alpha > 0$ above, hence it goes pass $\bar{\mathrm{v}}^{(k)-}$ and moves away from it.

b) On the other hand, if $\varepsilon^{(k)} \geq \xi_1$, the line search will have only one or two candidate solutions within Ω_2. In this case, the line search has no prior knowledge about the region under search, thus it degenerates to *randomly* selecting one or two $\alpha \in (-1,1)$ and setting $\bar{\mathrm{v}}^{(k+1)} = \bar{\mathrm{v}}^{(k)} - \alpha\xi_1\,\mathrm{d}^{(k)}$. Given an arbitrary bounded probability density distribution, the probability that a single point $\alpha = 0$ is chosen, however, is zero.

Moreover, due to the continuity, there also exists a $\zeta_2 > 0$ such that for all $\bar{\mathrm{v}} \in \Omega_3 = \left\{ \bar{\mathrm{v}} : \left\| \bar{\mathrm{v}} - \bar{\mathrm{v}}^{(k)-} \right\| < \xi_2 \right\}$, $g(\bar{\mathrm{v}})^T \mathrm{g}^{(k)-} > 0$. This mean that a gradient-based search starting in Ω_3 will go pass $\bar{\mathrm{v}}^{(k)-}$ following case (2a) stated above, hence $\bar{\mathrm{v}}^{(k)-}$ is not a point of attraction and after a few iterations, case (1) and case (2b) either change to case (2a), or change to the case that the descending direction does not drive any element towards zero.

This completes the proof that the saddle points cannot be reached with the designated gradient descent method.

Appendix 3

To solve the stated optimization problem, the well known Lagrangian optimization method is applied. The Lagrangian of (5.8) is given by

$$L = -(w_f)^T s + \lambda(\|w_f\|_2^2 - 1) + \sum_{i=1}^{m} \xi_l(-w_{fi}) \qquad \text{(A3.1)}$$

where λ and $\xi \geq 0$ are the Lagrange and Kuhn-Tucker multipliers.

Applying derivative to (A1) with respect to w_f and setting it to zero, a closed-form solution for w_f can be derived:

$$\frac{\partial L}{\partial w_f} = -s + 2\lambda w_f - \xi = 0 \Rightarrow w_f = \frac{1}{2\lambda(s + \xi)} \qquad \text{(A3.2)}$$

with the assumption (5.4.1), the positivity of λ is proved by contradiction. Suppose $\lambda<0$ and the assumption $s_i>0$ we have

$$s_i + \xi_l > 0$$

$$\Rightarrow \quad w_{fi} = \frac{(s_i + \xi_i)}{2\lambda} < 0$$

This result is contradictory with constraint $w_f \geq 0$, thus $\lambda>0$.

By application of Kuhn-Tucker condition, namely $\sum_i \xi_l w_{fi}=0$, the following three cases can be verified

1) $s_i=0 \Rightarrow \xi_l=0$ and $w_{fi}=0$
2) $s_i>0 \Rightarrow s_i + \xi_l>0 \Rightarrow w_{fi}>0 \Rightarrow \xi_l=0$
3) $s_i<0 \Rightarrow \xi_l>0 \Rightarrow w_{fi}=0 \Rightarrow s_i=-\xi_l$

Then, the optimum solution of w_f can be calculated in the following closed form:

$$w_{fi} = \begin{cases} 0 & \text{if } s_i \leq 0 \\ \dfrac{1}{2\lambda s_i} & \text{if } s_i > 0 \end{cases}$$

Therefore the normalized values are:

$$w_f^* = \frac{s^+}{\| s^+ \|} \qquad \text{(A3.4)}$$

with $s^+ = [\max(s_1,0), \ldots, \max(s_m,0)]^T$

Appendix 4

A4.1 Receiver operating characteristic

In signal detection theory, a **receiver operating characteristic** (**ROC**), or simply **ROC curve**, is a graphical plot of the sensitivity, or true positive rate, vs. false positive rate (1 − specificity or 1 − true negative rate), for a binary classifier system as its discrimination threshold is varied. The ROC can also be represented equivalently by plotting the fraction of true positives out of the positives (TPR = true positive rate) vs. the fraction of false positives out of the negatives (FPR = false positive rate). Also known as a Relative Operating Characteristic curve, because it is a comparison of two operating characteristics (TPR & FPR) as the criterion changes (Swets, 1996).

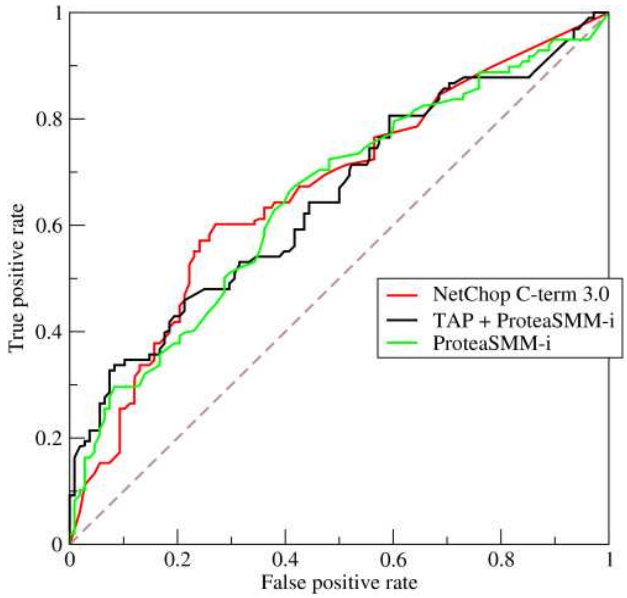


Fig. A4.1 An example of ROC curve. Image taken from Wikipedia.

ROC analysis provides tools to select possibly optimal models and to discard suboptimal ones independently from (and prior to specifying) the cost context or the class distribution. ROC analysis is related in a direct and natural way to cost/benefit analysis of diagnostic decision making. The ROC curve was first developed by electrical engineers and radar engineers during World War II for detecting enemy objects in battle fields, also known as the signal detection theory, and was soon introduced in psychology to account for perceptual detection of stimuli. ROC analysis since then has been used in medicine, radiology, and other areas for many decades, and it has been introduced relatively recently in other areas like machine learning and data mining.

A classification model (classifier or diagnosis) is a mapping of instances into a certain class/group. The classifier or diagnosis result can be in a real value (continuous output) in which the classifier boundary between classes must be determined by a threshold value, for instance to determine whether a person has hypertension based on blood pressure measure, or it can be in a discrete class label indicating one of the classes.

Let us consider a two-class prediction problem (binary classification), in which the outcomes are labeled either as positive (*p*) or negative (*n*) class. There are four possible outcomes from a binary classifier. If the outcome from a prediction is *p* and the actual value is also *p*, then it is called a *true positive* (TP); however if the actual value is *n* then it is said to be a *false positive* (FP). Conversely, a *true negative* has occurred when both the prediction outcome and the actual value are *n*, and *false negative* is when the prediction outcome is *n* while the actual value is *p*.

To get an appropriate example in a real-world problem, consider a diagnostic test that seeks to determine whether a person has a certain disease. A false positive in this case occurs when the person tests positive, but actually does not have the disease. A false negative, on the other hand, occurs when the person tests negative, suggesting they are healthy, when they actually do have the disease.

Let us define an experiment from P positive instances and N negative instances. The four outcomes can be formulated in a 2×2 *contingency table* or *confusion matrix*, as follows:

		Actual value		
		p	**n**	**total**
Prediction outcome	**p′**	True Positive	False Positive	**P′**
	n′	False Negative	True Negative	**N′**
	total	**P**	**N**	

Tab. A4.1

The contingency table can derive several evaluation “metrics”. To draw an ROC curve, only the true positive rate (TPR) and false positive rate (FPR) are needed. TPR determines a classifier or a diagnostic test performance on classifying positive instances correctly among all positive samples available during the test. FPR, on the other hand, defines how many incorrect positive results occur among all negative samples available during the test. A ROC space is defined by FPR and TPR as *x* and *y* axes respectively, which depicts relative trade-offs between true positive (benefits) and false positive (costs). Since TPR is equivalent with sensitivity and FPR is equal to 1−specificity, the ROC graph is sometimes called the sensitivity vs (1−specificity) plot. Each prediction result or one instance of a confusion matrix represents one point in the ROC space.

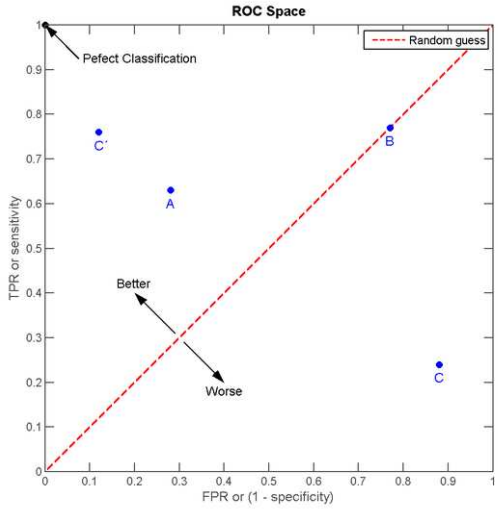


Fig. A4.2 An example of ROC curve. Image taken from Wikipedia.

The best possible prediction method would yield a point in the upper left corner or coordinate (0,1) of the ROC space, representing 100% sensitivity (no false negatives) and 100% specificity (no false positives). The (0,1) point is also called a *perfect classification*. A completely random guess would give a point along a diagonal line (the so-called *line of no-discrimination*) from the left bottom to the top right corners. An intuitive example of random guessing is a decision by flipping coins (head or tail).

The diagonal divides the ROC space. Points above the diagonal represent good classification results, points below the line poor results. Note that the output of a poor predictor could simply be inverted to obtain points above the line (See Figure A4.2).

A4.2 Kaplan-Meier Curve

The Kaplan–Meier estimator (Kaplan and Meier, 1958; Kaplan and Meier, 1983), also known as the product limit estimator, is an estimator for estimating the survival function from life-time data. In medical research, it is often used to measure the fraction of patients living for a certain amount of time after treatment. The estimator is named after Edward L. Kaplan and Paul Meier.

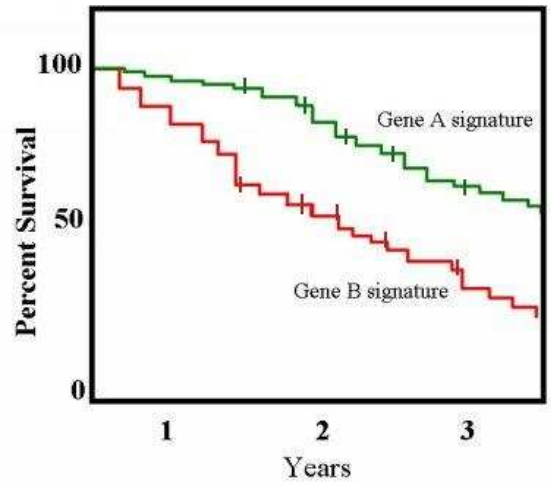


Fig. A4.3 An example of Kaplan-Meier curve. Image taken from Wikipedia.

A plot of the Kaplan–Meier estimate of the survival function is a series of horizontal steps of declining magnitude which, when a large enough sample is taken, approaches the true survival function for that population. The value of the survival function between successive distinct sampled observations ("clicks") is assumed to be constant.

An important advantage of the Kaplan–Meier curve is that the method can take into account some types of censored data, particularly *right-censoring*, which occurs if a patient withdraws from a study, i.e. is lost from the sample before the final outcome is observed. On the plot, small vertical tick-marks indicate losses, where a patient's survival time has been right-censored. When no truncation or censoring occurs, the Kaplan–Meier curve is equivalent to the empirical distribution function. In medical statistics, a typical application might involve grouping patients into categories, for instance, those with Gene A profile and those with Gene B profile. In the graph, patients with Gene B die much more quickly than those with gene A. After two years, about 80% of the Gene A patients survive, but less than half of patients with Gene B.

Let $S(t)$ be the probability that an item from a given population will have a lifetime exceeding t. For a sample from this population of size N let the observed times until death of N sample members be

$$t_1 \leq t_2 \leq t_3 \leq \ldots \leq t_N$$

Corresponding to each t_i is n_i, the number "at risk" just prior to time t_i, and d_i, the number of deaths at time t_i.

Note that the intervals between each time typically are not uniform. For example, a small data set might begin with 10 cases, have a death at Day 3, a loss (censored case) at Day 9, and another death at Day 11. Then we have ($t_1 = 3$, $t_2 = 11$), ($n_1 = 10$, $n_2 = 8$), and ($d_1 = 1$, $d_2 = 2$).

The Kaplan–Meier estimator is the nonparametric maximum likelihood estimate of $S(t)$. It is a product of the form

$$\hat{S}(t) = \prod_{t_i \prec t} \frac{n_i - d_i}{n_i} \qquad \text{A4.1}$$

When there is no censoring, n_i is just the number of survivors just prior to time t_i. With censoring, n_i is the number of survivors less the number of losses (censored cases). It is only those surviving cases that are still being observed (have not yet been censored) that are "at risk" of an (observed) death (Costella, 2010).

There is an alternative definition that is sometimes used, namely

$$\hat{S}(t) = \prod_{t_i \leq t} \frac{n_i - d_i}{n_i} \qquad \text{A4.2}$$

The two definitions differ only at the observed event times. The latter definition is right-continuous whereas the former definition is left-continuous.

Let T be the random feature that measures the time of failure and let $F(t)$ be its cumulative distribution function. Note that

$$S(t) = P[T > t] = 1 - P[T \leq t] = 1 - F(t) \qquad \text{A4.3}$$

Consequently, the right-continuous definition of $\hat{S}(t)$ may be preferred in order to make the estimate compatible with a right-continuous estimate of $F(t)$

References

Abba M., Hu Y., Sun H., *et al.*, Gene expression signature of estrogen receptor α status in breast cancer, BMC Genomics, 6, pp.74-81, 2005.

Abe S. and Thawonmas R., A fuzzy classifier with ellipsoidal regions, IEEE Tans. Fuzzy Syst., 5, pp. 358-368, 1997.

Aguado J.C., Aguilar-Martin J., A mixed qualitative-quantitative self-learning classification technique applied to diagnosis, The Thirteenth Int'l Workshop on Qualitative Reasoning Chris Price, pp. 124-128, 1999.

Aguilar J., Lopez De Mantaras R., The process of classification and learning the meaning of linguistic descriptors of concepts, Approximate reasoning in decision analysis, pp. 165-175, 1982.

Aha D.W., Incremental, instance-based learning of independent and graded concept descriptions, in Proced. of the 6th int'l Mach. Learning Workshop, pp. 387-391, 1989.

Aha D.W., Tolerating noisy, irrelevant and novel attributes in instance based learning algorithms, *Int. Man-Machine Studies* 36, pp. 267-287, 1992.

Alizadeh A.A., Eisen M.B., Davis R.E., *et al.*, Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, Nature, 403, pp. 503-511, 2000.

Andrews R., Mah R., Jeffery S., Guerrero M., papasin R., Reed C., The NASA smart probe project for real-time multiple microsensor tissue recognition, Int'l Congress Series, 1256, pp. 547-554, 2003.

Antman K., Shea S., screening mammography under age 50. The Journal of the American Medical Association, 2, pp. 281-1470, 1999.

Association pour la recherche sur le cancer http://www.arc-cancer.net/

Atkeson C.G., Moore A.W., Schaal S., Locally Weighted Learning, Artificial Intelligence Rev., 11 (15), pp. 11–73, 1997.

Ayers M, Symmans W.F., Stec J, *et al.*, Gene expression profiles predict complete pathologic response to neoadjuvant paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide chemotherapy in breast cancer, J Clin Oncol 22, pp. 2284-2293, 2004.

Baldi P., Brunak S., *Bioinformatics: The machine learning approach*, MIT Press, Cambridge, 2001.

Baraldi A., Blonda P., A survey of fuzzy clustering algorithms for pattern recognition, IEEE Trans. on Systems, Man, and Cybernetics, Part B, 29 (6), pp. 778-785,1999.

Bardou V.J., Arpino G., Alledge R.M., *et al.*, Progestrone receptor status significantly improves outcome prediction over estrogen receptor status alone for adjuvant endocrine therapy in two large breast cancer database, J Clin Oncol, 21, pp. 1973-1979, 2003.

Belle V.V., Calster B.V., Brouckaert O., *et al.*, Qualitative Assessment of the Progesterone Receptor and HER2 Improves the Nottingham Prognostic Index Up to 5 Years After Breast Cancer Diagnosis, J of Clin Onco, 28 (27), pp. 4129-4134, 2010.

Bellman R., *Adaptive control processes: A Guided Tour*, Princeton University Press, 1961.

Bertucci F., Nasser V., Granjeaud S., *et al.*, Gene expression profiles of poor-prognosis primary breast cancer correlate with survival, Oxford Journals Life Sciences & Medicine Human Molecular Genetics,11 (8), pp. 863-872, 2002.

Bezdec J.C., *Pattern Recognition with Fuzzy Objective Function Algorithms,* Plenum Press, New York, 1981.

Bhattacharjee A., Richards W., Staunton J., *et al.*, Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses, Proc. Nat'l Acad. Sc. USA, 98 (24), pp. 13790-13795, 2001.

Billard L., Somes Analyses of Interval Data, J of Comp and Info Tech, 16, pp. 225-233, 2008.

Blake C., Merz C., UCI Repository of Machine Learning Databases, 1998.

Blake W.J., Kærn M., Cantor C.R., Collins J.J., noise in eukaryotic gene expression, Nature, 422 (6932), pp. 633-637, 2003.

Blanco R., Larrañaga P., Inza I., Sierra B., Gene selection for cancer classification using wrapper approaches, Int'l J of Pattern Recognition and Artificial Intelligence (IJPRAI), 18 (8), pp. 1373-1390 , 2004.

Bloom H.J., Richardson W.W., Histological grading and prognosis in breast cancer; a study of 1409 cases of which 359 have been followed for 15 years, Br J Cancer, 11, pp. 359-377, 1957.

Bock H.H., Diday E., Analysis of Symbolic Data, Exploratory Methods for Extracting Statistical Information from Complex Data. Springer-Verlag, Heidelberg, 2000.

Bradley P.S. and Mangasarian O.L., Feature selection via concave minimization and support vector machines, In J. Shavlik, editor, Mach. Learn. Proc. of the 15th Intl. Conf. Mach. Learn., pp. 82-90, 1998.

Breiman L., Friedman J.H., Olshen R.A., Stone C.J., *Classification and regression trees*, Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software, 1984.

Buchholz T.A., Stivers D.N., Stec J., *et al.*, Global gene expression changes during neoadjuvant chemotherapy for human breast cancer, Cancer J, 8, pp. 461-468, 2002.

Buness A., Ruschhaupt M., Kuner R., Tresch A., Classification across gene expression microarray studies. BMC Bioinformatics, pp. 410-453, 2009.

Buyse M., Loi S., van't Veer L., Viale G., *et al.*, Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer, *J. N. Cancer Inst.*, 98(17), pp. 1183-1192, 2006.

Byrnes K., White S., Chu Q., *et al.*, High eIF4E, VEGF, and microvessel density in stage I to III breast cancer, Ann Surg, 243, pp. 684-690, 2006.

Cai Y., Sun Y., Li J., Goodison S., Online Feature Selection Algorithm with bayesian ℓ_1 Regularization, in T. Theeramunkong et al. (Eds): PAKDD, LNAI 5476, pp. 401-413, 2009.

Cai Y., Sun Y., Cheng Y., Li J., Goodison S., Fast Implementation of ℓ_1 Regularized Learning Algorithms Using Gradient Descent Methods, SDM, pp. 862-871, 2010a.

Cai Y., Hedjazi L., Sun Y., Goodison S., Fast Implementation of $\ell 1$ Regularized Learning Algorithms Using Gradient Descent Methods, submitted to IEEE Trans. on Pattern Analysis and Machine Intelligence 2010b.

Carter C.L., Allen C., Henson D.E., Relation of tumor size, lymph node status, and survival in 24740 breast cancer cases, Cancer, 63, pp. 181-187, 1989.

Chang J.C., Wooten E.C., Tsimelzon A., *et al.*, Gene expression profiling predicts therapeutic response to docetaxel (Taxotere) in breast cancer patients, Lancet 362, pp. 280-287, 2003a.

Chang J.C., Hilsenbeck S.G., Fuqua S.A.W., Genomic approaches in the management and treatment of breast cancer, British Journal of Cancer, 92, pp. 618-624, 2005.

Chang R-F., Wu W-J., Moon W.K., *et al.*, Support Vector Machines for Diagnosis of Breast Tumors on US Images, Academic Radiology, 10 (2), pp. 189-197, 2003b.

Chapelle O., Training a support vector machine in the primal, Neural Comput., 19, pp. 1155-1178, 2007.

Chen S-M., A weighted fuzzy reasoning algorithm for medical diagnosis, decision support systems, 11 (1), pp. 37-43, 1994.

Cheng C-J., Lin Y-C., Tsai M-T., *et al.*, SCUBE2 Suppresses Breast Tumor Cell Proliferation and Confers a Favorable Prognosis in Invasive Breast Cancer, Cancer Res 69, pp. 3634-3641, 2009.

Cheng Y. and Church G.M., Biclustering of expression data, Proc Int Conf Intell Syst Mol Biol, 8, pp. 93-103, 2000.

Cheung K.L., Graves C.R., Robertson J. F. R., Tumour marker measurements in the diagnosis and monitoring of breast cancer, Cancer Treatment Reviews, 26, pp. 91-102, 2000.

Chiu S.L., *Extracting fuzzy rules from data for function approximation and pattern classification*, In Fuzzy Information Engineering: A guided Tour of Applications, ed. D. Dubois, H. Prade, and R. Yager, John Wiley & Sons, 1997.

Cicchetti D.V., Neural networks and diagnosis in the clinical laboratory: state of the art. Clin Chem, 38, pp. 9-10, 1992.

Cleator S.J., Makris A., Ashley S.E., Lal R., Powles T.J., Good clinical responses of breast cancers to neoadjuvant chemoendocrine therapy is associated with improved overall survival, Annals of Oncology, 16 (2), pp. 267-272, 2004.

Cochran A.J., prediction of outcome for patients with cutaneous melanoma, Pigment Cell Res, 10, pp. 162-167, 1997

Colozza M., Azambuja E., Cardoso F., *et al.*, Proliferative markers as prognostic and predictive tools in early breast cancer: where are we now?, Ann Oncol, 16(11), pp. 1723-1739, 2005.

Costella J.P., *A simple alternative to Kaplan–Meier for survival curves*, 2010.

Covell D.G., Wallqvist A., Rabow A.A., Thanki N., Molecular Classification of Cancer: Unsupervised Self-Organizing Map Analysis of Gene Expression Microarray Data, Mol Cancer Ther, *2*, pp. 317-332, 2003.

Cover T., Hart P., Nearest neighbor pattern classification, IEEE Trans. Inf. Theory 13 (1967) 21-27.

Cross V.V. and Sudkamp T.A., Similarity and Compatibility in Fuzzy Set Theory: assessment and Applications, Physica-Verlag, New York (2002).

Cruz J.A., Wishart D.S., Application of Machine Learning in Cancer Prediction and Prognosis, Cancer Informatics, 2, pp. 59-77, 2006.

Das S., Filters, Wrappers, and a boosting-based Hybrid for feature selection, Proc. 18th Int'l Conf. Machine Learning, pp. 74-81, 2001.

Dash M., Liu H., Consistency-based search in feature selection, *Artif. Intell.* 151, 155-176, 2003.

De Carvalho F.A.T., De Souza R.M.C.R., Unsupervised Pattern Recognition Models for Mixed Feature-Type Symbolic Data, Pattern Recognition Letters, 31, pp. 430–443, 2010.

Deepa S., Claudine I., Utilizing Prognostic and Predictive Factors in Breast Cancer. Current Treatment Options in Oncology, 6, pp. 147-159. Current Science Inc., 2005.

Delen D. ,Walker G., Kadam A. , Predicting breast cancer survivability: a comparison of three data mining methods, 34, 2, pp. 113-127, 2005.

De Souto M.,Costa I., De Araujo D., *et al.*, Clustering cancer gene expression data: a comparative study. BMC Bioinformatics, 9(1), 497, 2008.

Dettling M, Gabrielson E, Parmigiani G: Searching for di®erentially expressed gene combinations. *Genome Biology* 6:R88, 2005,.

Diab S.G., Clark G.M., Osborne C.K., *et al*., Tumor characteritics and clinical outcome of tubular and mucinous breast carcinomas, J Clin Oncol, 17, pp. 1442-1448, 1999.

Dietterich T.G., Machine Learning Research: Four Current Directions, AI Magazine 18 (4), pp. 97-136, 1997.

Dubois D., Prade H., Testemale C., Weighted fuzzy pattern matching, Fuzzy Sets and Systems, 28, 3, pp. 313-331, 1988 .

Dubois D., Prade H., The three semantics of fuzzy sets, Fuzzy Sets and Syst., 90, pp. 141-150, 1997.

Duchi J., Shalev-Shwartz S., Singer Y., and Chandra T., Efficient projections onto the L1-ball for learning in high dimensions, Proc. 25th Intl. Conf. Mach. Learn., pp. 272-279, 2008.

Duda R.O., Hart P.E., Stork D.G., *pattern classification*, Wiley-interscience, 2001.

Dudoit S., Fridlyand J., Speed T.P, Comparison of discrimination methods for the classification of tumors using gene expression data, J. Am. Stat. Assoc, 97, pp. 77-87, 2002.

Duffy M.J., Duggan C., The urokinase plasminogen activator system: a rich source of tumor markers for the individualised management of patients with cancer, Clin Biochem, 37, pp. 541-548, 2004.

Duggan D.J., Bittner M., Chen Y., *et al.*, expression profiling using cDNA microarrays, Nature genetics, 21, pp. 10-37, 1999.

Dy J.G., Brodley C.E., Feature Selection for unsupervised learning, JMLR, 5, pp. 845-889, 2004.

Eifel P., Axelson J.A., Costa J., *et al.*, National institutes of health consensus development conference statement: Adjuvant therapy for breast cancer. Journal of National Cancer Institute, 93(13), pp. 979-989, 2001.

Evtushenkjo Y.G. and Zhadan V.G., Space-transformation technique: the state of the art, *In Nonlinear Optimization and Applications*, pp 101-123, 1996.

Fisher R. A., The use of multiple measurements in taxonomic problems, Annals of Eugenics, 7, pp. 179-188, 1936.

Faybusovich L., Dynamical systems which solve optimization problems with linear constraints, IMA J. Math. Ctrl. and Info., 8, pp. 135-149, 1991;

Filippone M., Camastra F., Masulli F., Rovetta S., A survey of kernel and spectral methods for clustering, Pattern recognition, 41, pp. 176-190, 2008.

Fisher B., Bryant J., Wolmark N., et al., Effect of preoperative chemotherapy on the outcome of women with operable breast cancer, J of Clin Onco, 16, pp. 2672-2685, 1998.

Fletcher R., *Practical methods of optimization*, John Wiley, New York, 1997.

Freund Y., Schapire R.E., A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting, J. Computer Systems Science, 55 (1), pp. 119-139, 1997.

Fujita T., Doihara H., Kawasaki K., *et al.*, PTEN activity could be a predicitive marker of trastuzumab efficacy in the treatment of ErbB2-overexpressing breast cancer, British Journal of Cancer, 94, pp. 247-252, 2006.

Fukunaga K., *Introduction to statistical pattern recognition*, 2nd ed. New York: Academic, 1990.

Fung G. and Mangasarian O.L., A feature selection Newton method for support vector machine classification, Computational Optimization and Applications, 28 (2), pp. 185-202, 2004.

Galea M.H., Blamey R.W., Elston C.E., and Ellis I.O., The nottingham prognostic index in primary breast cancer, Breast Cancer Research and Treatment, 22(3), pp. 207-219, 1992.

Gallardo-Caballero R., García-Orellana C.J., González-Velasco H.M., Macías-Macías M., Independent component analysis applied to detection of early breast cancer signs, in Proceedings of the 9th international work conference on Artificial neural networks, pp. 988-995, 2007.

Gevaert O., De Smet F., Timmerman D., Moreau Y., De Moor B., Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian network, Bioinformatics, 22 (14), 184-190, 2006.

Gilad-Bachrach R., Navot A., Tishby N., Margin Based feature selection-theory and algorithms, In proceed. 21st Int'l Conf. on Machine Learning, ACM Press, pp. 43-50, 2004.

Goldhirsh A., Wood W.C., Gelber R.D., *et al.*, Meeting highlights: Updated international expert consensus on the primary therapy of early breast cancer, Journal of Clinical Oncology, 21(17), pp. 3357–3365, 2003.

Golub T., Slonim D., Tamayo P., *et al.*, Molecular Classification of Cancer Class Discovery and Class Prediction by Gene Expression Monitoring, Science, 286 (5439), pp. 531-537, 1999.

Gómez-Ruiz J.A., Jerez-Aragonés J.M., Muñoz-Pérez J., Alba-Conejo E., A Neural Network Based Model for Prognosis of Early Breast Cancer, Applied Intelligence, 20, pp. 231-238, 2004.

Gowda K.C., Diday E., Symbolic clustering using a new similarity measure. IEEE Trans. Systems Man Cybernet. 22, pp. 368-378, 1992.

Grammer K., Gilad-Bachrach R., Navot A., Tishby N., Margin analysis of the lvq algorithm, In proceedings 17th Int'l Conf. on Neural Information Processing Systems, 2002.

Guyon I., Weston J., Barnhill S., Vapnik V., Gene selection for cancer classification using support vector machines, Mach. Learning, 46 (1–3), pp. 389-422, 2002.

Guyon I., Elisseeff A., An introduction to variable and feature selection, J. Mach. Learn. Res 3, pp. 1157-1182, 2003.

Guyon I., Gunn S.R., Ben-Hur A., and Dror G., Result analysis of the NIPS 2003 feature selection challenge, 17th Adv. Neu. Info. Proc. Sys., Vancouver, Canada, pp. 545-552, 2005.

Haffty B.G., Kornguth P., Fisher D., Beinfield M., McKhann C., Mammographically detected breast cancer: Results with conservative surgery and radiation therapy, Cancer 67, pp. 2801-2804, 1991.

Haibe-Kains M.B., *Identification and Assessment of Gene Signatures in Human Breast Cancer*, Thesis, Université Libre de Bruxelles, 2009.

Haibe-Kains M.B., Desmedt C., F. Rothé, *et al.*, A fuzzy gene expression-based computational approach improves breast cancer prognostication, Genome Biology, pp. 11-18, 2010.

Hall M.A., Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning, Int. Conf. Mach. Learning ICML, pp. 359-366, 2000.

Hayat M.A, *Methods of Cancer Diagnosis, Therapy and Prognosis: Breast Carcinoma*, 1, springer, 2008.

Hedjazi L., Kempowsky T., Le Lann M.V., Aguilar M.J., Classification floue des données intervallaires : Application au pronostic du cancer. Actes du XVIem Rencontres de la société francophone de classification SFC'09, pp. 165-168, 2009.

Hedjazi L., Aguilar-Martin J., Le Lann M.V., Kempowsky T., Fuzzy mechanisms for unified reasoning about heterogeneous data, 24th Int'l Workshop on Qualitative Reasoning, Portland, 2010a.

Hedjazi L., Kempowsky T., Le Lann M.V., Aguilar-Martin J., Prognosis of breast cancer based on a fuzzy classification method, 4th Int'l Joint Conf Biomed Eng Syst and Tech, 1st Int'l Conf. on Bioinformatics, pp. 123-130, 2010b.

Hedjazi L., Kempowsky T., Despenes L., Elgue S., Le Lann M.V., Aguilar M.J., Sensor placement and fault detection using an efficient fuzzy feature selection approach**,** 49th IEEE Int'l Conf. on Decision and Control CDC, pp. 6827-6832, 2010c.

Hedjazi L., Aguilar-Martin J., Le Lann M.V., Kempowsky T., Towards a Unified Principle for Reasoning about Heterogeneous Data: A Fuzzy Logic Framework, Submitted for publication in International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems (Under revision), 2011a.

Hedjazi L., Aguilar-Martin J., Le Lann M.V., Similarity-margin based feature selection for symbolic interval data, Pattern Recognition Letters, 32 (4), pp. 578-585, 2011b.

Hedjazi L., Aguilar-Martin J., Le Lann M.V., Kempowsky T., Membership-Margin based feature selection for Mixed-Type and High-Dimensional Data, submitted for publication in IEEE SMC, 2011c.

Hedjazi L., Kempowsky T., Le Lann M.V., F. Dalenc F., Favre G., Improved Breast Cancer Prognosis based on a Hybrid Marker Selection Approach, 5th Int'l Joint Conf Biomed Eng Syst and Tech 2nd Int'l Conf. on Bioinformatics, pp. 159-164, 2011d.

Hedjazi L., Kempowsky T., Le Lann M.V., Aguilar-Martin J., Une approche floue pour le placement des capteurs et la détection des fautes, QUALITA, 2011e.

Hedjazi L., Kempowsky T., Le Lann M.V., Aguilar M.J., Dalenc F., Favre G., Despenes L., Elgue S., From chemical process diagnosis to cancer prognosis: an integrated approach for diagnosis and sensor/marker selection, In: Pistikopoulos, E.N., Georgiadis, M.C., and Kokossis, A.C. (eds.) 21st European Symposium on Computer Aided Process Engineering (ESCAPE-21). Computer Aided Chemical Engineering, vol. 29. Amsterdam, Elsevier. pp. 1510-1514, 2011f.

Horn R., and Johnson C., *Matrix analysis*, Cambridge University Press, Cambridge, 1985.

Hu Q.H, Xie Z.X., Yu D.R., Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation, Pattern Recognition, 40, pp. 3509-3521, 2007.

Hu Q., Yu D., Liu J., Wu C., Neighborhood rough set based heterogeneous feature subset selection, Information sciences, 178, pp. 3577-3594, 2008.

Huber W., von Heydebreck A. , Sültmann H., Poustka A., Vingron M., Variance stabilization applied to microarray data calibration and to the quantification of differential expression, Bioinformatics, 18, pp. 96-104, 2002.

Hudelist G., Köstler W.J., Czerwenka K., *et al.*, HER2-neu/ and EGFR tyrosine kinase activation prediction the efficacy of trastuzumab-based therapy in patients with metastatic breast cancer, Int'l J of Cancer, 118 (5), pp. 1126-1134, 2005.

Institut de Veille Sanitaire, http://www.invs.sante.fr/

Institut National du Cancer http://www.e-cancer.fr

Institut National du Cancer, Rapport 2009 sur l'état des connaissance aux biomarqueurs uPA-PAI-1, Oncotype DX™ et Mammaprint® dans la prise en charge du cancer du sein, 2009.

Inza I., Larrañaga P., Blanco R. and Cerrolaza A.J., Filter versus wrapper gene selection approaches in DNA microarray domains, Artificial Intelligence in Medicine,31 (2), pp. 91-103, 2004.

Isaza C., Kempowsky T., Aguilar J., Gauthier A., Qualitative data Classification Using LAMDA and other Soft-Computer Methods, Recent Advances in Artificial Intelligence Research and Development, IOS Press, 2004.

Ishibuchi H., Nozaki K., Tanaka H., Distributed representation of fuzzy rules and its application to pattern classification, Fuzzy sets and syst., 52, pp. 21-32,1992.

Ishibuchi H., Nakashima T., Effect of rule weights in fuzzy rule-based classification systems, IEEE Tans. Fuzzy Syst., 9, pp. 506-515, 2001.

Ishibuchi H., Yamamoto T., Rule weight specification in fuzzy rule-based classification systems, IEEE Tans. Fuzzy Syst., 13, pp. 428-435, 2005.

Ivshina A.V., George J., Senko O., *et al.*, Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer, Cancer Res., 66 (21), pp. 10292-10301, 2006.

Jaccard P., Nouvelles recherches sur la distribution florale, Bulletin de la Société de Vaud des Sciences Naturelles, 44, 223, 1908.

Jahromi M.Z, Taheri M., A proposed method for learning rule weights in fuzzy rule-based classification systems, Fuzzy sets and syst., 159, pp. 494-459, 2008.

Jain A.K., Dubes R.C., *Algorithms for clustering data*, Prentice-Hall, 1988.

Jain A.K, Murty M.N., Flynn P.J., Data clustering: a review, Journal of ACM computing surveys, 31 (3), 1999.

Janicke F., Prechtl A., Thomssen C., *et al.* Randomized adjuvant chemotherapy trial in high-risk, lymph node-negative breast cancer patients identified by urokinase-type plasminogen activator and plasminogen activator inhibitor type 1, J Natl Cancer Inst, 93, pp. 913-920, 2001.

Jenssen T. and Hovig E., Gene-Expression Profiling in Breast Cancer, Lancet, 365, pp. 634-635, 2005.

Jerez J.M., Peláez J.I., Condoretty A., Alba E. , A Neuro-fuzzy Decision Model for Prognosis of Breast cancer relapse, In R. Conejo et al. (Eds.), LNAI 3040, pp. 638-645, 2004.

Kaplan E.L., In a retrospective on the seminal paper in "This week's citation classic", Current Contents, 24 (14),1983.

Kaplan E. L.; Meier P., Nonparametric estimation from incomplete observations, J. Amer. Statist. *Assn.*, 53, pp. 457-481, 1958.

Khan J., Wei J., Ringner M., *et al.*, Classification and Diagnostic Prediction of Cancers Using Gene Expression Profiling and Artificial Neural Networks, Nat. Med., 7 (6), pp. 673-679, 2001.

Khan M.U., Choi J.P., Shin H., Kim M., Predicting breast cancer survivability using fuzzy decision trees for personalized healthcare, Conf Proc IEEE Eng Med Biol Soc., pp. 5148-5151, 2008.

Kiefer J., Sequential minimax search for a maximum, Proc. 4th Amer Math. Soc., pp. 502-506, 1953.

Kira K., Rendell L., A practical approach to feature selection. In proced. 9th Int'l Workshop on Machine Learning, pp. 249-256, 1992a.

Kira K., Rendell L., The Feature Selection Problem: Traditional Methods and a New Algorithm, Proc. AAAI 129-134, 1992b.

Koh K., Kim S.J., and Boyd S., An interior-point method for large-scale ℓ_1-regularized logistic regression, J. Mach. Learn. Res., 8, pp. 1519-1555, 2007.

Kohavi R., John G.H., Wrapper for feature subset selection, Artificial Intelligence, 97, pp. 273-324, 1997.

Kohonen T., Self-Organaized formation of topologically feature maps, Biol Cybernetics, 43, pp. 59-69, 1982.

Konecny G.E., Meng Y.G., Untch M., *et al.*, Association between her-2/neu and vascular endothelial growth factor expression predicts clinical outcome in primary breast cancer patients. Clin Cancer Res, 10(5), pp. 1706-1716, 2004.

Kononenko I., Estimating Attributes: Analysis and Extentions of Relief, Proc. European Conf. Mach. Learning ECML, pp. 171-182, 1994.

Koscielny S., Critical review of microarray-based prognostic tests and trials in breast cancer, Current Opinion in Obstetrics and Gynecology, 20, pp. 47-50, 2008.

Kuipers B., *Qualitative Reasoning: Modeling and simulation with Incomplete Knowledge*, The MIT Press, Cambridge, Massachusetts, London, 1994.

Lafferty J., Wasserman L., Challenges in statistical machine learning, Stat. Sinica, 16, pp. 307-322, 2006.

Land W.H., Verheggen E.A., Multiclass primal support vector machines for breast density classification, Int J Comput Biol Drug Des., 2(1), 21-57, 2009.

Lee C.C., Fuzzy logic in control systems: Fuzzy logic controller-Part I and Part II, IEEE Tans. Syst., Man, Cybern., 20 (2), pp. 404-435, 1990.

Lee J.K., Havaleshko D.M., Cho H-J., *et al.*, A strategy for predicting the chemosensitivity of human cancers and its application to drug discovery, Proceedings of the National Academy of Sciences of the United States of America, 104 (32), pp. 13086–13091, 2007.

Lee M-L.T., Kuo F.C., Whitemore G.A., Sklar J., Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations, Proc. Natl. Acad. Sci. USA 97, pp. 9834-9839, 2000.

Lee M-Y., Yang Chi-Shih, Entropy-based feature extraction and decision tree induction for breast cancer diagnosis with standardized thermograph images, Computer methods and programs in biomedicine, 100 (100), pp. 269-282, 2010a.

Lee S.I., Lee H., Abbeel P., and Ng A.Y., Efficient ℓ_1 regularized logistic regression, Proc. 21st AAAI Conf. Artif. Intell., Boston, MA, USA, pp. 1-9, 2006.

Lee Y., Support vector machines for classification: a statistical portrait, Methods Mol Biol., 620, pp. 347-368, 2010b.

Li J., Zhang Z., Rosenzweig J., Wang Y.Y., Chan D.W., Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer, Clinical chemistry 48 (8), pp. 1296-1304, 2002.

Li T., Zhang C., Ogihara M., A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression, Bioinformatics, 20 (15), pp. 2429-2437, 2004.

Li Y., Wu Z-F, Fuzzy feature selection based on min-max learning rule and extension matrix, Pattern Recognition , 41, pp. 217-226, 2008.

Li Y., Lu B-L, Feature selection based on loss-margin of nearest neighbor classification, Pattern Recognition, 42, 1914-1921, 2009.

Liu H., Hussian F., Tan C.L., Dash M., Discretization: an enabling technique, J. Data Mining and Knowledge Discovery, 6(4), pp. 393-423, 2002.

Liu H. , Yu L., Toward Integrating Feature Selection Algorithms for Classification and Clustering, IEEE Transactions on Knowledge and Data Engineering, 17 (4), pp. 491-502, 2005.

Liu H.X., Zhang R.S., Luan F., *et al.*, Diagnosing Breast Cancer Based on Support Vector Machines, J. Chem. Inf. Comput. Sci., 43 (3), pp 900–907, 2003.

Lucas P.J.F., Model-based diagnosis in medicine. Artif. Intell. Med., 10, pp. 201-208, 1997.

Ma X., Wang J.Z., Ryan P.D., *et al.*, A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen, Cancer Cell, 5, pp. 607–616, 2004.

Maclin P.S., Dempsey J., Brooks J., and Rand J., Using neural networks to diagnose cancer , Journal of Medical Systems,15 (1), pp. 11-19, 1991

MacQueen J., Some methods for classification and analysis of multivariate observations, Proceedings of the 5th Berkely Symposium on Mathematical Statistics and Probability, 1, University of California, Berkeley, USA, pp. 281-297, 1967.

Makris A., Powles T.J. , Ashley S.E. , *et al.*, A reduction in the requirements for mastectomy in a randomized trial of neoadjuvant chemoendocrine therapy in primary breast cancer. Ann Oncol, 9 (11), pp. 1179-1184, 1998.

Mangasarian O.L., Exact 1-Norm support vector machines via unconstrained convex differentiable minimization, J. Mach. Learn. Res., 7, pp. 1517-1530, 2006.

Matsumura Y., Tarin D., Significance of CD44 gene products for cancer diagnosis and disease evaluation, The Lancet, 340 (8827), pp. 1053-1058, 1992.

Mauri D., Pavlidis N., Ioannidis J.P.A., Neoadjuvant Versus Adjuvant Systemic Treatment in Breast Cancer: A Meta-Analysis, *J.* Natl. Cancer Inst., 97(3), pp. 188–194, 2005.

Medasani S., Kim J., An overview of membership function generation techniques for pattern recognition, Int'l J. of Approximate Reasoning, 19, 391–417, 1998.

Mian S., Ugurel S., Parkinson E., Serum proteomic fingerprinting between clinical stages and predict disease progression in melanoma patients, J Clin Oncol, 23, pp. 5088-5093, 2005.

Michalski R.S. and Stepp R.E., Automated construction of classifications: Conceptual clustering versus numerical taxonomy, IEEE Trans. Pattern Anal. Machine Intell., 5(4), , pp. 396-410, 1980.

Michiels S., Koscielny D., Hill C., Prediction of cancer outcome with microarrays: a multiple random validation strategy, The Lancet, 365 (9458), pp. 488-492, 2005.

Mika S., Ratsch G., Weston J., Scholkopf B., Mullers K.R.**,** Fisher discriminant analysis with kernels, Workshop Neural Networks for Signal Processing IX, proceedings IEEE Signal Processing Society, 1999.

Miller L.D., Smeds J., George J., *et al.*, An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival, PNAS September 20, 102 (38), pp. 13550-13555, 2005.

Mitchell T.M., *Machine learning*, McGraw, 1997.

Mitra P., Murthy C.A., Pal S.K., Unsupervised Feature Selection Using Feature Similarity, IEEE Transactions on Pattern Analysis and Machine Intelligence, 24 (3), pp. 301-312, March, 2002.

Mohri T., Hidehiko T., An optimal Weighting Criterion of case indexing for both numeric and symbolic attributes, in *D.W. Aha (Ed.)*, Case-based Reasoning workshop. Menlo Park, CA:AIII Press, pp. 123-127, 1994.

Ng A.Y., Feature selection, L1 vs. L2 regularization, and rotational invariance, Proc. 21st Intl. Conf. Mach. Learn. , Canada, pp. 78-86, 2004.

Nocedal J., and Wright S.J., *Numerical optimization*, Springer Verlag, New York, NY, 1999.

Novak J.P., Sladek R., Hudson T.J., Characterization of variability in large-scale gene expression data: implications for study design, Genomics, 79, pp. 104-113, 2002.

Nykter M., Aho T., Ahdesmäki M., Ruusuvuori P., *et al.*, Simulation of microarray data with realistic characteristics, BMC Bioinformatics, 7(1), pp. 332-349, 2006.

Olivotto I.A., Bajdik C.D., Ravdin P.M., *et al.*, Population-based validation of the prognostic model adjuvant! for early breast cancer, J of Clin Onco, 23(12), pp. 2716-2725, 2005.

Olshen A.B., Jain A.N., Deriving quantitative conclusions from microarray expression data, Bioinformatics,18 (7), pp. 961-970, 2002.

Paik S., Shak S., Tang G., *et al.*, A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. New England Journal of Medicine, 351(27), pp. 2817-2826, 2004.

Parry R.M., Jones W., Stokes T.H., *et al.*, *k*-Nearest neighbor models for microarray gene expression analysis and clinical outcome prediction, Bioinformatics, 18(7), pp. 961-970, 2002.

Perez E.A., Suman V.J., Davidson N.E., *et al.*, HER2 Testing by Local, Central, and Reference Laboratories in Specimens From the North Central Cancer Treatment Group N9831 Intergroup Adjuvant Trial. J Clin Oncol, 24(19), pp. 3032-3038, 2006.

Perou C.M., Sorlie T., Eisen M.B., *et al.*, Molecular portraits of human breast tumours, Nature 52, pp. 406-747, 2000.

Piera N. and Aguilar-Martin J., Controlling selectivity in non-standard pattern recognition algorithms, IEEE Trans. on Systems, Man and Cybernetics, 21 (1), 1991.

Quinlan J.R., Induction of decision trees, Machine Learning, 1, pp. 81-106, 1986.

Raemaekers T., Ribbeck K., Beaudouin J., *et al.*, NuSAP, a novel microtubule-associated protein involved in mitotic spindle organization, J Cell Biol, 162 (6), pp. 1017-1029, 2003

Ramaswamy S., Tamayo P., Rifkin R., *et al.*, MultiClass Cancer Diagnosis Using Tumor Gene Expression Signatures, Proc. Nat'l Acad. Sc. USA, 98 (26), pp. 15149-15154, 2001.

Rasmani K.A., Shen Q., Modifying Weighted Fuzzy Subsethood-based Rule Models with Fuzzy Quantifiers, In Proceedings of the IEEE international conference on fuzzy systems, pp. 1679-1684, 2004.

Reid J.F. , Lusa L., De Cecco L., *et al.*, Limits of Predictive Models Using Microarray Data for Breast Cancer Clinical Treatment Outcome, JNCI J Natl Cancer Inst , 97 (12), pp. 927-930, 2005.

Reis-Filho J.S., Westbury C., Pierga J-Y., The impact of expression profiling on prognostic and predictive testing in breast cancer, J. Clin. Pathol., 59, pp. 225-231, 2006.

Rennie J.D., and Srebro N., Fast maximum margin matrix factorization for collaborative prediction, Proc. 22nd Intl. Conf. Mach. Learn., Bonn, Germany, pp. 713-719, 2005.

Ressom H., Reynolds R., Varghese R.S., Increasing the efficiency of fuzzy logic-based gene expression data analysis, Physiol Genomics, 13, pp. 107-117, 2003.

Ripley R.M., Harris A.L., Tarassenko L., Non-linear survival analysis using neural networks, Stat. Med., 23, pp. 825-842, 2004.

Robnik-Šikonja M., and Kononenko I., Theoretical and empirical analysis of ReliefF and RReliefF, Machine Learning, 53, pp. 23-69, 2003.

Rodvold D.M., McLeod D.G., Brandt J.M., Snow P.B., Murphy G.P., introduction to artificial neural networks for physicians: taking the lid of the black box, Prostate, 46, pp. 39-44, 2001.

Rosenblatt F., Principle of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms, Spartan Books, 7 (3), pp. 218-219, 1962.

Rosset S., Zhu J., and Hastie T., Boosting as a regularized path to a maximum margin classifier, J. Mach. Learn. Res., 5, pp. 941-973, 2004.

Rouzier R., Perou C.M., Symmans W.F., *et al.*, Breast Cancer Molecular Subtypes Respond Differently to Preoperative Chemotherapy, Clin Cancer Res, 11, pp. 5678-5685, 2005.

Rumelhart D.E., Hinton G.E., Williams R.J., Learning representations by back-propagation errors, Nature, 323, pp. 533-536, 1986.

Saeys Y., Inza I., Larrañaga P., A review of feature selection techniques in bioinformatics, Bioinformatics, 23(19), pp. 2507-2517, 2007.

Sanz J.A., Fernández A., Bustince H., Herrera F., Improving the performance of fuzzy rule-based classification systems with interval-valued fuzzy sets and genetic amplitude tuning, Inform. Sci., 180, pp. 3674-3685, 2010.

Schmidt M., Fung G., and Rosales R., Fast optimization methods for L1 regularization:a comparative study and two new approaches, Proc. 18th Euro. Conf. Mach. Learn., pp. 286-297, 2007.

Schnieders F., Dörk T., Arnemann J., *et al.*, Testis-specific protein, Y-encoded (TSPY) expression in testicular tissues, Hum. Mol. Gent., 5, 1801-1807, 1996.

Sheng Q., Moreau Y., and De Moor B., Biclustering microarray data by Gibbs sampling. Bioinformatics, 19(2), pp. 196–205, 2003.

Shipp M., Ross K., Tamayo P., *et al.*, Diffuse Large b-Cell Lymphoma Outcome Prediction by Gene-Expression Profiling and Supervised Machine Learning, Nat. Med., 8 (1), pp. 68-74, 2002.

Simes R.J., Treatment selection for cancer patients: application of statistical decision theory to the treatment of advanced ovarian cancer, J Chronic Dis, 38, pp. 171-186, 1985.

Singh D., Febbo P., Ross K., *et al.*, Gene Expression Correlates of Clinical Prostate Cancer Behavior, Cancer Cell, 1 (2), pp. 203-209, 2002.

Singletary S.E., Allred C., Ashley P., *et al.*, Revision of the American joint Committee on cancer staging system for breast cancer, J Clin Oncol, 20, pp. 3628-3636, 2002.

Smith R.E, A review of selective estrogen receptor modulators and national surgical adjuvant breast and bowel project clinical trails, Semin Oncol, 30, pp. 4-13, 2003.

Sotiriou C., Neo S-Y., McShane L.M. , *et al.* , Breast cancer classification and prognosis based on gene expression profiles from a population-based study, PNAS, 100 (18), pp. 10393-10398, 2003.

Stephenson A.J., Smith A., Kattan M.W., *et al.*, Integration of gene expression profiling and clinical variables to predict prostate carcinoma recurrence after radical prostatectomy, Cancer, 104, pp. 290-298, 2005.

Straver M.E., Glas A.M., Hannemann J., *et al.*, The 70-gene signature as a response predictor for neoadjuvant chemotherapy in breast cancer, Breast Cancer Res. Treat., 119 (3), pp. 551-558, 2009.

Sugeno M., An introductory survey of fuzzy control, Inf. Sci., 36 (1/2), pp. 59-83, 1997.

Sun Y., Todorovic S., Li J., Wu D., Unifying Error-Correcting and Output-Code AdaBoost through the Margin Concepts, In proceedings of 22nd Int'l Conf. on Machine Learning, pp. 872-879, 2005.

Sun Y., Goodison S., Li J., Liu L., Farmerie W., Improved breast cancer prognosis through the combination of clinical and genetic markers. Bioinformatics, Gene expression, 23 (1), 30-37, 2007a.

Sun Y., Iterative RELIEF for Feature Weighting: Algorithms, Theories, and Applications, IEEE TPAMI, 2 (6), pp. 1035-1051, 2007b.

Swets J.A., Signal detection theory and ROC analysis in psychology and diagnostics: collected papers, Scientific psychology series, Hillsdale, NJ, England: Lawrence Erlbaum Associates, 1996.

Tamayo P., Slonim D., Mesirov J., *et al.*, Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation, Proc Natl Acad Sci USA, 96(6), pp. 2907-2912, 1999.

Tibshirani R., Regression shrinkage and selection via the LASSO, J. R. Stat. Soc. Ser. B, 58, pp. 267-288, 1996.

Tu Y., Stolovitzky G., Klein U., Quantitative noise analysis for gene expression microarray experiments, pnas, 99 (22), pp. 14031-14036, 2002.

Van de Vijver M.J., He Y.D., Van't Veer L.J., *et al.*, A Gene expression signature as a predictor of survival in breast cancer, N Engl J Med, 347 (25), pp. 1999-2009, 2002.

Van't Veer L.J., Dai H., van de Vijver M.J., Gene expression profiling predicts clinical outcome of breast cancer, Nature 415, pp. 530-536, 2002.

Vapnik V.N., *Statistical Learning Theory*, John Wiley & Sons, 1998.

Vellido A., Lisboa P.J.G., Neural and Other Machine Learning Methods in Cancer Research, In F. Sandoval et al. (Eds.), LNCS 4507, pp. 964-971, 2007.

Vogel C.L., Cobleigh M.A., Tripathy D., *et al.*, Efficacy and safety of trastuzumab as a single agent in first-line treatment of HER2-overexpressing metastatic breast cancer, J of Cl Onco, 20 (3) , pp. 719-726, 2002.

Wang J., Bø T.H., Jonassen I., *et al.*, Tumor classification and marker gene prediction by feature selection and fuzzy c-means clustering using microarray data, BMC bioinformatics, 4 (60), 2003.

Wang W. and Zhang Y., On fuzzy cluster validity indices, Fuzz. Set and sys., 158 (19), pp. 2095-2117, 2007

Wang X., Wang Y., Wang L., Improving fuzzy c-means clustering based on feature-weight learning, Pattern Recognition Letters, 25, pp. 1123-1132, 2004.

Wang Y., Klijn J.G., Zhang Y., *et al.*, Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet, 365, pp. 671-679, 2005a.

Wang Y., Tetko I.V., Hall M.A., *et al.*, Gene selection from microarray data for cancer classification—a machine learning approach, Computational Biology and Chemistry, 29 (1), pp. 37-46, 2005b.

Wei H-L., Billings S.A., Feature Subset Selection and Ranking for Data Dimensionality Reduction, IEEE Transactions on Pattern Analysis and Machine Intelligenc, 29 (1), pp. 162-166, 2007.

Weston J., Mukherjee S., Chapelle O., Pontil M., Vapnik V., Feature Selection for SVMs, Advances in Neural Information Processing Systems, pp. 668-674, 2001.
Wettschereck D., Aha D.W., Weighting features, Case-Based Reasoning, Research and Development Lecture Notes in Computer Science, 1010, pp. 347-358, 1995.

Wettschereck D., Aha D.W., and Mohri T., A Review and Empirical Evaluation of Feature Weighting Methods for a Class of Lazy Learning Algorithms, Artificial Intelligence Review, 11 (1-5), pp. 273-314, 1997.

Wiseman S.M., Makretsov N., Nielsen T.O., *et al.*, Coexpression of the Type 1 Growth Factor Receptor Family Members HER-1, HER-2, and HER-3 has a Synergistic Negative Prognostic Effect on Breast, Cancer, 23 (9), pp. 1770-1777, 2005.

Wolberg W.H., Street W.N., and Mangasarian O.L.. Machine learning techniques to diagnose breast cancer from fine-needle aspirates. Cancer Letters, 77, pp. 163-171, 1994.

Wong S.L., Chao C., Edwards M.J., *et al.*, Frequency of sentinel lymph node metastases in patients with favorable breast cancer histologic subtypes, Am J Surg, 184, pp. 492-498, 2002.

Wygralak M., An axiomatic approach to scalar cardinalities of fuzzy sets, Fuzzy Sets and Systems, 110, pp. 175-179, 2000.

Xing E., Jordan M., Karp R., Feature selection for high dimensional genomic microarray data, Proc. 15th Int'l Conf. Machine Learning, pp. 601-608, 2001.

Xu R., Wunch D.I.I., Survey of clustering algorithms, IEEE Trans. Neural Networks, 16(3), pp. 645-678, 2005.

Yang R-B. , Ng C.K.D., Wasserman S.M., *et al.*, Identification of a novel family of cell-surface proteins expressed in human vascular endothelium, J. Biol. Chem., 227, pp. 46364-46373, 2002.

Yao J., Dash M., Tan S.T., Liu H., Entropy-based fuzzy clustering and fuzzy modeling, Fuzz. Set and Sys., 113, pp. 381-388, 2000.

Zadeh L.A., Fuzzy sets and systems theory, In: Fox J editor, Polytechnic press, pp. 29-37, 1965.

Zadeh L.A., Biological applications of the theory of fuzzy sets and systems. In proced. Of an international symposium on Biocybernetics of the Central Nervous System, pp. 199-206, 1969.

Zhang L., Zhou W., Velculescu V.E. , *et al.*, Gene Expression Profiles in Normal and Cancer Cells, Science, 276 (5316), pp. 1268-1272, 1997.

Zheng B., Wang X., Lederman D., Tan J., Gur D., Computer-aided detection; the effect of training databases on detection of subtle breast masses, Pharmacogenomics J., 10(4), pp. 292-309, 2010.

Zhou X., Tan M., Stone Hawthorne V., *et al.*, Activation of the Akt/mammalian target of rapamycin/4E-BP1 pathway by ErbB2 overexpression predicts tumor progression in breast cancers, Clin Cancer Res, pp. 6779-6788, 2004.

Zhu J., Rosset S., Hastie T., and Tibshirani R., 1-norm support vector machines, Proc. 16th Adv. Neu. Info. Proc. Sys., pp. 49-56, 2003.

Zindy P., Bergé Y., Allal B.C., *et al.*, Formation of the eIF4F translation initiation complex determines sensitivity to anti-cancer drugs targeting the EGFR and HER2 receptors, Cancer Research, doi: 10.1158/0008-5472.CAN-11-0420, 2011.

Zwick R., Carlstein E., Budescu D.V., Measures of similarity among fuzzy concepts: A comparative analysis, Int'l J. Approx. Reason., 1 (2), pp. 221-242, 1987.