Classification multi-labels graduée : découverte des relations entre les labels, et adaptation à la reconnaissance des odeurs et au contexte big data des systèmes de recommandation

Khalil Laghmari

Sorbonne Université

Université Hassan II de Casablanca

École Doctorale Informatique, Télécommunication et Électronique

Laboratoire Informatique de Paris 6

Classification multi-labels graduée : découverte des relations entre les labels, et adaptation à la reconnaissance des odeurs et au contexte big data des systèmes de recommandation

Par Khalil Laghmari

Thèse de doctorat d'informatique

Dirigée par Christophe Marsala et Mohammed Ramdani

Présentée et soutenue publiquement le 23/03/2018

Devant un jury composé de :

Abdelaziz BERRADO, Professeur, Rapporteur

Abdelkrim BEKKHOUCHA, Professeur, Examinateur

Anne LAURENT, Professeur, Rapporteur

Bernadette BOUCHON-MEUNIER, Directrice de recherche CNRS émérite, Président

Christophe MARSALA, Professeur, Co-directeur de thèse

Mohammed RAMDANI, Professeur, Co-directeur de thèse

Dédicace

A ma femme Yasmina

A mes parents Abderrahmane et Fatima

A ma sœur Samia, et à mes deux frères Amine et Yassine

Aux autres membres de la famille, aux amis et à tous ceux qui me sont chers et proches

Je dédie ce travail

Remerciements

Il me sera très difficile de remercier tout le monde car c'est grâce à l'aide de nombreuses personnes que j'ai pu mener cette thèse à son terme.

Je voudrais tout d'abord remercier grandement mes deux directeurs de thèse, Christophe Marsala et Mohammed Ramdani, pour toute leur aide. Je suis ravi d'avoir travaillé en leur compagnie car outre l'appui scientifique, ils ont toujours été présents pour me soutenir et me conseiller au cours de l'élaboration de cette thèse.

Je remercie également Anne Laurent et Abdelaziz Berrado qui m'ont fait l'honneur d'être rapporteurs de ma thèse. Leurs remarques m'ont permis d'améliorer mon travail et m'ont apporté différentes perspectives. Pour tout cela je les remercie.

Je tiens à remercier Bernadette Bouchon-Meunier et Abdelkrim Bekkhoucha pour avoir accepté de participer à mon jury de thèse, et pour avoir consacré du temps à ce travail de recherche.

Mes derniers remerciements vont au Ministère de l'Enseignement Supérieur De la Recherche Scientifique et de la Formation des Cadres, et à l'Institut Français du Maroc pour avoir financé conjointement ce travail de thèse.

Table des matières

Introduction **1**

1 Classification multi-labels graduée **9**

1.1 Classification mono-label 10

1.1.1 Description formelle de la classification mono-label 13

1.1.2 Classifieurs mono-label 13

1.1.3 Transformation d'un problème multi-classes en un problème binaire 14

1.1.4 Conclusion 16

1.2 Classification multi-labels 17

1.2.1 Description formelle de la classification multi-labels 21

1.2.2 Mesures d'évaluation de la classification multi-labels 21

1.2.3 Approches d'adaptation au cas multi-labels 24

1.2.4 Approches de transformation au cas mono-label 25

1.2.5 Conclusion 33

1.3 Classification multi-labels graduée 34

1.3.1 Description formelle du problème de la classification multi-labels graduée . . 37

1.3.2 Interaction avec les autres types de classification 37

1.3.3 Décomposition du problème de classification multi-labels graduée 38

1.3.4 Conclusion 48

1.4 Synthèse 48

2 Nouvelles approches de classification multi-labels graduée **51**

2.1 L'approche PSI 54

2.1.1 Algorithme PSI 54

2.1.2 Exemple d'exécution de l'approche PSI 55

2.2 L'approche PSI2 59
2.2.1 Description de l'approche PSI2 59
2.2.2 Exemple d'exécution de l'approche PSI2 63
2.3 L'approche CLR_PSI 67
2.3.1 Description de l'approche CLR_PSI 67
2.3.2 Exemple d'exécution de l'approche CLR_PSI 67
2.4 L'approche Stacked_RPC_PSI 70
2.4.1 Description de l'approche Stacked_RPC_PSI 70
2.4.2 Exemple d'exécution de l'approche Stacked_RPC_PSI 70
2.5 Expérimentation 73
2.5.1 Sélection des paramètres des approches PSI et PSI2 73
2.5.2 Évaluation des approches PSI, PSI2, CLR_PSI, et Stacked_RPC_PSI sur des données multi-labels 74
2.5.3 Évaluation des approches PSI, CLR_PSI, et Stacked_RPC_PSI sur des données multi-labels graduées 79
2.6 Conslusion 82

3 Systèmes de recommandation et classification multi-labels graduée 85
3.1 Introduction 86
3.2 Systèmes de recommandation 88
3.2.1 Description formelle de la tâche de recommandation 88
3.2.2 Approches d'élaboration des systèmes de recommandation 88
3.3 Systèmes de recommandations basés sur la classification 89
3.3.1 Modes d'apprentissage des classifieurs 89
3.3.2 Arbres de décision incrémentales 90
3.3.3 Gestion des changements de concept 91
3.4 Une nouvelle approche pour les systèmes de recommandation 91
3.4.1 Représentation des données 92
3.4.2 Optimisation de l'accès au disque 93
3.4.3 Paramètres de l'apprentissage incrémental 94
3.4.4 Spécification du classifieur multi-labels gradué 94
3.5 Expérimentation 95
3.5.1 Données utilisées 95

3.5.2 Méthodes d'évaluation de la classification incrémentale 95
3.5.3 Paramètres de l'approche proposée 96
3.5.4 Résultats et discussion 97
3.6 Conclusion 99

4 Reconnaissance des odeurs et classification multi-labels graduée 101
4.1 Introduction 101
4.2 Comparaison des approches de classification MLG sur les données des molécules odorantes 104
4.2.1 Préparation des données 104
4.2.2 Expérimentations préliminaires 105
4.3 Adaptation au jeu de données des molécules odorantes 107
4.4 Expérimentation et résultats 108
4.5 Conclusion 113

Conclusion et perspectives 115

Bibliographie 130

Liste des figures

1 Classification multi-labels graduée **9**

1.1 Types de classification 10

1.2 Arbre de décision de l'odeur 'JASMIN' à partir des données du Tableau 1.1 14

1.3 Arbre de décision pour prédire l'absence ou la présence de l'odeur 'HERBAC' à partir des données du Tableau 1.8 19

1.4 Arbre de décision pour prédire l'absence ou la présence de l'odeur 'FRUITY' à partir des données du Tableau 1.8 19

1.5 Arbre de décision pour prédire l'absence ou la présence de l'odeur 'JASMIN' à partir des données du Tableau 1.8 20

1.6 Structure de dépendance entre les labels 'HERBAC', 'FRUITY', et 'JASMIN' à partir des arbres de décision illustrés dans les figures 1.3, 1.4, et 1.5 20

1.7 Approches de transformations au cas mono-label 34

1.8 Arbre de décision pour prédire le degré d'association de l'odeur 'HERBAC' à partir des données du Tableau 1.16 35

1.9 Arbre de décision pour prédire le degré d'association de l'odeur 'FRUITY' à partir des données du Tableau 1.16 36

1.10 Arbre de décision pour prédire le degré d'association de l'odeur 'JASMIN' à partir des données du Tableau 1.16 36

2 Nouvelles approches de classification multi-labels graduée **51**

2.1 Verrous levés par les nouvelles approches proposées (PSI, PSI2, CLR_PSI, Stacked_RPC_PSI 53

2.2 Compatibilité des approches proposées avec les stratégies de décomposition de la classification multi-labels graduée 53

2.3 Arbres de décision dans $\mathscr{H}$ 57

2.4 Structure de dépendance initiale des classifieurs de $\mathscr{H}$ 57
2.5 Le nouveau classifieur $\mathbb{H}_2$ 60
2.6 Structures de dépendances après le retrait de H_2 60
2.7 Structures de dépendances après l'ajout de H_4 dans $\mathbb{H}$ 60
2.8 Le nouveau classifieur $\mathbb{H}_1$ 61
2.9 Structures de dépendances après le retrait de H_1 61
2.10 Structures de dépendances après l'ajout de H_3 dans $\mathbb{H}$ 61
2.11 L'approche PSI2: arbres de décision à l'étape initiale 64
2.12 L'approche PSI2: structure de dépendance des classifieurs initiaux 64
2.13 L'approche PSI2: structure de dépendance finale 64
2.14 L'ensemble d'arbres de décision construits par l'approche CLR_PSI selon l'approche CLR à partir des données d'apprentissage du Tableau 2.4 68
2.15 L'ensemble d'arbres de décision construits par l'approche CLR_PSI selon l'approche PSI à partir des données d'apprentissage du Tableau 2.4 69
2.16 Structure de dépendances correspondant aux classifieurs construits par l'approche CLR_PSI à partir des données d'apprentissage du Tableau 2.4 69
2.17 L'ensemble d'arbres de décision construits par l'approche Stacked_RPC_PSI selon l'approche PSI à partir des données d'apprentissage du Tableau 2.6 72
2.18 Structure de dépendances correspondant aux classifieurs construits par l'approche Stacked_RPC_PSI à partir des données d'apprentissage du Tableau 2.6 72

3 Systèmes de recommandation et classification multi-labels graduée 85

3.1 Exemple d'une matrice d'évaluation 86
3.2 Evaluation de l'approche proposée sur les données 'MovieLens' 97
3.3 Evaluation de l'approche proposée sur les données 'Jester' 99

4 Reconnaissance des odeurs et classification multi-labels graduée 101

4.1 Nombre de molécules pour chaque cardinalité de labels associés 105
4.2 Nombre de molécules pour chaque odeur 105
4.3 Nombre de molécules pour chaque ensemble d'odeurs 109
4.4 Performance en prédiction pour les qualités olfactives 110
4.5 L'arbre de décision de la 'VANILL' obtenu par l'approche BR 110
4.6 L'arbre de décision de la 'VANILL' obtenu par l'approche *PSI** 111

4.7 Performance en prédiction pour les ensembles de qualités olfactives générées 111
4.8 L'arbre de décision de la 'ROSE' obtenu par l'approche PSI à la 10ème étape de la validation croisée 112
4.9 Graphe de dépendances obtenu par l'approche PSI à la 10ème étape de la validation croisée 113

Liste des tableaux

1 Classification multi-labels graduée **9**

1.1 Données d'apprentissage pour la prédiction de l'absence ou la présence de l'odeur 'JASMIN' 11

1.2 Données d'apprentissage pour la prédiction d'une odeur dans l'ensemble {JASMIN, MINTY, MUSK} 12

1.3 Données d'apprentissage pour la prédiction du degré d'association de l'odeur 'JASMIN' 12

1.4 Instances multi-classes 16

1.5 Transformation-OVA des données du Tableau 1.4 16

1.6 Transformation-OVO des données du Tableau 1.4 17

1.7 Transformation ordinale $\rightarrow$ binaire des données du Tableau 1.4 en considérant $c_1 < c_2 < c_3$ 17

1.8 Données d'apprentissage pour la prédiction de sous-ensembles d'odeurs 18

1.9 Exemple de données multi-labels 25

1.10 Transformation par la méthode LP 26

1.11 Transformation par la méthode PPT avec un seuil égal à deux 27

1.12 Transformation par la méthode PPT-n avec un seuil égal à deux 27

1.13 Transformation par la méthode BR 28

1.14 Transformation par la méthode CC 30

1.15 Transformation par la méthode RPC 32

1.16 Données d'apprentissage pour la prédiction de sous-ensembles gradués d'odeurs . . 35

1.17 Exemple de données multi-labels graduées 38

1.18 Décomposition verticale des données multi-labels graduées du Tableau 1.17 39

1.19 Décomposition horizontale des données multi-labels graduées du Tableau 1.17 40

1.20 Décomposition complète des données multi-labels graduées du Tableau 1.17 42
1.21 Données multi-labels graduées avec les labels virtuels 43
1.22 Décomposition par l'approche Horizontal_CLR pour le degré d'association m_2 . . . 45
1.23 Vue sur la décomposition par l'approche Full-CLR 47

2 Nouvelles approches de classification multi-labels graduée 51

2.1 Ensemble de données multi-labels 55
2.2 Impact des mesures PSI sur la structure de dépendance obtenue 62
2.3 Structures de dépendances pour différents paramètres de l'approche PSI2 64
2.4 Exemple de données d'apprentissage pour l'approche CLR_PSI 68
2.5 Exemple de données d'apprentissage pour l'approche Stacked_RPC_PSI 71
2.6 Données d'apprentissage mises à jour pour l'approche Stacked_RPC_PSI après l'introduction des prédictions des classifieurs H_{c_1,c_2}, H_{c_1,c_3}, et H_{c_2,c_3} 71
2.7 Données multi-labels réelles 73
2.8 Évaluation de la prédiction pour les données 'emotions', 'scenes', et 'yeast' 77
2.9 Evaluation de la complexité du modèle appris et des dépendances entre les labels pour les données 'emotions', 'scenes', et 'yeast' 78
2.10 Evaluation de la propagation de l'erreur de prédiction pour les données 'emotions', 'scenes' et 'yeast' 78
2.11 Description des données multi-labels graduées 'BelaE' 80
2.12 Evaluation de la prédiction pour les données 'BelaE' pour chaque degré d'association 81

3 Systèmes de recommandation et classification multi-labels graduée 85

3.1 Description des données 95
3.2 Évaluation détaillée de l'approche proposée sur les données 'MovieLens' 98
3.3 Evaluation détaillée de l'approche proposée sur les données 'Jester' 98
3.4 Comparaison de l'approche proposée avec des approches existantes sur les données 'MovieLens' 99

4 Reconnaissance des odeurs et classification multi-labels graduée 101

4.1 Tendances dans le domaine de la reconnaissance de l'odeur 103
4.2 Préparation des données 104

4.3 Description du jeu de données des molécules odorantes 104

4.4 Evaluation de la prédiction des odeurs pour chaque sous-problème de classification multi-labels 106

4.5 Evaluation de la prédiction en moyennant les résultats obtenus dans les sous-problèmes de classification multi-labels 107

Introduction

Les données provenant de différentes sources sont continuellement collectées et entreposées grâce au développement continu des outils de stockage. La préoccupation de plusieurs chercheurs et industriels est l'extraction des informations intéressantes et potentiellement utiles à partir de données entreposées. Plusieurs méthodes scientifiques et outils informatiques permettent l'analyse et l'extraction de connaissances à partir des données. Le choix d'une méthode d'exploration de données (data-mining) dépend de la nature des données et de l'objectif souhaité.

Dans cette thèse, nous nous sommes intéressés au cas où les données sont classées selon des catégories, dites aussi classes ou labels, et décrites par des attributs, dits attributs descriptifs ou conditionnels. L'objectif est d'extraire des liens reliant les valeurs prises par les attributs descriptifs aux catégories qui sont les valeurs prises par l'attribut de décision. L'intérêt d'extraire ces liens est de pouvoir prédire la catégorie de nouvelles données non encore classées.

Lorsque chaque donnée ne peut être associée qu'à un seul label, la tâche de prédiction du label associé est dite classification mono-label. Par exemple, lorsqu'un document ne peut être associé qu'à un seul thème ('économie', 'politique', etc), la prédiction du thème associé est une tâche de classification mono-label. Elle est dite classification multi-labels lorsque chaque donnée peut être affectée à plusieurs labels simultanément (Herrera et al.,2016,[48]). Par exemple, le cas où un document peut être associé à la fois au thème 'économie' et au thème 'politique'. La classification est dite multi-labels graduée (CMLG) lorsque l'association à chaque label se fait avec un degré d'association appartenant à une échelle graduelle de degrés d'appartenance (Cheng et al.,2010,[25];Brinker et al.,2014,[20]). Par exemple, un document peut être associé fortement au label 'économie', faiblement au label 'politique', et très faiblement au label 'société'. Ce document est donc associé à l'ensemble gradué de labels {'économie' | fortement, 'politique' | faiblement, 'société' | très faiblement}. Les données associées à des ensembles gradués de labels (les documents dans l'exemple précédent) sont dites multi-labels graduées. La classification est dite floue (Zadeh,1965,[125]) lorsque les degrés d'association sont des valeurs numériques dans l'intervalle $[0, 1]$ dont la somme est égale à 1 pour chaque donnée. Par exemple, un document peut être associé à l'ensemble flou de thèmes {'économie' | 0.7, 'politique' | 0.2, 'société' | 0.1}. L'information sur la distance entre les degrés d'association est donc fournie pour les données floues. Par contre, pour les données multi-labels graduées, nous disposons uniquement d'un ordre total de degrés d'association qui ne sont pas nécessairement numériques (par exemple: très faiblement < faiblement < fortement).

L'introduction de la classification multi-labels graduée (Cheng et al.,2010,[25]) a permis de formaliser des approches de base pour tenir en compte cet aspect de gradualité des degrés d'association. Dans cette thèse, nous nous intéressons à la classification multi-labels graduée dans l'objectif de proposer des pistes d'amélioration des approches existantes par rapport à la précision des prédictions, à la complexité, et à l'interprétabilité de l'approche de classification.

Dans ce qui suit, nous introduisons des exemples réels de données multi-labels graduées correspondant à des domaines différents. Nous discutons ensuite les verrous à lever en classification multi-labels graduée et les limites des approches existantes. Nous synthétisons ensuite le périmètre et la contribution de cette thèse par rapport aux verrous à lever, puis nous détaillons le plan de la thèse. La liste des publications faites dans le cadre de cette thèse est donnée après le plan de la thèse.

Sources de données multi-labels graduées

Une source riche en données multi-labels graduées est le web. Plusieurs sites-web offrent aux internautes la possibilité d'exprimer leurs avis (degrés de satisfaction) par rapport à une variété de produits ou de services (labels). Ces sites-web diffèrent par rapport à l'échelle de degrés de satisfaction utilisée. Par exemple, l'échelle des degrés d'association peut être linguistique allant de 'très insatisfaisant' à 'très satisfaisant', ou symbolique allant d'une à cinq étoiles, ou numérique allant de la note 0 à la note 10. Ces données multi-labels graduées peuvent être exploitées pour l'apprentissage d'un système de recommandations (Aguilar et al.,2017,[3]) dont l'objectif est de proposer un contenu adapté qui a le plus de chance d'intéresser l'internaute. Ceci permet de garder l'internaute plus longtemps sur une même page pour augmenter les chances qu'il effectue une action d'achat.

La biologie et la chimie sont aussi des domaines riches en données multi-labels graduées. Un exemple intéressant est celui des molécules odorantes où chaque molécule est décrite par des variables physico-chimiques, et peut émettre plusieurs odeurs en même temps selon une échelle graduelle d'intensité allant de 'très faible' à 'très forte'. Les industries du parfum sont à l'écoute de nouveaux modèles de prédiction des odeurs à partir des propriétés physico-chimiques des molécules. Ces modèles peuvent guider les experts pour la synthèse de nouveaux parfums multi-odeurs avec les intensités souhaitées. Le secteur de la sécurité et de la santé est aussi intéressé par les modèles de classification des odeurs. En effet, ces modèles permettent d'élaborer des capteurs d'odeurs appelés 'nez électroniques'. Les nez électroniques sont utiles dans les cas où le nez humain n'est pas performant ou si son utilisation est dangereuse pour la santé. Par exemple, les nez électroniques peuvent être utilisés pour suivre la qualité de production du café ou du thé à partir de l'odeur (Chen et al.,2013,[24]), pour détecter les fuites de gaz (Mahlke et al.,2007,[74]), ou pour l'auto-dépistage des diabètes à partir de l'odeur d'urine (Seesaard et al.,2016,[98]).

Les documents multi-médias (textes, images, vidéos) constituent une grande base de données multi-labels graduée. Chaque document peut être affecté à plusieurs thèmes ordonnés selon le degré d'appartenance du document à chaque thème (Tai and Lin,2012,[103];Jindal and Taneja,2015,[54]). Par exemple, un document peut être associé aux thèmes 'économie', 'politique', et 'sport', mais être

plus focalisé sur le thème 'économie'. Les moteurs de recherche sur le web sont les plus intéressés par les approches de classification multi-labels graduée. Chaque nouveau document ajouté dans le web est injecté dans le modèle de classification afin de prédire les labels associés. Lorsqu'un internaute effectue une recherche, le résultat correspond aux documents ayant les plus grands degrés d'association aux thèmes de recherche.

Les données multi-labels graduées présentent souvent des relations entre les labels pouvant être divisées en deux types principaux :

- relations de dépendance permettant de prédire le degré d'association d'un label en se basant sur le degré d'association d'un autre label. Par exemple, une forte odeur de 'musc' est souvent perçue en présence d'une faible odeur parmi l'ensemble {'animale', 'boisée', 'ambrée'} ;

- relations de préférence exprimant une préférence entre deux labels en se basant sur les attributs descriptifs. Par exemple, un morceau de musique dont le signal numérique contient plusieurs piques a plus de chance d'être associé au label 'heureux' qu'au label 'relaxant'.

Défis de la classification multi-labels graduée :

* Relations entre les labels :

Les relations entre les labels sont une connaissance cachée dans les données qui intéresse les experts et qui peut guider la prédiction. En effet, la prédiction exacte de l'ensemble gradué de labels est une tâche difficile. Les approches de classification multi-labels graduée sont souvent évaluées par rapport à la similarité entre la prédiction et l'ensemble gradué de labels effectivement associés à la donnée. L'inconvénient des approches ignorant les relations entre les labels (Luaces et al.,2012,[73]) est qu'elles peuvent fournir des prédictions partiellement correctes mais incohérentes par rapport aux relations entre les labels. Par exemple, le fait de prédire pour une molécule l'ensemble gradué d'odeurs {'fruité' | faible, 'citron' | modérée, 'musc' | forte} semble incohérent par rapport à la relation liant l'odeur 'musc' aux odeurs 'animale', 'boisée', et 'ambrée'. Cette prédiction est donc probablement fausse ou juste partiellement correcte. En effet, soit l'odeur 'musc' n'est pas associée à la molécule, soit l'odeur 'musc' est associée à la molécule mais simultanément avec une autre odeur qui n'est pas prédite ('animale', 'boisée', ou 'ambrée'). L'avantage des approches tenant en compte les relations entre les labels (Read et al.,2014,[93];Soonsiripanichkul and Murata,2016,[101]) est qu'elles permettent l'obtention des prédictions cohérentes. Cependant l'inconvénient des prédictions basées sur les relations entre les labels est le risque de la propagation d'erreur de prédiction (une erreur de prédiction engendrée par une autre erreur de prédiction) (Senge et al.,2014,[99]). Par exemple, dans le cas où une molécule n'est pas associée à l'odeur 'musc', l'erreur de prédiction de l'odeur 'musc' risque d'engendrer l'erreur de prédiction d'une odeur faible 'animale', 'boisée', ou 'ambrée'.
D'un côté, l'apprentissage d'un maximum de relations entre les labels permet de maximiser la cohérence des prédictions. D'un autre côté, l'apprentissage d'un minimum de relations entre les labels

permet de minimiser le risque de la propagation des erreurs de prédiction. Ce compromis est l'un des verrous à lever en classification multi-labels graduée auquel nous nous intéressons dans cette thèse.

L'apprentissage des relations de dépendance entre les labels présente un défi supplémentaire dans le cas des dépendances cycliques. Par exemple, un film associé fortement au genre 'action' est souvent associé simultanément au genre 'suspense' et vice versa. Cette dépendance peut se traduire par les deux règles de classification suivantes:

- prédire le genre 'action' si l'association au genre 'suspense' est forte
- prédire le genre 'suspens' si l'association au genre 'action' est forte

Dans cet exemple, il ne serait pas possible de prédire ni le genre 'action' ni le genre 'suspense' parce que la dépendance entre les deux genres est cyclique (la prédiction d'un genre dépend de la prédiction de l'autre). Certaines approches de classification existantes gèrent ce problème en fixant d'abord une structure de dépendance acyclique entre les labels (Read et al.,2015,[95]). Les relations entre les labels sont ensuite apprises uniquement en accord avec la structure de dépendance fixée. L'inconvénient de ces approches est qu'elles ne vérifient pas le cas favorable où l'apprentissage des dépendances entre les labels sans une restriction au préalable n'engendre pas des dépendances cycliques. Une autre stratégie des approches de l'état de l'art est de combiner une approche ignorant les relations entre les labels, et une approche permettant l'apprentissage des dépendances entre les labels sans une restriction au préalable (Alvares-Cherman et al.,2012,[5];Montañés et al.,2011,[79];Montañés et al.,2014,[80]). L'approche ignorant les dépendances entre les labels permet de fournir une prédiction préliminaire. L'approche tenant compte des relations entre les labels fournit la prédiction finale en se basant sur la prédiction préliminaire en cas de dépendance cyclique. L'inconvénient de cette stratégie est la complexité due à la combinaison de deux approches. Dans cette thèse, nous nous intéressons à ce défi d'apprentissage des dépendances entre les labels sans restriction au préalable et en gardant une complexité minimale.

Les approches de l'état de l'art permettent d'apprendre soit des relations de préférence ((Fürnkranz et al.,2008,[37];Loza Mencía and Fürnkranz,2008,[72]);Brinker et al.,2014,[20]), soit des relations de dépendance entre les labels (Younes et al.,2011,[124];Wang et al.,2014,[116];Mena et al.,2016,[78]). Dans cette thèse, nous faisons l'hypothèse que la combinaison de l'apprentissage des deux types de relations entre les labels permet d'améliorer la prédiction. Nous étudions en conséquence différentes nouvelles approches de combinaison des relations de préférence et de dépendance entre les labels afin de valider cette hypothèse.

* Contraintes supplémentaires de certains domaines d'application :

Les systèmes de recommandation constituent un domaine d'application directe de la classification multi-labels graduée. Les données collectées par les systèmes de recommandation ont la particularité

d'être en perpétuelle croissance et d'avoir plusieurs valeurs manquantes. En effet, selon le contexte du système de recommandation (recommandation de films, de produits, de livres, etc) chaque internaute ne peut évaluer qu'un ensemble limité d'éléments disponibles. Ceci explique le fait que les données d'apprentissage pour le système de recommandation ne sont pas complètes. Les données d'apprentissage pour la classification multi-labels graduée sont supposées être complètes et n'évoluant pas avec le temps. Les approches de la classification multi-labels graduée ne peuvent donc pas être appliquées directement aux systèmes de recommandation. Dans cette thèse, nous nous intéressons au défi de construire un système de recommandation basé sur des approches de la classification multi-labels graduée en les adaptant aux contraintes des données manquantes et arrivant en flux.

La perception des odeurs est un domaine de recherche intéressant qui a fait l'objet du prix Nobel en 2004 (Buck and Axel,1991,[22]). Le prix a été attribué pour la découverte d'un ensemble de capteurs au niveau du nez permettant de coder le signal de l'odeur. Le cerveau reçoit ce signal via des neurones et le traduit en une sensation d'odeur. En effet, chaque odeur permet d'activer selon son intensité un sous ensemble différent de capteurs. De nombreuses recherches ont été menées par des biologistes, des chimistes, et des informaticiens afin d'expliquer l'odeur par les propriétés des molécules odorantes activant les capteurs du nez (Bosc et al.,2015,[16];Jha and Hayashi,2017,[52];Ucar:et:Recep:2017;). Cependant, les propriétés physico-chimiques qui ont étés étudiées ne sont pas assez discriminantes pour prédire l'ensemble gradué d'odeurs des molécules odorantes (Bosc et al.,2016,[17];de March et al.,2015,[28]). Nous nous intéressons dans cette thèse à ce défi du manque de variables discriminantes à travers un jeu de données multi-labels graduées de molécules odorantes (Arctander,1969,[8]).

Périmètre et contribution de la thèse

L'objectif de ce travail est d'établir de nouvelles approches de classification multi-labels graduée permettant de dépasser certaines limites des approches existantes.

Notre approche pour apprendre les relations entre les labels sans fixer un ordre de dépendance au préalable, appelée PSI, est basée sur trois mesures: Pré-sélection, Sélection, et Intérêt de chaînage. L'approche PSI consiste à construire pour chaque label un classifieur permettant de prédire son degré d'association. La prédiction peut être basée à la fois sur les attributs descriptifs et sur les prédictions des autres classifieurs. Le problème des dépendances cycliques est résolu grâce aux trois mesures de l'approche PSI. La mesure de pré-sélection permet de retrouver les classifieurs impliqués dans une dépendance cyclique. La mesure de sélection permet de sélectionner un classifieur à remplacer pour supprimer ses dépendances qui bloquent la prédiction. La mesure d'intérêt de chaînage permet de sélectionner des dépendances alternatives que le classifieur peut apprendre sans risquer de causer de nouvelles dépendances cycliques. Le remplacement d'un seul classifieur ne garantit pas la suppression de toutes les dépendances cycliques. Les trois mesures de l'approche PSI sont donc utilisées itérativement jusqu'à l'élimination de toutes les dépendances cycliques. L'avantage de l'approche PSI est qu'elle permet d'apprendre des relations de dépendance différentes selon les mesures PSI

utilisées. Les mesures PSI permettent donc de manipuler le compromis entre l'apprentissage d'un maximum de relations entre les labels pour assurer des prédictions cohérentes, et d'un minimum de relations entre les labels pour réduire le risque de la propagation de l'erreur de prédiction.

Notre approche PSI2 est une amélioration de l'approche PSI permettant de réduire davantage le risque de la propagation de l'erreur de prédiction. En effet, l'approche PSI permet d'apprendre une relation de dépendance lorsque le degré d'association d'un label est mieux identifié en se basant sur le degré d'association d'un autre label que sur les attributs descriptifs. Cependant, dans certains cas l'apport d'une dépendance par rapport à l'utilisation d'un attribut descriptif est minimal. La dépendance apprise dans ce cas ne fait qu'augmenter le risque de la propagation des erreurs de prédiction pour un label aux labels qui en dépendent. L'amélioration apportée par notre approche PSI2 consiste à introduire certaines conditions permettant d'éviter l'apprentissage des dépendances dont l'apport est minimal par rapport aux attributs descriptifs.

Nos deux approches CLR_PSI et Stacked_RPC_PSI permettent d'apprendre à la fois les relations de dépendance et de préférence entre les labels. L'idée est de combiner une approche existante permettant l'apprentissage des relations de préférence avec notre approche PSI permettant l'apprentissage des relations de dépendance. Cette idée est motivée par l'hypothèse que l'exploitation de plusieurs types de relations entre les labels peut améliorer la qualité des prédictions. Les deux nouvelles approches que nous proposons diffèrent au niveau de la stratégie de combinaison et au niveau de l'approche d'apprentissage de préférence utilisée.

Notre stratégie d'adaptation des approches de la classification multi-labels graduée est basée sur la combinaison de deux modèles de classification pour construire un système de recommandation. Le premier modèle de classification apprend à prédire le degré de satisfaction d'un utilisateur par rapport à un produit en se basant sur les informations de l'utilisateur, et sur ses degrés de satisfaction collectés auparavant. Le deuxième modèle de classification apprend à prédire le degré de satisfaction pour un produit par rapport à un utilisateur en se basant sur les informations du produit, et sur les degrés de satisfaction affectés auparavant à ce produit par d'autres utilisateurs. Chaque modèle de classification multi-labels graduée est incrémental (se met à jour à l'arrivée de nouvelles données d'apprentissage) afin de gérer le défi des données massives et arrivant en flux.

Notre approche de classification adaptée à la contrainte du manque de variables discriminantes dans le jeu de données des molécules odorantes s'effectue en deux étapes. La première étape consiste à construire d'abord des sous-ensembles de labels plus faciles à prédire en se basant sur les attributs descriptifs. Un premier modèle de classification est appris afin d'établir des prédictions pour les sous-ensembles de labels générés. La deuxième étape de l'approche consiste à construire un deuxième modèle de classification afin de fournir une prédiction finale des odeurs en se basant sur les prédictions du premier modèle.

Plan de la thèse

La suite de ce document est organisée comme suit:

Le chapitre 1 présente les approches existantes en classification multi-labels graduée. Les approches de l'état de l'art sont étudiées et comparées par rapport à certains critères tels que la complexité et la capacité d'apprendre les relations entre les labels .

Le chapitre 2 présente les nouvelles approches de classification que nous proposons (PSI, PSI2, CLR_PSI, et Stacked_RPC_PSI). Nos approches sont comparées aux approches existantes par rapport à des jeux de données multi-labels et multi-labels graduées.

Le chapitre 3 présente d'abord une synthèse des approches existantes pour les systèmes de recommandation. Ensuite, notre approche permettant de construire un système de recommandation basé sur une adaptation de la classification multi-labels graduée est détaillée. Notre approche est comparée aux approches de l'état de l'art par rapport à des jeux de données différents.

Le chapitre 4, présente d'abord une synthèse des approches existantes pour la prédiction des odeurs. Ensuite notre approche adaptée à la contrainte du manque de variables discriminantes est détaillée. Notre approche est comparée aux approches de l'état de l'art par rapport à un jeu de données de molécules odorantes multi-odeurs.

Nous discutons dans la conclusion l'efficacité des nouvelles approches que nous avons proposé et la validation de nos hypothèses. Nous terminons par la présentation des perspectives de notre travail et des pistes d'amélioration possibles.

Liste des articles publiés

- Laghmari et al.,2016,[60]: L'approche PSI ;
- Laghmari et al.,2017,[63] : L'approche PSI (description plus affinée) ;
- Laghmari et al.,2017,[62]: L'approche de construction d'un système de recommandation basé sur la classification multi-labels graduée ;
- Laghmari et al.,2017,[61] : La version distribuée de l'approche de construction d'un système de recommandation basé sur la classification multi-label graduée.

Liste des articles acceptés

- Laghmari et al.,2018,[66] : L'approche Stacked_RPC_PSI ;
- Laghmari et al.,2018,[64] : Les approches PSI, PSI2, CLR_PSI, et Stacked_RPC_PSI.

Liste des articles soumis

Laghmari et al.,2018,[65] : L'approche PSI2 (description détaillée).

Chapitre 1

Classification multi-labels graduée

La classification est une branche de l'apprentissage supervisé qui consiste à apprendre à partir de données associées à des classes un classifieur pour prédire la classe des données non encore associées à une classe. La classification est dite mono-label lorsque chaque instance peut être associée à exactement une classe. La classification mono-label est dite binaire lorsque l'ensemble des classes disponibles contient exactement deux éléments. Elle est dite multi-classes lorsque l'ensemble des classes disponibles contient plus de deux éléments. La classification multi-classes est dite ordinale lorsque l'ensemble des classes disponibles est ordonné (Figure 1.1).

La classification est dite multi-labels lorsque chaque instance peut être associée à plusieurs classes (labels) en même temps. Elle est dite floue lorsque l'association entre une instance et un label possède un degré dans l'intervalle $[0, 1]$. *La classification est dite multi-labels graduée lorsque les degrés d'association appartiennent à un ensemble ordonné de degrés d'appartenance. La classification multi-labels graduée peut être décomposée en plusieurs tâches de classification multi-labels et de classification ordinale qui peuvent eux même être décomposées en plusieurs tâches de classification binaire (Figure 1.1).*

Ce chapitre présente formellement les différents types de classification et les approches de l'état de l'art qui y sont adaptées. Ensuite il présente les stratégies permettant de réduire le problème de la classification multi-labels graduée à des problèmes de classification multi-labels et de classification ordinale. Les approches de l'état de l'art sont comparées selon leur complexité et par rapport à leur capacité d'apprendre les relations entre les labels.

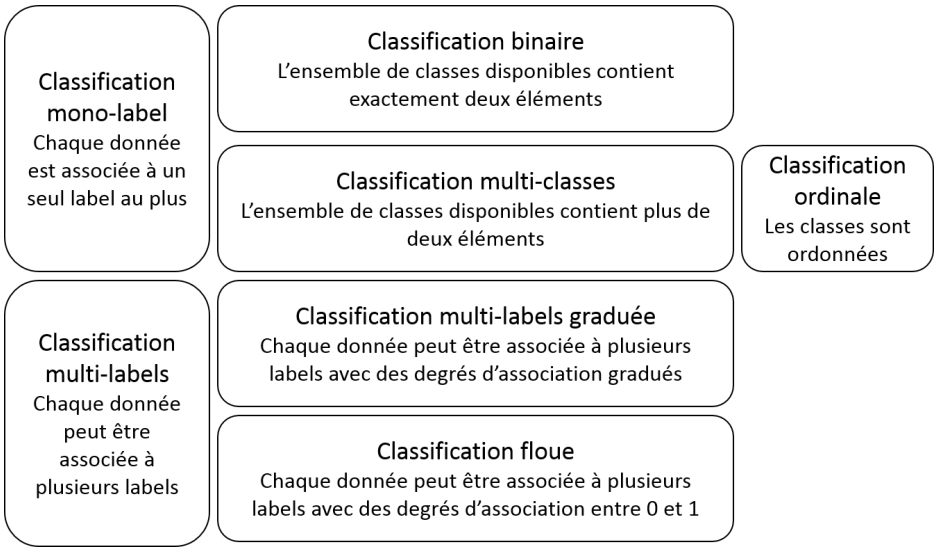


Figure 1.1: Types de classification

1.1 Classification mono-label

La reconnaissance des formes constitue une compétence d'intelligence fondamentale. Par exemple, en considérant un ensemble d'images et un ensemble de classes {coucher du soleil, montagne, plage}, l'intelligence humaine est capable d'associer chaque image à la classe correspondante. La reconnaissance de formes s'étend à d'autres domaines tels que la reconnaissance d'odeurs. Par exemple, en considérant un ensemble de molécules odorantes, et un ensemble de classes correspondant à des qualités olfactives: {'JASMIN', 'MINTY', 'MUSK'}, l'intelligence humaine est capable d'associer chaque molécule odorante à la qualité olfactive ressentie. D'une façon générale, en considérant un ensemble de classes disponibles, l'intelligence humaine est capable d'associer chaque instance (image, molécule odorante, etc) à la classe correspondante.

Le défi de transférer la compétence de reconnaissance de formes aux machines fait partie du domaine de l'apprentissage artificiel. L'apprentissage est dit supervisé lorsqu'un ensemble d'instances dont les classes associées sont connues est disponible. L'apprentissage consiste à établir des liens entre les attributs descriptifs des instances et les classes associées. L'objectif de l'apprentissage est de généraliser les liens appris afin de pouvoir prédire la classe associée à chaque instance en se basant sur les attributs descriptifs.

La classification mono-label est une branche de l'apprentissage supervisé où chaque instance est décrite par des attributs descriptifs, et est associée à exactement une classe parmi un ensemble de classes disponibles. Le nombre de classes disponibles et le fait que les classes soient ordonnées ou non déterminent la nature de la tâche de classification mono-label.

Le Tableau 1.1 illustre un jeu de données où les attributs descriptifs correspondent aux propriétés physico-chimique des molécules odorantes, et les classes disponibles correspondent à la présence ou l'absence de l'odeur du jasmin (Arctander,1969,[8]). L'attribut 'Structure' est catégorique et peut

prendre les valeurs {A,C,L}. L'attribut 'Masse moléculaire' est numérique et peut prendre des valeurs positives. L'ensemble de classes disponibles contient deux éléments: la valeur 0 correspondant à l'absence de l'odeur du jasmin, et la valeur 1 correspondant à la présence de l'odeur du jasmin. L'attribut 'JASMIN' dont les valeurs correspondent aux classes disponibles est dit attribut de décision. La tâche de classification mono-label dans le Tableau 1.1 est dite binaire car l'ensemble de classes disponibles contient exactement deux éléments {0,1}.

Le Tableau 1.2 illustre un jeu de données où l'ensemble de classes disponibles contient trois éléments {JASMIN, MINTY, MUSK} correspondant à l'odeur du jasmin, de la menthe, et du musc. La tâche de classification mono-label dans ce cas est dite multi-classes parce que le nombre de valeurs distinctes prises par l'attribut de décision 'Odeur' est supérieur à deux.

Le Tableau 1.3 illustre un jeu de données où l'ensemble de classes disponibles est ordonné {Weak < Moderate < Strong}. La tâche de classification mono-label dans ce cas est dite ordinale.

Molécule	Structure	Masse moléculaire	JASMIN
$C_8H_8O_2$	A	136.16	1
$C_9H_{10}O_2$	A	150.19	1
$C_9H_{12}O$	A	182.29	1
$C_{11}H_{14}O_3$	A	194.25	1
$C_{14}H_{14}O$	A	306.43	1
$C_7H_8O_3$	C	140.15	0
$C_9H_{16}O$	C	140.25	0
$C_{10}H_{16}O$	C	154.28	1
$C_9H_{16}O_2$	C	156.25	1
$C_{13}H_{22}O$	C	194.35	1
$C_{13}H_{20}O_3$	C	224.33	1
$C_8H_{16}O_2$	L	144.24	0
$C_{10}H_{16}O$	L	152.26	0
$C_{10}H_{16}O_2$	L	168.26	0

Tableau 1.1: Données d'apprentissage pour la prédiction de l'absence ou la présence de l'odeur 'JASMIN'

Formule	Structure	Masse moléculaire	Odeur
$C_{12}H_{18}O$	A	266.47	JASMIN
$C_{13}H_{22}O$	C	194.35	JASMIN
$C_{11}H_{14}O_3$	A	194.25	JASMIN
$C_{12}H_{22}O_2$	C	198.34	MINTY
$C_{13}H_{20}O_2$	C	208.33	MINTY
$C_{10}H_{14}O_2$	C	302.45	MINTY
$C_{18}H_{34}O_4$	L	257.39	MUSK
$C_6H_{10}O_4$	L	146.16	MUSK
$C_{16}H_{30}O_2$	L	254.46	MUSK

Tableau 1.2: Données d'apprentissage pour la prédiction d'une odeur dans l'ensemble {JASMIN, MINTY, MUSK}

Formule	Structure	Masse moléculaire	JASMIN
$C_{10}H_{16}O$	C	154.28	Strong
$C_9H_{16}O_2$	C	156.25	Strong
$C_{11}H_{14}O_2$	A	178.25	Weak
$C_{12}H_{22}O$	C	182.34	Weak
$C_{12}H_{16}O_2$	A	192.28	Moderate
$C_{13}H_{20}O$	A	192.33	Moderate
$C_{13}H_{22}O$	C	194.35	Strong
$C_{14}H_{18}O$	A	202.32	Moderate
$C_{12}H_{22}O_3$	L	214.34	Weak
$C_{13}H_{16}O_3$	A	220.29	Weak
$C_{18}H_{28}O_2$	A	276.46	Weak

Tableau 1.3: Données d'apprentissage pour la prédiction du degré d'association de l'odeur 'JASMIN'

1.1.1 Description formelle de la classification mono-label

Soit $X = \{x_i\}_{1\leq i\leq n}$ un ensemble d'instances (base d'apprentissage). Soit $A = \{a_j\}_{1\leq j\leq p}$ un ensemble d'attributs descriptifs. Chaque instance x_i est un vecteur de valeurs d'attributs descriptifs $x_i = (x_{i,a_1}, \ldots, x_{i,a_p}) = (x_{i,a_j})_{1\leq j\leq p}$.

Soit $C = \{c_l\}_{1\leq l\leq k}$ un ensemble de classes. Chaque instance x_i est associée à exactement une classe $y_i \in C$.

Soit $\lambda : X \to C$ la fonction qui associe à chaque instance $x_i \in X$ sa classe correspondante: $\lambda(x_i) = y_i$. La fonction λ est dite fonction de supervision.

La classification mono-label est dite binaire lorsque $k = 2$. La classification mono-label est dite multi-classes lorsque $k > 2$. La classification multi-classes est dite ordinale lorsque l'ensemble de classes disponibles est ordonné: $c_1 < \ldots < c_k$.

Soit $E : C \times C \to [0,1]$ une fonction objectif à optimiser (minimiser ou maximiser).

La classification mono-label consiste à apprendre à partir de l'ensemble d'apprentissage muni de la fonction de supervision (X, λ), un classifieur $H : a_1 \times \ldots \times a_p \to C$ permettant de prédire pour tout $x \in a_1 \times \ldots \times a_p$ la classe correspondante $H(x)$ de façon à optimiser la fonction objectif $E(y, H(x))$, où y étant la classe effectivement associée à l'instance x.

1.1.2 Classifieurs mono-label

Plusieurs classifieurs sont proposés dans l'état de l'art pour relever le défi de la classification binaire (Kumari and Srivastava,2017,[59]) et de la classification multi-classes (Mehra and Gupta,2013,[77]). Parmi les classifieurs les plus utilisés se trouvent les réseaux de neurones (Bhardwaj et al.,2016,[13]), les séparateurs à vaste marge (Ding et al.,2017,[30]), les k-plus proches voisins (Tang and Xu,2016,[105]), les classifieurs de bayes (Šuch and Barreda,2016,[111]), les règles de décision (Tan et al.,2005,[104]), et les arbres de décision (Yan-yan and Ying,2015,[122]).

La Figure 1.2 illustre un arbre de décision (Quinlan,1993,[89]) correspondant aux données du Tableau 1.1. La racine de l'arbre est un nœud qui correspond à l'attribut descriptif 'Structure'. Les arcs sortants d'un nœud sont étiquetés par les valeurs prises par l'attribut correspondant. Les nœuds qui n'ont pas d'arcs sortants sont appelés nœuds de décision ou feuilles et permettent d'établir des prédictions.
Au niveau de la racine dans la Figure 1.2, 5 instances sont associées à la classes 0 (absence de l'odeur 'JASMIN'), et 9 instances sont associées à la classe 1 (présence de l'odeur 'JASMIN'). Les cinq instances de l'ensemble d'apprentissage ayant la valeur 'A' pour l'attribut 'Structure' sont tous associées à la classe 1. La décision à prendre pour les instances ayant la valeur 'A' pour l'attribut 'Structure' est de prédire la classe 1 pour l'attribut de décision 'JASMIN' (Figure 1.2). L'arbre de décision de la Figure 1.2 permet d'extraire l'ensemble des règles de décision suivant:

- si 'Structure' = 'A' alors 'JASMIN' = '1'

- si 'Structure' = 'C' et 'Masse moléculaire' $\leq$ '140.25' alors 'JASMIN' = '0'
- si 'Structure' = 'C' et 'Masse moléculaire' > '140.25' alors 'JASMIN' = '1'
- si 'Structure' = 'L' alors 'JASMIN' = '0'

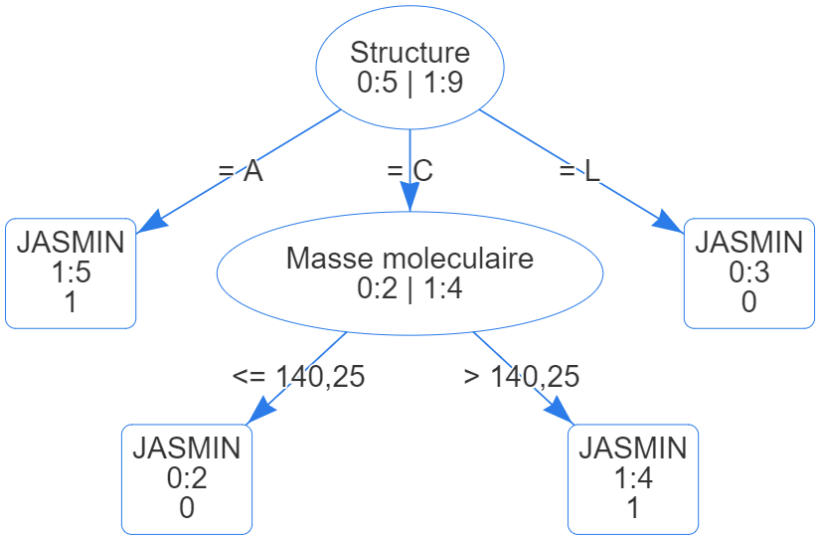


Figure 1.2: Arbre de décision de l'odeur 'JASMIN' à partir des données du Tableau 1.1

L'idée de la méthode des séparateurs à vaste marge (Boser et al.,1992,[18]) en classification binaire est de trouver l'hyperplan optimal qui sépare les deux classes, et qui maximise la distance (marge) entre ce plan et les plus proches instances (vecteurs de support) (Vapnik and Lerner,1963,[115]).

La méthode des séparateurs à vaste marge fonctionne uniquement pour les attributs descriptifs numériques. Les attributs catégoriques doivent être convertis en attributs numériques. Lorsque les instances ne sont pas linéairement séparables, l'idée est de transformer l'espace de représentation des instances en un espace de plus grande dimension dans lequel les instances sont linéairement séparables. Cette technique est connue sous le nom de *l'astuce du noyau* (kernel trick) (Aizerman et al.,1964,[4]).

1.1.3 Transformation d'un problème multi-classes en un problème binaire

Certains classifieurs tels que les arbres de décision sont adaptés au cas multi-classes, d'autres classifieurs tels que les séparateurs à vaste marge sont adaptés uniquement au cas de la classification binaire. Les approches de transformation des données multi-classes en données de classification binaires permettent d'exploiter les classifieurs binaires dans des cas multi-classes.

Les deux méthodes de transformation les plus connues sont la méthode *un contre tout* (OVA: One Vs All) (Friedman,1996,[36]), et la méthode *un contre un* (OVO: One Vs One) (Hastie and Tibshirani,1998,[47]).

La méthode OVA consiste à construire un classifieur multi-classes H à partir de k classifieurs binaires $\{H_l\}_{1\leq l\leq k}$, un pour chaque classe.

Soit $\lambda_l : X \to \{0,1\}$ la fonction qui associe à chaque instance $x_i \in X$ la valeur 1 si l'instance x_i est associée au label c_l, et la valeur 0 sinon. La fonction λ_l est dite fonction de pertinence du label c_l (relevance of c_l), et constitue une fonction de supervision pour le classifieur H_l.

Chaque classifieur $H_l : a_1 \times \ldots \times a_p \to \{0,1\}$ apprend à partir de l'ensemble supervisé (X, λ_l) à prédire la pertinence du label c_l pour toute instance $x \in a_1 \times \ldots \times a_p$.

La prédiction $H_l(x) \in \{0,1\}$ du classifieur H_l pour une instance x est donnée avec une confiance associée $Conf(H_l(x)) \in [0,1]$. La classe finalement prédite par H est celle correspondant à la plus grande confiance: $H(x) = c_l$ telle que $\forall l' \neq l : Conf(H_l(x)) \geq Conf(H_{l'}(x))$.

La méthode OVO consiste à construire un classifieur multi-classes H à partir de $\frac{(k-1)\times k}{2}$ classifieurs binaires $\{H_{l,l'}\}_{1\leq l<l'\leq k}$, un pour chaque paire de classes.

Soit $X_{l,l'} = \{x_i \in X, \lambda(x_i) = c_l \; ou \; \lambda(x_i) = c_{l'}\}$ l'ensemble d'instances associées soit au label c_l soit au label $c_{l'}$.

Soit $\lambda_{l,l'} : X_{l,l'} \to \{c_l, c_{l'}\}$ la fonction qui associe à chaque instance $x_i \in X_{l,l'}$ le label qui lui est associé (c_l ou $c_{l'}$). La fonction $\lambda_{l,l'}$ est dite fonction de préférence entre c_l et $c_{l'}$, et constitue une fonction de supervision pour le classifieur $H_{l,l'}$.

Chaque classifieur $H_{l,l'} : a_1 \times \ldots \times a_p \to \{c_l, c_{l'}\}$ apprend à partir de l'ensemble supervisé $(X_{l,l'}, \lambda_{l,l'})$ à prédire pour chaque instance $x \in a_1 \times \ldots \times a_p$ le label préféré entre c_l et $c_{l'}$.

Soit $V_{c_l} : a_1 \times \ldots \times a_p \to [\![0, k-1]\!]$ la fonction qui donne pour chaque instance $x \in a_1 \times \ldots \times a_p$ le nombre de fois où le label c_l est préféré par les classifieurs $\{H_{l',l''}\}_{1\leq l'<l''\leq k}$ (nombre de votes donnés par les classifieurs $\{H_{l',l''}\}_{1\leq l'<l''\leq k}$ au label c_l).

La classe finalement prédite par H est celle qui a reçue le maximum de votes par les classifieurs binaires: $H(x) = c_l$ telle que $\forall l' \neq l$: $V_{c_l}(x) \geq V_{c_{l'}}(x)$

Dans le cas de la classification ordinale, l'apprentissage et la prédiction doivent tenir compte de l'ordonnancement des classes. Une des approches les plus utilisées pour transformer une tâche de classification ordinale en une tâche de classification binaire consiste à construire un classifieur ordinal H à partir de $k-1$ classifieurs binaires $\{H_l\}_{2\leq l\leq k}$ (Frank and Hall,2001,[35]).

Soit $\lambda_{\geq c_l} : X \to \{0,1\}$ la fonction qui associe à chaque instance $x_i \in X$ la valeur 1 si x_i est associée à un label $c_{l'}$ d'ordre au moins égal à celui de c_l ($c_{l'} \geq c_l$), et la valeur 0 sinon.

Chaque classifieur $H_l : a_1 \times \ldots \times a_p \to \{0,1\}$ apprend à partir de l'ensemble supervisé $(X, \lambda_{\geq c_l})$ à prédire pour chaque instance $x \in a_1 \times \ldots \times a_p$ si elle est associée à un label $c_{l'} \geq c_l$ ($H_l(x) = 1$) ou pas ($H_l(x) = 0$).

Chaque classifieur H_l renvoie pour toute instance $x \in a_1 \times \ldots \times a_p$ la probabilité de prédiction de la valeur 1: $Pr(H_l(x) = 1)$. Cette probabilité correspond à la probabilité que l'instance x soit associée à un label d'ordre supérieur ou égale à celui de c_l: $Pr(H_l(x) = 1) = Pr(H(x) \geq c_l)$.

La probabilité qu'une instance x soit associée à une classe $c_l \in C$ est calculée de la façon suivante (Frank and Hall,2001,[35]) :

- $Pr(H(x) = c_1) = 1 - Pr(H(x) \geq c_2) = 1 - Pr(H_2(x) = 1)$
- $Pr(H(x) = c_k) = Pr(H(x) \geq c_k) = Pr(H_k(x) = 1)$
- $\forall l \in]1,k[: Pr(H(x) = c_l) = Pr(H(x) \geq c_l) - Pr(H(x) \geq c_{l+1})$
  $= Pr(H_l(x) = 1) - Pr(H_{l+1}(x) = 1)$

La classe prédite finalement est celle ayant la plus grande probabilité: $H(x) = c_l$ telle que $\forall l' \in [1,k]$: $Pr(H(x) = c_l) \geq Pr(H(x) = c_{l'})$.

Le Tableau 1.4 illustre un jeu de données multi-classes de six instances et trois classes disponibles. Le Tableau 1.5 illustre la transformation par la méthode OVA. Le Tableau 1.6 illustre la transformation par la méthode OVO. Le Tableau 1.7 illustre la transformation de la tâche de classification ordinale en plusieurs tâches de classification binaire en considérant que l'ensemble de classes dans le Tableau 1.4 est ordonné $\{c_1 < c_2 < c_3\}$.

Instances	Classe
x_1	c_1
x_2	c_1
x_3	c_2
x_4	c_2
x_5	c_3
x_6	c_3

Tableau 1.4: Instances multi-classes

Instances	c_1 Vs All
x_1	1
x_2	1
x_3	0
x_4	0
x_5	0
x_6	0

Instances	c_2 Vs All
x_1	0
x_2	0
x_3	1
x_4	1
x_5	0
x_6	0

Instances	c_3 Vs All
x_1	0
x_2	0
x_3	0
x_4	0
x_5	1
x_6	1

Tableau 1.5: Transformation-OVA des données du Tableau 1.4

1.1.4 Conclusion

Les approches de transformation de la classification multi-classes (nominale ou ordinale) en classification binaire ont deux intérêts principaux:

Instances	c_1 Vs c_2
x_1	c_1
x_2	c_1
x_3	c_2
x_4	c_2

Instances	c_1 Vs c_3
x_1	c_1
x_2	c_1
x_5	c_3
x_6	c_3

Instances	c_2 Vs c_3
x_3	c_2
x_4	c_2
x_5	c_3
x_6	c_3

Tableau 1.6: Transformation-OVO des données du Tableau 1.4

Instances	$c_l \geq c_2$?
x_1	0
x_2	0
x_3	1
x_4	1
x_5	1
x_6	1

Instances	$c_l \geq c_3$?
x_1	0
x_2	0
x_3	0
x_4	0
x_5	1
x_6	1

Tableau 1.7: Transformation ordinale $\rightarrow$ binaire des données du Tableau 1.4 en considérant $c_1 < c_2 < c_3$

- elles permettent d'exploiter les classifieurs binaires existants dans des cas multi-classes.
- elles constituent la base des approches de transformation de la classification multi-labels en classification mono-label (Section 1.2), et de la classification multi-labels graduée en classification multi-labels ou en classification ordinale (Section 1.3).

1.2 Classification multi-labels

En classification multi-labels, chaque instance est décrite par des attributs descriptifs et elle est associée à une ou plusieurs classes (labels) parmi un ensemble de labels disponibles.

Le Tableau 1.8 illustre un jeu de données multi-labels où chaque molécule odorante (instance) peut être associée à une ou plusieurs odeurs (labels) appartenant à l'ensemble d'odeurs {HERBAC,FRUITY,JASMIN} (Arctander,1969,[8]).

Les cas multi-labels sont généralement traités selon deux types d'approches: les approches permettant d'adapter un classifieur mono-label au cas multi-labels (Section 1.2.3), et les approches permettant de transformer une tâche de classification multi-labels en une ou plusieurs tâches de classification mono-label (Section 1.2.4) (Tsoumakas and Katakis,2007,[107] ; Zhang and Zhou,2014,[130]). Chaque approche se distingue par sa complexité et sa capacité à apprendre les relations entre les labels.

La Figure 1.3 illustre un arbre de décision construit à partir des données du Tableau 1.8 pour prédire l'absence ou la présence de l'odeur 'HERBAC' dans les molécules odorantes. La particularité de cet arbre de décision est qu'il considère le label 'JASMIN' en tant qu'un attribut descriptif aidant à la prédiction du label 'HERBAC' (nœud coloré en bleu dans l'arbre de décision de la Figure 1.3).

Le nœud 'JASMIN' dans l'arbre de décision de l'odeur 'HERBAC' est appelé un nœud label. La présence d'un nœud label dans un arbre de décision reflète une relation de dépendance entre les labels. Par exemple, la prédiction de l'odeur 'HERBAC' en utilisant l'arbre de décision de la Figure 1.3 dépend de la prédiction de l'odeur 'JASMIN'.

Afin de fournir des prédictions multi-labels en se basant sur les données d'apprentissage du Tableau 1.8, un arbre de décision est construit pour chaque label (Figure 1.3, Figure 1.4, et Figure 1.5). Les dépendances entre les labels extraites à partir des arbres de décision appris permettent d'établir une structure de dépendance (Figure 1.6). L'arc dont le sens est orienté du noeud 'JASMIN' vers le noeud 'HERBAC' dans la Figure 1.6 indique que la prédiction du label 'HERBAC' dépend de la prédiction du label 'JASMIN'. Les arcs dans une structure de dépendance indiquent l'ordre de prédiction qu'il faut suivre.

Par exemple, selon la structure de dépendance de la Figure 1.6, il faut d'abord prédire l'absence ou la présence du label 'JASMIN' pour pouvoir prédire l'absence ou la présence du label 'HERBAC'.

Les dépendances entre les labels constituent une information supplémentaire qui peut intéresser les experts du domaine des données d'apprentissage. Cependant, leur inconvénient est qu'elle permettent la propagation de l'erreur en prédiction pour un label aux labels qui en dépendent (Senge et al.,2014,[99]).

La structure de dépendance de la Figure 1.6 présente une dépendance cyclique entre les labels 'FRUITY' et 'JASMIN'. L'ordre de prédiction ne peut pas être identifié dans ce cas. Plusieurs approches de classification multi-labels de l'état de l'art évitent ce problème en fixant au préalable les dépendances que peut apprendre chaque classifieur (Read et al.,2011,[94];Read et al.,2015,[95]).

L'un des objectifs de l'étude des approches existantes en classification multi-labels est de comparer leurs complexités et les structures de dépendance qu'elles permettent d'apprendre. Cette étude permet de guider la sélection d'une approche existante ou la proposition d'une nouvelle approche optimisée en terme de complexité et de capacité d'apprendre les relations entre les labels.

Formule	Structure	Masse moléculaire	Odeurs
$C_{12}H_{18}O$	A	266.47	{HERBAC,JASMIN}
$C_{10}H_{18}O$	L	154.28	{FRUITY,JASMIN}
$C_9H_{16}O_2$	C	156.25	{FRUITY,JASMIN}
$C_{12}H_{16}O_2$	A	192.28	{HERBAC,FRUITY,JASMIN}
$C_{12}H_{22}O$	C	182.34	{HERBAC,FRUITY,JASMIN}
$C_{14}H_{14}O$	A	306.43	{JASMIN}
$C_{11}H_{14}O_3$	A	194.25	{JASMIN}
$C_{18}H_{32}O_2$	L	280.5	{HERBAC,FRUITY}
$C_{14}H_{28}O_2$	L	228.42	{HERBAC,FRUITY}
$C_9H_{12}O_2$	C	152.21	{HERBAC}
$C_{10}H_{16}O_2$	C	168.26	{HERBAC}

Tableau 1.8: Données d'apprentissage pour la prédiction de sous-ensembles d'odeurs

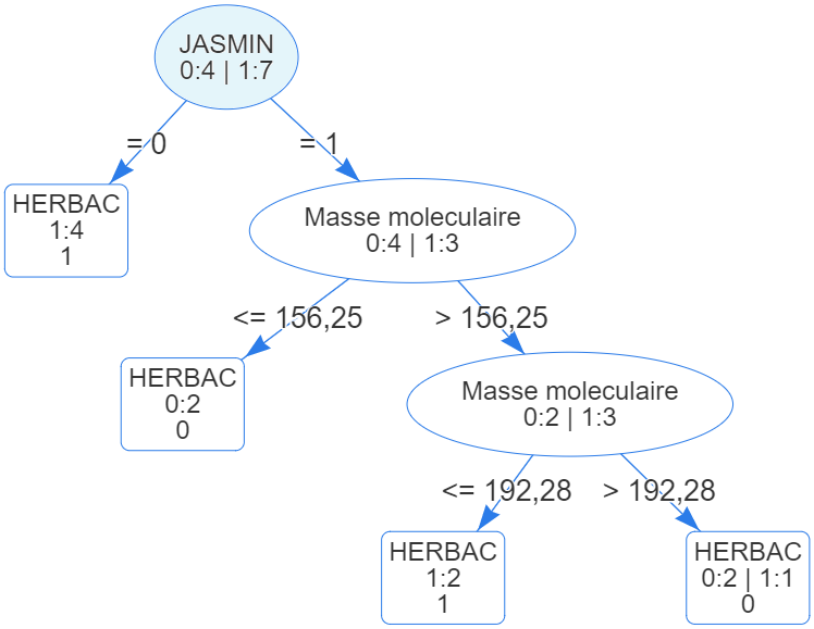


Figure 1.3: Arbre de décision pour prédire l'absence ou la présence de l'odeur 'HERBAC' à partir des données du Tableau 1.8

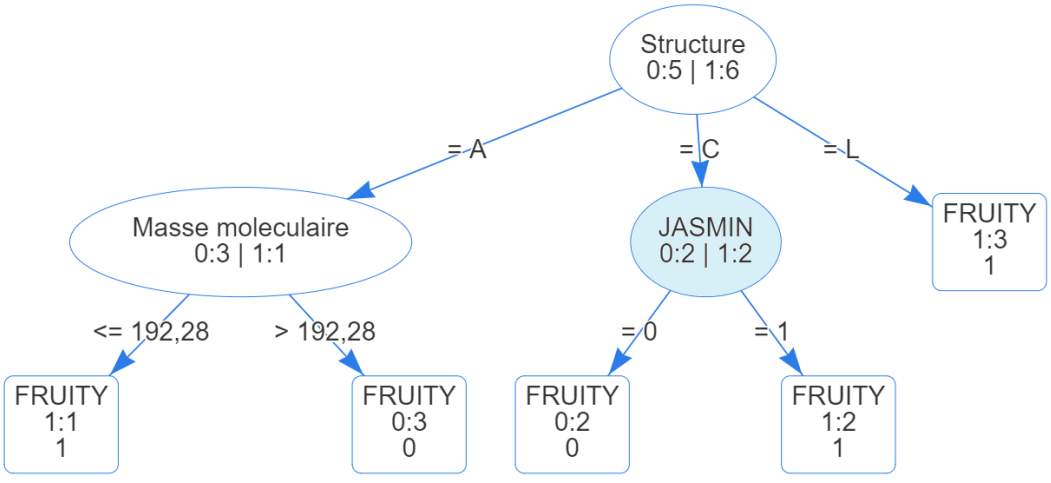


Figure 1.4: Arbre de décision pour prédire l'absence ou la présence de l'odeur 'FRUITY' à partir des données du Tableau 1.8

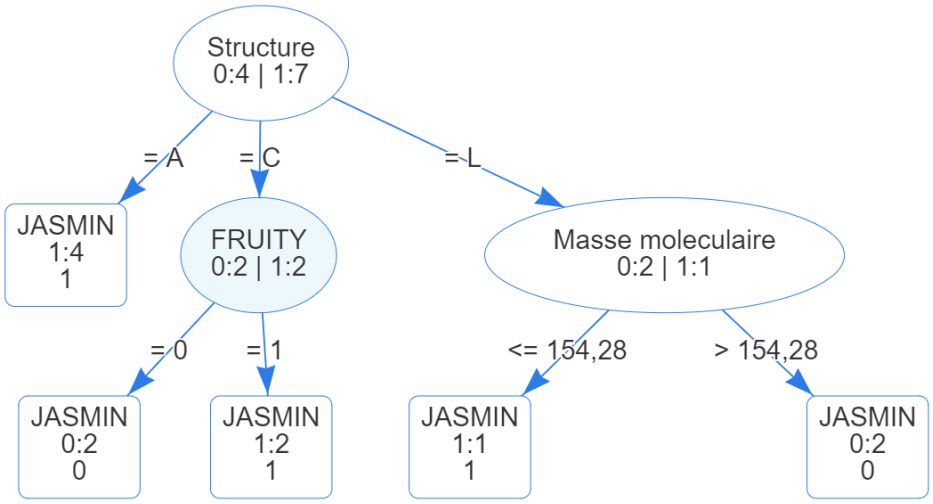


Figure 1.5: Arbre de décision pour prédire l'absence ou la présence de l'odeur 'JASMIN' à partir des données du Tableau 1.8

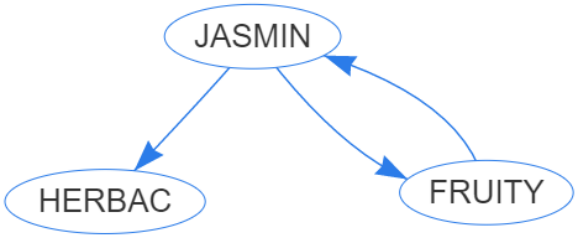


Figure 1.6: Structure de dépendance entre les labels 'HERBAC', 'FRUITY', et 'JASMIN' à partir des arbres de décision illustrés dans les figures 1.3, 1.4, et 1.5

1.2.1 Description formelle de la classification multi-labels

Soit $X = \{x_i\}_{1 \leq i \leq n}$ un ensemble d'instances. Soit $A = \{a_j\}_{1 \leq j \leq p}$ un ensemble d'attributs descriptifs. Chaque instance x_i est un vecteur de valeurs d'attributs descriptifs $(x_{i,a_j})_{1 \leq j \leq p}$. Soit $C = \{c_l\}_{1 \leq l \leq k}$ un ensemble de labels. Chaque instance x_i est associée à un sous-ensemble de labels $y_i \subseteq C$.

L'ensemble de sous-ensembles de C est noté $\mathscr{P}(C)$.

Soit $E : \mathscr{P}(C) \times \mathscr{P}(C) \rightarrow [0,1]$ une fonction objectif à optimiser (à minimiser ou à maximiser).

Soit $\lambda : X \rightarrow \mathscr{P}(C)$ la fonction qui associe à chaque instance $x_i \in X$ le sous-ensemble de labels correspondant $\lambda(x_i) = y_i$.

La classification multi-labels consiste à apprendre un classifieur $H : a_1 \times \ldots \times a_p \rightarrow \mathscr{P}(C)$ à partir de l'ensemble supervisé (X, λ) permettant de prédire pour chaque instance $x \in a_1 \times \ldots \times a_p$ l'ensemble de labels correspondant $H(x)$ de façon à optimiser l'évaluation de la fonction objectif $E(y, H(x))$, où y est le sous-ensemble de labels effectivement associé à l'instance x.

1.2.2 Mesures d'évaluation de la classification multi-labels

Dans le cadre de la classification multi-labels, nous avons besoin d'introduire des mesures d'évaluation mettant en avant l'aspect multi-labels des données, la complexité de l'approche de classification, les relations entre les labels, le risque de la propagation des erreurs de prédiction, et la performance de l'approche de classification en prédiction.

Complexité de données

La performance des classifieurs multi-labels peut dépendre de la répartition des labels par rapport aux données. Trois différentes mesures permettent d'évaluer la complexité d'un jeu de données multi-labels (Tsoumakas and Katakis,2007,[107]):

- la cardinalité (label cardinality) qui évalue la moyenne du nombre de labels associés à une instance: $\mathbb{LC} = \frac{1}{n} \sum_{i=1}^{n} |\lambda(x_i)|$.
- la densité (label density) qui évalue la moyenne du nombre de labels associés à une instance par rapport au nombre total de labels : $\mathbb{LD} = \frac{1}{n} \sum_{i=1}^{n} \frac{|\lambda(x_i)|}{k} = \frac{\mathbb{LC}}{k}$.
- le nombre de combinaisons distinctes de labels (distinct label combinations) qui évalue le nombre de sous-ensembles différents de labels qui sont associés aux instances : $\mathbb{DLC} = |\{\lambda(x_i)\}_{1 \leq i \leq n}|$.

Complexité du classifieur

Plusieurs mesures peuvent être utilisées pour évaluer la complexité des classifieurs multi-labels.

Par exemple, dans le cas d'un arbre de décision, la complexité peut être évaluée en fonction du nombre de nœuds et de feuilles, et en fonction de la profondeur des branches de l'arbre de décision.

Dépendances entre les labels

L'évaluation des dépendances entre les labels dépend du classifieur utilisé. Par exemple, dans le cas d'un arbre de décision, le nombre de nœuds labels dans l'arbre de décision, et le nombre de nœud labels dans une même branche de l'arbre de décision peuvent être utilisés comme mesures descriptives des dépendances apprises.

Propagation de l'erreur de prédiction

Soit $H_l : a_1 \times \ldots \times a_p \rightarrow \{0,1\}$ et $H_{l'} : a_1 \times \ldots \times a_p \rightarrow \{0,1\}$ deux classifieurs permettant de prédire respectivement la pertinence (présence ou absence) du label c_l et du label $c_{l'}$. Soit $x \in a_1 \times \ldots \times a_p$ une instance associée au sous-ensemble de labels $y \subseteq C$. La prédiction de la pertinence du label c_l pour l'instance x est correcte dans les deux cas suivants:

- $H_l(x) = 1$ et $c_l \in y$.
- $H_l(x) = 0$ et $c_l \notin y$.

Soit R : "$\forall x \in a_1 \times \ldots \times a_p$: *si* $H_{l'}(x) = 1$ *alors* $H_l(x) = 1$" un exemple d'une règle de décision permettant de prédire la pertinence du label c_l en se basant sur la pertinence du label $c_{l'}$. Dans le cas de la règle R, c_l est dit attribut de décision et $c_{l'}$ est dit attribut label puisqu'il ne fait pas partie des attributs descriptifs. Quatre cas différents peuvent être distingués lorsque la règle R est utilisée:

- une prédiction correcte $H_{l'}(x)$ qui permet d'obtenir une prédiction correcte $H_l(x)$.
- un prédiction incorrecte $H_{l'}(x)$ qui permet quand même d'obtenir une prédiction correcte $H_l(x)$.
- malgré une prédiction correcte $H_{l'}(x)$ la prédiction $H_l(x)$ est incorrecte.
- une prédiction incorrecte $H_{l'}(x)$ qui cause une prédiction incorrecte $H_l(x)$ (propagation de l'erreur de prédiction pour le label $c_{l'}$ au label dépendant c_l).

Une erreur de prédiction pour l'attribut de décision c_l est notée 'err-decision', et une erreur de prédiction pour l'attribut label $c_{l'}$ est notée 'err-label'. Nous avons mis en place des mesures permettant d'évaluer le risque de la propagation des erreurs de prédiction que nous détaillons dans la suite.

Soit X' un ensemble de données différent de l'ensemble d'apprentissage X. Les mesures suivantes permettant d'évaluer la propagation de l'erreur de prédiction pour les instances dans X' sont directement basées sur les quatre cas précédents:

- le nombre des prédictions incorrectes basées sur une valeur incorrecte d'un attribut label noté $|\textit{err-label} \Rightarrow \textit{err-decision}|_{X'}$.

- le nombre de prédictions correctes basées sur une valeur incorrecte d'un attribut label noté $|\textit{err-label} \Rightarrow \textit{not-err-decision}|_{X'}$.
- le nombre de prédictions incorrectes basées sur une valeur correcte d'un attribut label noté $|\textit{not-err-label} \Rightarrow \textit{err-decision}|_{X'}$.
- le nombre de prédictions correctes basées sur une valeur correctes d'un attribut label noté $|\textit{not-err-label} \Rightarrow \textit{not-err-decision}|_{X'}$.

L'inconvénient des mesures précédentes est qu'elles ne sont pas normalisées et dépendent du nombre de dépendances entre les labels. Les mesures suivantes permettent de remédier à cet inconvénient:

- le risque de la propagation d'erreur (Risk of Error Propagation) mesuré par la probabilité qu'une erreur de prédiction pour un label soit propagée à un autre :
  $REP = \dfrac{|\textit{err-label} \Rightarrow \textit{err-decision}|_{X'}}{|\textit{err-label} \Rightarrow \textit{err-decision}|_{X'} + |\textit{err-label} \Rightarrow \textit{not-err-decision}|_{X'}}$.
- l'utilité des relations entre les labels (Utility of Label Relations) mesurée par la probabilité qu'une prédiction correcte pour un attribut label conduit à une prédiction correcte pour l'attribut de décision:
  $ULR = \dfrac{|\textit{not-err-label} \Rightarrow \textit{not-err-decision}|_{X'}}{|\textit{not-err-label} \Rightarrow \textit{not-err-decision}|_{X'} + |\textit{not-err-label} \Rightarrow \textit{err-decision}|_{X'}}$.

Evaluation de la qualité de prédiction

La qualité de prédiction peut être évaluée pour chaque instance $x \in a_1 \times \ldots \times a_p$, ensuie la moyenne par rapport à un ensemble d'instances X' peut être calculée. Plusieurs mesures peuvent être utilisées pour évaluer une prédiction multi-labels (Tsoumakas et al.,2010,[108]):

- la mesure de l'erreur de hamming (Hamming-loss) (Destercke,2014,[29]) donnée par $\mathbb{HL} = \dfrac{|y \triangle H(x)|}{k}$, avec $y \triangle H(x)$ étant la différence symétrique entre l'ensemble correcte de labels associés et l'ensemble de labels prédit:

  $y \triangle H(x) = \{c_l \in y - H(x)\}_{1 \leq l \leq k} \bigcup \{c_l \in H(x) - y\}_{1 \leq l \leq k}$.

  L'erreur de Hamming évalue la proportion des erreurs entre les labels effectivement associés à l'instance et les labels prédits par rapport au nombre total de labels disponibles. Le nombre de labels disponibles représente le nombre maximal des erreurs possibles entre les labels effectivement associés à l'instance et les labels prédits. L'inconvénient de cette mesure est qu'elle est sensible à la cardinalité et à la densité des labels.

- le score de Hamming (closely related Hamming score) (Godbole and Sarawagi,2004,[42]) n'est pas sensible à la cardinalité et à la densité des labels. C'est une mesure qui évalue le nombre de labels prédits correctement par rapport au nombre de l'union des labels prédits et des labels effectivement associés à l'instance: $\mathbb{CRHS} = \dfrac{|y \cap H(x)|}{|y \cup H(x)|}$.

- la précision (precision) mesure la probabilité qu'un label prédit soit effectivement associé à l'instance:
  $\mathbb{PRECISION} = \frac{|y \cap H(x)|}{|H(x)|}$.

- le rappel (recall) mesure la probabilité qu'un label associé à l'instance soit prédit :
  $\mathbb{RECALL} = \frac{|y \cap H(x)|}{|y|}$.

- La mesure F_β (Pillai et al.,2017,[86]) combinant la précision et le rappel est donnée pour chaque $\beta > 0$ par :
  $\mathbb{F}_\beta = (1+\beta^2)\frac{\mathbb{PRECISION} \times \mathbb{RECALL}}{\beta^2 \times \mathbb{PRECISION} + \mathbb{RECALL}}$
  Plus d'importance est donnée à la précision pour les valeurs $\beta < 1$, et plus d'importance est donnée pour le rappel pour les valeurs $\beta > 1$. La même importance est donnée pour la précision et le rappel pour la valeur $\beta = 1$.

- la moyenne géométrique (GMEAN) (Kubat et al.,1997,[58]) est une mesure d'évaluation de la qualité de prédiction adaptée aux données avec un déséquilibre de classes. En effet, toutes les mesures précédentes favorisent un classifieur qui prédit la classe majoritaire en cas de déséquilibre de classes. La moyenne géométrique combine la précision positive donnée par $acc^+ = \frac{|\{c_l \in C, c_l \in y \ \& \ c_l \in H(x)\}|}{|y|} = \mathbb{RECALL}$, et la précision négative donnée par $acc^- = \frac{|\{c_l \in C,\ c_l \notin y \ \& \ c_l \notin H(x)\}|}{|C-y|}$ en une seule mesure donnée par : $\mathbb{GMEAN} = \sqrt{acc^+ \times acc^-}$.

- la correspondance exacte (exact match: $\mathbb{EM}$) est la mesure d'évaluation la plus stricte considérant la prédiction d'un ensemble de labels correcte seulement si l'ensemble prédit correspond exactement à l'ensemble de labels effectivement associé à l'instance: $\mathbb{EM}$ fournit la valeur 1 si $y = H(x)$, et la valeur 0 sinon.

1.2.3 Approches d'adaptation au cas multi-labels

L'adaptation des classifieurs mono-labels au cas multi-labels consiste en la modification de certaines étapes, paramètres, ou mesures de l'algorithme afin de tenir compte du problème de la multiplicité de labels. Parmi les méthodes qui ont bénéficié de cette adaptation se trouvent les séparateurs à vaste marge (Elisseeff and Weston,2001,[33] ; Sun et al.,2016,[102]), l'algorithme des k-plus proches voisins (Zhang and Zhou,2005,[131] ; Zhang and Zhou,2007,[132] ; Liu and Cao,2015,[69]), les classifieurs de bayes (Yan et al.,2016,[121]), les réseaux de neurones (Agrawal et al.,2016,[2]), et les arbres de décision (Amanda and King,2001,[6] ; Wang et al.,2015,[117]). L'inconvénient des approches d'adaptation est que pour modifier la stratégie d'apprentissage des relations entre les labels il faut modifier l'algorithme lui même. L'intérêt des approches de transformation au cas mono-label

est que l'algorithme de classification n'est pas modifié et seules les données d'apprentissage sont adaptées.

1.2.4 Approches de transformation au cas mono-label

Afin d'analyser les différentes approches de transformation au cas mono-label, nous nous basons sur l'exemple illustratif du Tableau 1.9. Nous considérons un jeu de données de 10 instances $X = \{x_i\}_{1 \leq i \leq 10}$, et un ensemble de trois labels disponibles $C = \{c_l\}_{1 \leq l \leq 3}$. Chaque instance $x_i \in X$ peut être associée à un ou plusieurs labels parmi les labels disponibles.

Instances	Labels
x_1	$\{c_1, c_2\}$
x_2	$\{c_1, c_2\}$
x_3	$\{c_1, c_2\}$
x_4	$\{c_1, c_2, c_3\}$
x_5	$\{c_1, c_2, c_3\}$
x_6	$\{c_1, c_3\}$
x_7	$\{c_2\}$
x_8	$\{c_2, c_3\}$
x_9	$\{c_3\}$
x_{10}	$\{c_3\}$

Tableau 1.9: Exemple de données multi-labels

1.2.4.1 Label Power set (LP)

L'approche LP consiste à considérer pour chaque instance l'ensemble de labels auxquels elle est associée en tant qu'un seul nouveau label. Le problème de la classification multi-labels devient donc une tâche de classification mono-label multi-classes. L'avantage de cette méthode est qu'elle fonctionne avec un seul classifieur multi-classes, par contre elle a l'inconvénient de ne pas prendre en compte certaines relations entre les labels.

Par exemple, si deux ensembles de labels non identiques contiennent des labels en commun, ils sont considérés comme totalement différents dans le problème transformé. De plus, cette transformation peut produire un déséquilibre de classes par le fait d'avoir plusieurs nouveaux labels peu récurrents (associés à très peu d'instances).

Le Tableau 1.10 illustre le résultat de la transformation de l'exemple du Tableau 1.9 en appliquant cette méthode. Le nouvel ensemble de labels produit $\mathscr{C}$ contient les nouveaux labels suivants:

- $\mathscr{C}_1 = \{c_1, c_2\}$
- $\mathscr{C}_2 = \{c_1, c_2, c_3\}$

- $\mathscr{C}_3 = \{c_1, c_3\}$
- $\mathscr{C}_4 = \{c_2\}$
- $\mathscr{C}_5 = \{c_2, c_3\}$
- $\mathscr{C}_6 = \{c_3\}$

Instances	Labels
x_1	$\mathscr{C}_1$
x_2	$\mathscr{C}_1$
x_3	$\mathscr{C}_1$
x_4	$\mathscr{C}_2$
x_5	$\mathscr{C}_2$
x_6	$\mathscr{C}_3$
x_7	$\mathscr{C}_4$
x_8	$\mathscr{C}_5$
x_9	$\mathscr{C}_6$
x_{10}	$\mathscr{C}_6$

Tableau 1.10: Transformation par la méthode LP

1.2.4.2 Pruned Problem Transformation (PPT)

L'approche PPT (Read,2008,[92]) propose une solution au problème de déséquilibre de classes que présente l'approche LP. L'idée est d'éliminer les combinaisons de labels peu récurrentes afin de ne garder que les combinaisons récurrentes. La fréquence seuil pour l'élimination est sélectionnée par l'utilisateur et elle est souvent égale à deux ou à trois.

L'inconvénient de cette approche est qu'elle est paramétrique, et elle mène à une perte d'information par l'élimination des labels peu récurrents qui peut influencer la qualité du classifieur appris.

Le Tableau 1.11 illustre le résultat de la transformation de l'exemple du Tableau 1.9 en appliquant cette méthode avec un seuil égal à deux.

1.2.4.3 Pruned Problem Transformation with no information loss (PPT-n)

L'approche PPT-n (Read,2008,[92]) reprend les mêmes idées que l'approche PPT mais au lieu de tronquer les combinaisons de labels peu fréquentes, PPT-n les divise en sous combinaisons plus fréquentes. Chaque instance est dupliquée autant de fois que de sous-combinaisons générées.

Le Tableau 1.12 illustre le résultat de la transformation de l'exemple du Tableau 1.9 en appliquant la méthode PPT-n avec un seuil égal à deux. La combinaison de labels $\{c_1, c_3\}$ est associée uniquement à l'instance x_6. L'instance x_6 est divisée en $x_{6'}$ associée à $\{c_1\}$ et $x_{6''}$ associée à $\{c_3\}$. La même stratégie est appliquée au cas de la combinaison $\{c_2, c_3\}$ associée uniquement à l'instance x_8.

Instances	Labels
x_1	$\{c_1, c_2\}$
x_2	$\{c_1, c_2\}$
x_3	$\{c_1, c_2\}$
x_4	$\{c_1, c_2, c_3\}$
x_5	$\{c_1, c_2, c_3\}$
x_9	$\{c_3\}$
x_{10}	$\{c_3\}$

Tableau 1.11: Transformation par la méthode PPT avec un seuil égal à deux

Instances	Labels
x_1	$\{c_1, c_2\}$
x_2	$\{c_1, c_2\}$
x_3	$\{c_1, c_2\}$
x_4	$\{c_1, c_2, c_3\}$
x_5	$\{c_1, c_2, c_3\}$
$x_{6'}$	$\{c_1\}$
$x_{6''}$	$\{c_3\}$
x_7	$\{c_2\}$
$x_{8'}$	$\{c_2\}$
$x_{8''}$	$\{c_3\}$
x_9	$\{c_3\}$
x_{10}	$\{c_3\}$

Tableau 1.12: Transformation par la méthode PPT-n avec un seuil égal à deux

1.2.4.4 Pruned Problem Transformation -Extended (PPT-ext)

L'inconvénient des approches LP, PPT, et PPT-n est qu'elles ne peuvent pas prédire une combinaison de labels qui n'est pas associée à une instance de l'ensemble d'apprentissage. L'approche PPT-ext (Read,2008,[92]) propose une extension permettant la prédiction de nouvelles combinaisons de labels. L'idée est de prédire une probabilité pour plusieurs combinaisons de labels. Ensuite chaque label est associé à la somme de probabilités prédites des combinaisons qui le contiennent. Les labels prédits au final sont ceux associés à une valeur supérieure à un seuil de probabilité souvent égal à 0.5.

L'inconvénient de cette méthode est qu'elle ajoute un second paramètre de probabilité pour les méthodes PPT et PPT-n qui utilisent déjà un paramètre seuil de fréquence.

1.2.4.5 Binary Relevance (BR)

Soit $\lambda_l : X \rightarrow \{0,1\}$ la fonction de pertinence du label c_l ($\lambda_l(x) = 1 \Leftrightarrow c_l \in y$).

L'approche BR consiste à construire un classifieur multi-labels H à partir d'un ensemble de k classifieurs mono-labels $\{H_l\}_{1 \leq l \leq k}$ (Tsoumakas and Katakis,2007,[107]). Chaque classifieur

Instances	λ_1
x_1	1
x_2	1
x_3	1
x_4	1
x_5	1
x_6	1
x_7	0
x_8	0
x_9	0
x_{10}	0

Données d'apprentissage pour H_1

Instances	λ_2
x_1	1
x_2	1
x_3	1
x_4	1
x_5	1
x_6	0
x_7	1
x_8	1
x_9	0
x_{10}	0

Données d'apprentissage pour H_2

Instances	λ_3
x_1	0
x_2	0
x_3	0
x_4	1
x_5	1
x_6	1
x_7	0
x_8	1
x_9	1
x_{10}	1

Données d'apprentissage pour H_3

Tableau 1.13: Transformation par la méthode BR

$H_l : a_1 \times \ldots \times a_p \rightarrow \{0,1\}$ apprend de l'ensemble X muni de la fonction de supervision λ_l à prédire la pertinence du label c_l pour toute donnée $x \in a_1 \times \ldots \times a_p$. Le classifieur multi-labels H est donné par : $H(x) = \{c_l, H_l(x) = 1\}_{1 \leq l \leq k}$.

Le Tableau 1.13 illustre le résultat de la transformation de l'exemple du Tableau 1.9 en appliquant l'approche BR. Cette approche est similaire à l'approche OVA (Section 1.1.3) qui permet de transformer un problème mono-label multi-classes en un problème mono-label binaire. La différence est que dans l'approche OVA chaque instance dans le problème initial est associée à exactement un seul label, alors que dans l'approche BR chaque instance dans le problème initial est associée à un ou plusieurs labels. Une autre différence est que dans l'approche OVA la prédiction correspond au label prédit avec la confiance la plus grande, alors que dans l'approche BR la prédiction correspond à la combinaison de tous les labels prédits par les classifieurs binaires: $H(x) = \{c_l, H_l(x) = 1\}_{1 \leq l \leq k}$.

Bien que l'approche BR nécessite k classifieurs, le temps d'exécution peut être optimisé car chaque classifieur peut être appris et fournir sa prédiction indépendamment des autres.

L'inconvénient de cette approche est qu'elle ne prend pas du tout en considération les relations entre les labels et suppose qu'ils sont indépendants.

Un autre inconvénient de cette approche est le risque de déséquilibre des classes. En effet, plusieurs combinaisons de classes apparaissent dans le problème initial, mais dans le problème transformé il n'y a que deux valeurs de classes: 0 et 1. Donc si une classe est incluse dans presque toutes les combinaisons (beaucoup plus de 1 que de 0), ou si elle n'est incluse que dans très peu de combinaisons (beaucoup plus de 0 que de 1), alors l'ensemble d'apprentissage correspondant à cette classe présentera un déséquilibre de classes.

1.2.4.6 Band-removal (BandSVM)

Puisque l'approche BR est basée sur la classification binaire d'une classe c_l contre toutes les autres classes, le cas le plus attendu de déséquilibre de classes est d'avoir beaucoup moins d'instances associée à c_l que d'instances associées aux autres classes $\{c_{l'}\}_{l' \neq l}$. L'approche BandSVM (Tsoumakas et al.,2010,[109]) permet de réduire l'effet du déséquilibre de classes en supprimant certaines instances non associées à c_l. Cette approche est faite pour fonctionner avec les classifieurs de type SVM à l'origine, mais l'idée peut être adaptée à d'autres types de classifieurs.

L'approche BandSVM s'effectue en deux étapes: la première est d'apprendre pour chaque label c_l un classifieur binaire H_l de type SVM pour prédire la pertinence de c_l. H_l est donc un hyperplan séparant les instances associées à c_l des instances non associées à c_l. La deuxième étape est de refaire l'apprentissage de chaque classifieur après la suppression des instances non associées à c_l se situant à une distance inférieure à un seuil par rapport à l'hyperplan trouvé par le classifieur H_l.

L'avantage de cette approche est qu'elle permet de gérer le problème du déséquilibre de classes. Cependant, elle a l'inconvénient de nécessiter l'apprentissage de $2k$ classifieurs et de ne pas prendre en compte les relations entre les labels.

1.2.4.7 Classifier Chains (CC)

L'approche CC est une extension de l'aproche BR permettant l'apprentissage des relations entre les labels en introduisant un ensemble d'attributs descriptifs supplémentaires $B = \{b_l\}_{1 \leq l \leq k}$. Chaque instance $x_i \in X$ est étendue telle que $x_i^e = (x_{i,a_1}, \ldots, x_{i,a_p}, \lambda_1(x_i), \ldots, \lambda_k(x_i))$. L'ensemble d'apprentissage X_l de chaque classifieur H_l est construit par une projection de l'ensemble d'apprentissage étendu $X^e = \{x_i^e\}_{1 \leq i \leq n}$ sur l'espace d'attributs descriptifs $A \cup \{b_{l'}\}_{1 \leq l' < l}$. Les attributs $\{b_{l'}\}_{l' \geq l}$ sont donc ignorés par le classifieur H_l.
Le Tableau 1.14 illustre le résultat de la transformation de l'exemple du Tableau 1.9 en appliquant l'approche CC.
Le classifieur $H_2 : a_1 \times \ldots \times a_p \times b_1 \rightarrow \{0,1\}$ ne peut pas fournir directement une prédiction pour une instance $x \in a_1 \times \ldots \times a_p$. En effet, la valeur de l'attribut b_1 est inconnue pour les instances qui ne font pas partie de l'ensemble d'apprentissage. L'instance x est donc d'abord étendue par la prédiction du classifieur H_1: $x = (x_{a_1}, \ldots, x_{a_p}, H_1(x_i))$ avant d'être reçue par le classifieur H_2. Chaque classifieur H_l a donc la possibilité de fournir des prédictions en se basant sur les prédictions des autres classifieurs qui le précèdent: $\{H_{l'}\}_{1 \leq l' < l \leq k}$. L'inconvénient de l'approche CC est que les relations de co-occurrences qui peuvent être apprises dépendent de l'ordre initial des labels. Par conséquent, la prédiction d'un label c_l ne peut pas dépendre d'une relation de co-occurrences avec les labels $c_{l'}, l' > l$.

Instances	λ_1
x_1	1
x_2	1
x_3	1
x_4	1
x_5	1
x_6	1
x_7	0
x_8	0
x_9	0
x_{10}	0

Données d'apprentissage pour H_1

Instances	b_1	λ_2
x_1	1	1
x_2	1	1
x_3	1	1
x_4	1	1
x_5	1	1
x_6	1	0
x_7	0	1
x_8	0	1
x_9	0	0
x_{10}	0	0

Données d'apprentissage pour H_2

Instances	b_1	b_2	λ_3
x_1	1	1	0
x_2	1	1	0
x_3	1	1	0
x_4	1	1	1
x_5	1	1	1
x_6	1	0	1
x_7	0	1	0
x_8	0	1	1
x_9	0	0	1
x_{10}	0	0	1

Données d'apprentissage pour H_3

Tableau 1.14: Transformation par la méthode CC

1.2.4.8 Aggregating Independent and Dependent Models (AID)

L'approche AID permet d'apprendre les relations entre les labels sans dépendre de l'ordre initial des labels en se basant sur deux ensembles de classifieurs (Montañés et al.,2011,[79]). Le premier ensemble de classifieurs $\{h_l\}_{1\leq l\leq k}$ est construit par l'approche BR (chaque classifieur h_l est indépendant des autres classifieurs). Le deuxième ensemble de classifieurs $\{H_l\}_{1\leq l\leq k}$ est construit de façon similaire à l'approche CC. La différence est que l'ensemble d'apprentissage X_l pour le classifieur H_l est construit en projetant l'ensemble étendu X^e sur tous les attributs initiaux et supplémentaires sauf l'attribut b_l à prédire: $A \cup B - \{b_l\}$. Ceci permet au classifieur H_l d'établir sa prédiction en se basant sur la pertinence de tous les autres labels $c_{l'}, l' \neq l$. Chaque instance donnée $x \in a_1 \times \ldots \times a_p$ est étendue par les prédictions du premier ensemble de classifieurs $x^e = (x_{a_1}, \ldots, x_{a_p}, h_1(x), \ldots, h_{l-1}(x), h_{l+1}(x), \ldots, h_k(x))$ avant d'être fournie en entrée au classifieur H_l. Chaque classifieur H_l fournie sa prédiction en se basant sur les prédictions initiales $\{h_{l'}(x)\}_{l'\neq l}$ et en ignorant les prédictions finales des autres classifieurs $\{H_{l'}(x)\}_{l'\neq l}$. La prédiction du classifieur multi-labels H est donnée par : $H(x) = \{c_l, H_l(x) = 1\}_{1\leq l\leq k}$.
L'approche AID possède deux inconvénients remarquables :

- elle nécessite l'apprentissage de $2k$ classifieurs.

- l'ensemble de labels prédit finalement n'est pas nécessairement en accord avec les relations apprises contrairement à l'approche CC. En effet, chaque label prédit finalement est en accord avec les relations apprises uniquement par rapport aux prédictions initiales $\{h_{l'}(x)\}_{l'\neq l}$ qui sont remplacées par les prédictions finales $\{H_{l'}\}_{l'\neq l}$.

1.2.4.9 Binary relevance with label dependence (BR+)

L'approche BR+ (Montañés et al.,2014,[80]) est une amélioration de l'approche AID permettant d'obtenir des prédictions tenant en compte les dépendances entre les labels. L'idée est d'obtenir d'abord les prédictions indépendantes de l'approche AID $\{h_l(x)\}_{1\leq l\leq k}$, ensuite un ordre de labels est établi par rapport à la fréquence de présence des labels ou par rapport à la confiance des prédictions obtenues par les classifieurs $\{h_l(x)\}_{1\leq l\leq k}$.

Afin de simplifier la notation, nous considérons que l'ordre des labels établi est le même ordre initial. Les classifieurs dépendants $\{H_l(x_i)\}_{1\leq l\leq k}$ sont utilisés pour mettre à jour les prédictions des classifieurs $\{h_l(x)\}_{1\leq l\leq k}$. Chaque classifieur H_l fournit sa prédiction en se basant sur $l-1$ prédictions des classifieurs dépendants $\{H_{l'}\}_{l'<l}$ et sur $k-l$ prédictions des classifieurs indépendants $\{H_{l'}\}_{l'>l}$. L'approche BR+ permet donc de produire des prédictions tenant en compte les dépendances entre les labels mais seulement d'une façon partielle. Le seul classifieur dont la prédiction respecte entièrement les dépendances entre les labels est le dernier selon l'ordre établi H_k.

L'inconvénient de toutes les approches basées sur la prédiction de la pertinence de chaque label (BR, CC, AID) est que seules les relations de dépendance peuvent être apprises. Les relations exprimant une préférence entre les labels sont donc ignorées par ce type d'approches.

1.2.4.10 Ranking by Pairwise Comparison (RPC)

L'approche 'Ranking by Pairwise Comparisons' (RPC) est basée sur $\frac{k(k-1)}{2}$ classifieurs $\{H_{l,l'}\}_{1\leq l<l'\leq k}$ permettant la prédiction d'une préférence entre chaque deux labels $(c_l, c_{l'})$ (Hüllermeier et al.,2008,[50]).

L'ensemble d'apprentissage $X_{l,l'}$ du classifieur $H_{l,l'}$ est le sous-ensemble de X contenant uniquement les instances associées exclusivement à l'un des deux labels c_l ou $c_{l'}$: $X_{l,l'} = \{x_i \in X, (c_l \in y_i \ et \ c_{l'} \notin y_i) \ ou \ (c_l \notin y_i \ et \ c_{l'} \in y_i)\}$.
Soit $\lambda_{l,l'} : X_{l,l'} \rightarrow \{c_l, c_{l'}, \emptyset\}$ la fonction donnée par :

- $\lambda_{l,l'}(x_i) = c_l$ si $(c_l \in y_i \ et \ c_{l'} \notin y_i)$
- $\lambda_{l,l'}(x_i) = c_{l'}$ si $(c_l \notin y_i \ et \ c_{l'} \in y_i)$

La fonction $\lambda_{l,l'}$ fournit le label préféré entre c_l et $c_{l'}$ pour les instances de l'ensemble d'apprentissage, et constitue une fonction de supervision pour le classifieur $H_{l,l'}$. Chaque classifieur $H_{l,l'}$ apprend à partir de l'ensemble supervisé $(X_{l,l'}, \lambda_{l,l'})$ à prédire le label préféré entre c_l et $c_{l'}$ pour toute instance $x \in a_1 \times \ldots \times a_p$.
Soit $V_{c_l} : a_1 \times \ldots \times a_p \rightarrow [\![0, k-1]\!]$ la fonction qui fournit pour le label $c_l \in C$ le nombre de fois qu'il a été préféré pour une instance x (le nombre de votes pour le label c_l): $V_{c_l}(x) = |\{(c_{l'}, c_{l''}), H_{l',l''}(x) = c_l\}_{1\leq l'<l''\leq k}|$.
L'approche RPC ne fixe pas une méthode de prédiction et permet juste d'ordonner les labels selon le

nombre de fois où ils ont été préférés par les classifieurs $\{H_{l,l'}\}_{1\leq l<l'\leq k}$. Il est possible, par exemple, de prédire les labels dont le nombre de votes est supérieur à un seuil fixé v. Le classifieur multi-labels H dans ce cas est donné par : $H(x) = \{c_l \in C, V_{c_l}(x) \geq v\}$.

Le Tableau 1.15 illustre le résultat de la transformation de l'exemple du Tableau 1.9 en appliquant l'approche RPC. Cette approche est similaire à l'approche OVO (Section 1.1.3) permettant de transformer un problème mono-label multi-classes en un problème mono-label binaire. La différence est que dans l'approche OVO chaque instance dans le problème initial est associée à exactement un seul label, alors que dans l'approche RPC chaque instance dans le problème initial peut être associée à un ou plusieurs labels. Une autre différence est que dans l'approche OVO la prédiction correspond au label sélectionné par un vote majoritaire, alors que dans l'approche RPC les prédictions des classifieurs binaires permettent seulement d'obtenir une liste ordonnée de tous les labels selon le nombre de fois qu'ils ont été préférés.

L'approche RPC permet d'apprendre des préférences entre les labels deux à deux. Cependant, le fait qu'un label est préféré ne permet pas de conclure que l'instance est associée à ce label. En effet, les labels c_l et $c_{l'}$ peuvent être préférés le plus de fois mais être en même temps incompatibles. L'approche RPC a l'inconvénient de ne pas apprendre les dépendances entre les labels permettant de tenir compte de ce cas.

Instances	$\lambda_{1,2}$
x_6	c_1
x_7	c_2
x_8	c_1

Données d'apprentissage du classifieur $H_{1,2}$

Instances	$\lambda_{1,3}$
x_1	c_1
x_2	c_1
x_3	c_1
x_8	c_3
x_9	c_3
x_{10}	c_3

Données d'apprentissage du classifieur $H_{1,3}$

Instances	$\lambda_{2,3}$
x_1	c_2
x_2	c_2
x_3	c_2
x_6	c_3
x_7	c_2
x_9	c_3
x_{10}	c_3

Données d'apprentissage du classifieur $H_{2,3}$

Tableau 1.15: Transformation par la méthode RPC

1.2.4.11 Calibrated Label Ranking (CLR)

L'approche CLR est une extension de l'approche RPC qui permet de sélectionner les labels à prédire en utilisant un label virtuel au lieu d'un paramètre seuil pour le nombre de votes (Fürnkranz et al.,2008,[37]). L'approche CLR introduit un label virtuel c_0 et apprend k classifieurs de plus $\{H_{l,0}\}_{1\leq l\leq k}$ par rapport à l'approche RPC. La fonction de supervision correspondant à un classifieur $H_{l,0}$ est donnée par :

- $\lambda_{l,0}(x_i) = c_l$ si $c_l \in y_i$.

- $\lambda_{l,0}(x_i) = c_0$ sinon.

Le classifieur multi-labels H prédit tous les labels qui reçoivent plus de votes que le label virtuel: $H(x) = \{c_l \in C, V_{c_l}(x) \geq V_{c_0}(x)\}$.

L'approche CLR résout le problème de séparation entre les labels à prédire et à ne pas prédire dans la liste ordonnée de labels obtenue par l'approche RPC. Par contre, elle ne résout pas le problème du risque d'incompatibilité entre les labels prédits.

1.2.5 Conclusion

Les approches de transformation du cas multi-labels au cas mono-label peuvent être regroupées en deux catégories (Figure 1.7): les approches de transformation en classification mono-label multi-classes (LP, PPT, PPT-n, et PPT-ext), et les approches de transformation en classification mono-label binaire, qui eux même peuvent être regroupées en deux catégories: Celles utilisant un classifieur binaire pour prédire la pertinence de chaque label (BR, BandSVM, CC, AID, BR+), et celles utilisant un classifieur binaire pour prédire la préférence entre chaque deux labels distincts (RPC, CLR).

Certaines approches de transformation au cas mono-label ne prennent pas du tout en compte les relations entre les labels (BR). D'autres approches ne considèrent que les relations entre des combinaisons de labels et ignorent les relations entre les sous-ensembles de ces combinaisons de labels (LP). Certaines méthodes permettent l'apprentissage des relations de dépendance entre les labels mais seulement par rapport à une seule structure de dépendance fixée au préalable (CC). D'autre méthodes apprennent uniquement des relations de préférence entre chaque deux labels distincts en ignorant les relations de dépendance entre les labels (RPC, CLR). D'autres approches de l'état d'art nécessitent l'apprentissage de plusieurs couches de classifieurs, et ne permettent de tenir en compte les dépendance entre les labels que d'une façon partielle (AID, BR+).

L'un des objectifs de ce travail est de proposer une nouvelle approche de transformation permettant de remédier aux limites des approches de l'état de l'art:

- permettre l'apprentissage de plusieurs structures de dépendance entre les labels avec une complexité minimale.
- tenir en compte les deux types de relations entre les labels: les relations de dépendance et les relations de préférence.

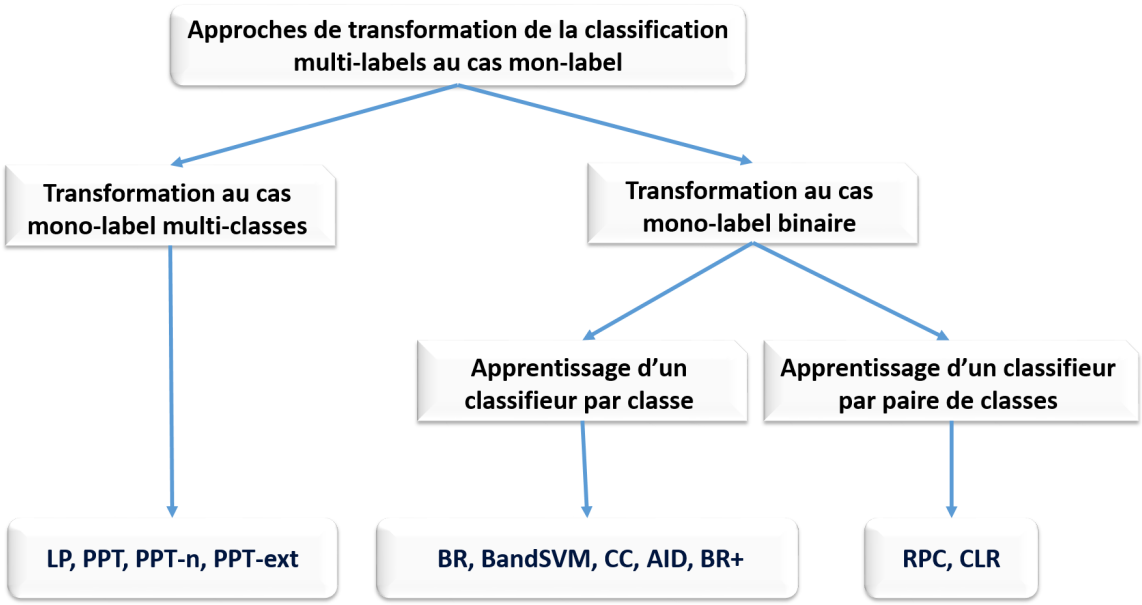


Figure 1.7: Approches de transformations au cas mono-label

1.3 Classification multi-labels graduée

En classification multi-labels graduée chaque instance peut être associée à un ou plusieurs labels. L'association d'une instance à chaque label est effectuée avec un degré appartenant à un ensemble ordonné de degrés d'association.

Le Tableau 1.16 illustre un jeu de données de molécules odorantes où chaque molécule peut émettre jusqu'à 7 odeurs ordonnées selon leurs intensités (Arctander,1969,[8]). Il s'agit du même exemple de données multi-labels dans le Tableau 1.8 mais avec l'information supplémentaire sur les degrés d'association. L'odeur d'intensité maximale est associée à la molécule avec le degré d'association 7. L'odeur d'intensité minimale est associée à la molécule avec le degré d'association 1. Les odeurs qui ne sont pas émises par la molécule ont le degré d'association 0. Certaines odeurs associées aux molécules dans le Tableau 1.16 ont été éliminées pour ne garder que les odeurs {HERBAC, FRUITY, JASMIN} afin de simplifier l'exemple.

Les figures 1.8, 1.9, et 1.10 illustrent les arbres de décision permettant de prédire respectivement le degré d'association de l'odeur 'HERBAC', de l'odeur 'FRUITY', et de l'odeur 'JASMIN'.

La classification multi-labels graduée se distingue par le fait que les dépendances entre les labels ne sont pas que des relations de co-occurrence, mais aussi des relations d'ordre impliquant les degrés d'association des labels.

Par exemple l'arbre de décision de la Figure 1.9 permet d'extraire la règle suivante:

si 'Structure' = 'C' et 'degré d'association au label HERBAC' ≤ 4 et 'Masse moléculaire' > 156.25 alors 'degré d'association au label FRUITY' = 5.

L'un des objectifs d'étudier l'état de l'art des approches de classification multi-labels graduée est de comparer leurs complexités, et leurs capacités à apprendre les différents types de relations entre les labels. Cette étude permet de mettre en avant les avantages et les inconvénients de chaque approche afin de proposer des améliorations possibles ou de nouvelles approches plus performantes.

Formule	Structure	Masse moléculaire	Ensemble gradué d'odeurs	Degré HERBAC	Degré FRUITY	Degré JASMIN
$C_{12}H_{18}O$	A	266.47	{HERBAC | 6 , JASMIN | 5}	6	0	5
$C_{10}H_{18}O$	L	154.28	{FRUITY | 7 , JASMIN | 5}	0	7	5
$C_9H_{16}O_2$	C	156.25	{FRUITY | 7 , JASMIN | 5}	0	7	5
$C_{12}H_{16}O_2$	A	192.28	{HERBAC | 5 , FRUITY | 6 , JASMIN | 4}	5	6	4
$C_{12}H_{22}O$	C	182.34	{HERBAC | 4 , FRUITY | 5 , JASMIN | 3}	4	5	3
$C_{14}H_{14}O$	A	306.43	{JASMIN | 6}	0	0	6
$C_{11}H_{14}O_3$	A	194.25	{JASMIN | 6}	0	0	6
$C_{18}H_{32}O_2$	L	280.5	{HERBAC | 6 , FRUITY | 7}	6	7	0
$C_{14}H_{28}O_2$	L	228.42	{HERBAC | 6 , FRUITY | 7}	6	7	0
$C_9H_{12}O_2$	C	152.21	{HERBAC | 7}	7	0	0
$C_{10}H_{16}O_2$	C	168.26	{HERBAC | 7}	7	0	0

Tableau 1.16: Données d'apprentissage pour la prédiction de sous-ensembles gradués d'odeurs

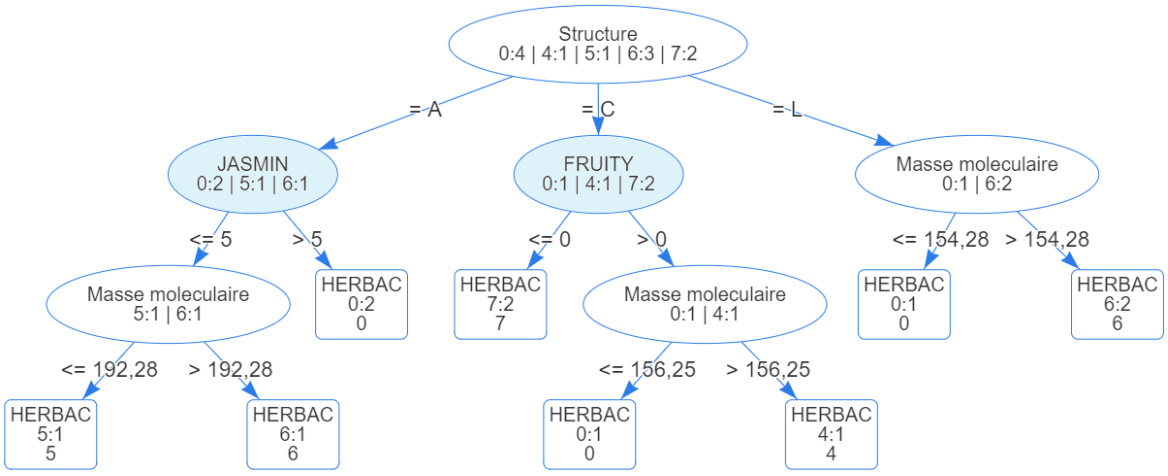


Figure 1.8: Arbre de décision pour prédire le degré d'association de l'odeur 'HERBAC' à partir des données du Tableau 1.16

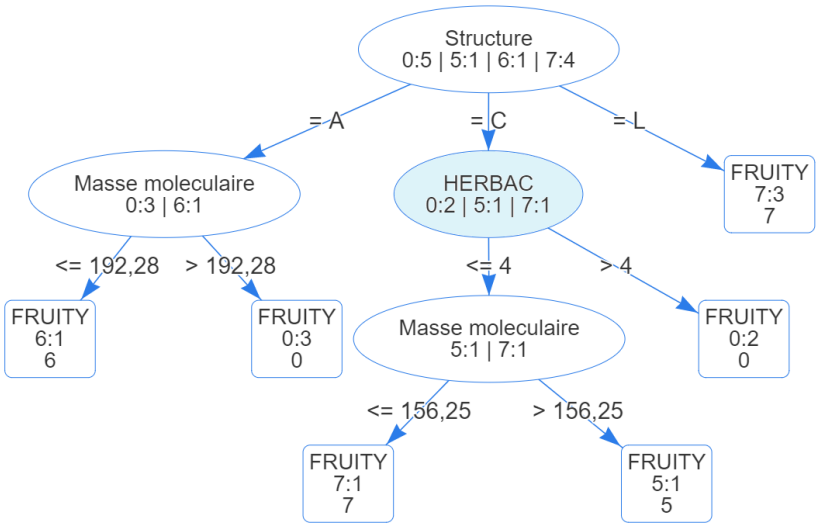


Figure 1.9: Arbre de décision pour prédire le degré d'association de l'odeur 'FRUITY' à partir des données du Tableau 1.16

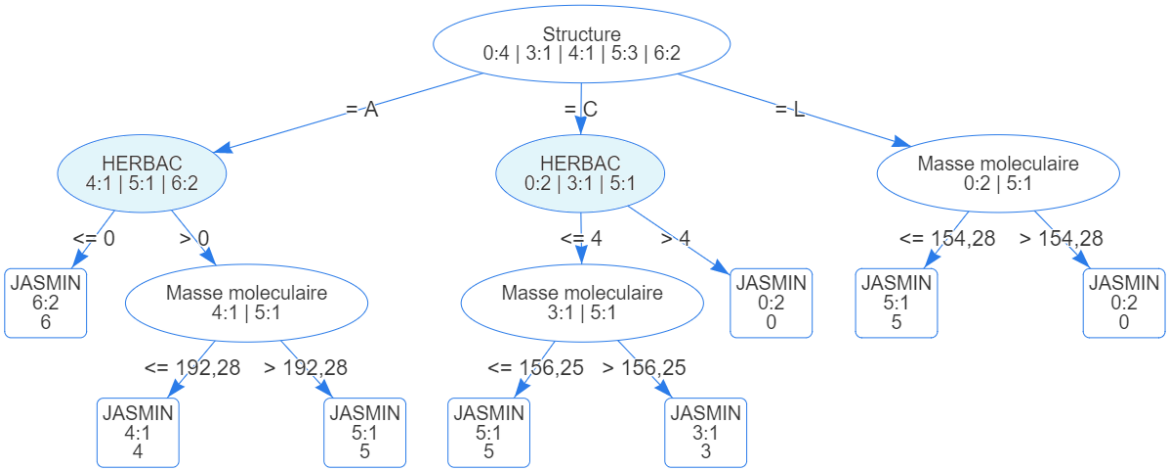


Figure 1.10: Arbre de décision pour prédire le degré d'association de l'odeur 'JASMIN' à partir des données du Tableau 1.16

1.3.1 Description formelle du problème de la classification multi-labels graduée

Soit $A = \{a_j\}_{1 \leq j \leq p}$ un ensemble d'attributs descriptifs. Soit $X = \{x_i\}_{1 \leq i \leq n}$ un ensemble d'instances. Chaque instance est décrite par un vecteur de valeurs d'attributs descriptifs $x_i = (x_{ij})_{1 \leq j \leq p}$.

Soit $C = \{c_l\}_{1 \leq l \leq k}$ un ensemble de labels. Soit $M = \{m_g\}_{1 \leq g \leq s}$ un ensemble ordonné de degrés d'appartenance tel que $m_1 < \ldots < m_s$.

L'ensemble de toutes les pairs (label | degré d'appartenance) est noté $C|M$: $C|M = \{c_l|m_g\}_{\substack{1 \leq l \leq k \\ 1 \leq g \leq s}}$. L'ensemble de tous les sous-ensembles de pairs (label | degré d'appartenance) est noté $\mathscr{P}(C|M)$.

Chaque instance $x_i \in X$ est associée à un sous-ensemble gradué de labels $y_i \in \mathscr{P}(C|M)$. Les labels ayant un degré d'association nul à une instance x_i peuvent être omis de l'ensemble gradué de labels y_i.

Soit $\lambda : X \rightarrow \mathscr{P}(C|M)$ la fonction qui associe à chaque instance $x_i \in X$ le sous-ensemble gradué de labels correspondant $\lambda(x_i) = y_i$.

Soit $E : \mathscr{P}(C|M) \times \mathscr{P}(C|M) \rightarrow [0,1]$ une fonction objectif à optimiser.

La classification multi-labels graduée consiste à apprendre à partir de l'ensemble supervisé (X, λ) un classifieur $H : a_1 \times \ldots \times a_p \rightarrow \mathscr{P}(C|M)$ permettant de prédire pour chaque instance $x \in a_1 \times \ldots \times a_p$ un ensemble gradué de labels $H(x) \in \mathscr{P}(C|M)$ optimisant l'évaluation de la fonction objectif $E(y, H(x))$, où y étant l'ensemble gradué de labels effectivement associé à l'instance x.

1.3.2 Interaction avec les autres types de classification

La classification floue où les degrés d'association sont des valeurs réelles dans l'intervalle $[0,1]$ (Prati,2015,[87] ; Nápoles et al.,2016,[81] ; Gaied et al.,2017,[38]) est une généralisation de la classification multi-labels graduée (Cheng et al.,2010,[25] ; Brinker et al.,2014,[21] ; Lastra et al.,2014,[68]). Cependant, l'intérêt de la classification multi-labels graduée est qu'en pratique il est plus simple pour un humain d'associer aux instances des labels avec des degrés d'association graduels.

La gradualité dans la classification multi-labels graduée est une généralisation de la classification multi-labels où l'association aux labels est binaire: $M = \{0,1\}$ (Section 1.2).

La multiplicité de labels dans la classification multi-labels graduée est une généralisation de la classification mono-label ordinale où chaque instance ne peut être associée qu'à un seul label en même temps (Section 1.1.3).

La tâche de classification multi-labels graduée peut être décomposée en plusieurs tâches de classification multi-labels et de classification ordinale. Les approches de classification multi-labels graduée se distinguent par rapport à la stratégie de décomposition et par rapport à la stratégie d'agrégation des sous-tâches générées.

1.3.3 Décomposition du problème de classification multi-labels graduée

Afin d'illustrer les différentes approches de décomposition du problème de la classification multi-labels graduée, nous considérons l'exemple illustratif du Tableau 1.17. L'exemple est constitué de 12 instances $X = \{x_i\}_{1 \leq i \leq 12}$. Chaque instance peut être associée à trois labels différents c_1, c_2, c_3 selon une échelle graduelle de quatre degrés d'appartenance $m_1 < m_2 < m_3 < m_4$.

Instances	Ensemble gradué de labels
x_1	$\{c_1\|m_1, c_2\|m_2, c_3\|m_3\}$
x_2	$\{c_1\|m_1, c_2\|m_2, c_3\|m_3\}$
x_3	$\{c_1\|m_3, c_2\|m_1, c_3\|m_2\}$
x_4	$\{c_1\|m_3, c_2\|m_1, c_3\|m_2\}$
x_5	$\{c_1\|m_2, c_2\|m_3, c_3\|m_1\}$
x_6	$\{c_1\|m_2, c_2\|m_3, c_3\|m_1\}$
x_7	$\{c_1\|m_4, c_2\|m_1, c_3\|m_1\}$
x_8	$\{c_1\|m_4, c_2\|m_1, c_3\|m_1\}$
x_9	$\{c_1\|m_2, c_2\|m_4, c_3\|m_2\}$
x_{10}	$\{c_1\|m_2, c_2\|m_4, c_3\|m_2\}$
x_{11}	$\{c_1\|m_3, c_2\|m_3, c_3\|m_4\}$
x_{12}	$\{c_1\|m_3, c_2\|m_3, c_3\|m_4\}$

Tableau 1.17: Exemple de données multi-labels graduées

1.3.3.1 Décomposition verticale

Soit $\mu_{c_l} : X \rightarrow M$ la fonction qui donne pour chaque instance $x_i \in X$ son degré d'association au label c_l.

Une façon directe de décomposer le problème de la classification multi-labels graduée est de générer une tâche de classification ordinale pour chaque label. Dans ce cas, pour chaque sous-tâche générée, les degrés d'appartenance sont considérés en tant que classes ayant un ordre (Section 1.1.3). Cette décomposition appelée 'décomposition verticale' (Cheng et al.,2010,[25]) permet d'apprendre un classifieur multi-labels gradué H à partir d'un ensemble de classifieurs ordinaux $\{H_{c_l}\}_{1 \leq l \leq k}$. Chaque classifieur $H_{c_l} : a_1 \times \ldots \times a_p \rightarrow M$ apprend à partir de l'ensemble supervisé (X, μ_{c_l}) à prédire le degré d'association du label c_l à toute instance $x \in a_1 \times \ldots \times a_p$. La prédiction multi-labels graduée pour une instance x est donnée par : $H(x) = \{c_l|H_{c_l}(x)\}_{1 \leq l \leq k}$.

Le Tableau 1.18 illustre la décomposition verticale appliquée au jeu de données du Tableau 1.17.

Instances	μ_{c_1}
x_1	m_1
x_2	m_1
x_3	m_3
x_4	m_3
x_5	m_2
x_6	m_2
x_7	m_4
x_8	m_4
x_9	m_2
x_{10}	m_2
x_{11}	m_3
x_{12}	m_3

Données pour H_{c_1}

Instances	μ_{c_2}
x_1	m_2
x_2	m_2
x_3	m_1
x_4	m_1
x_5	m_3
x_6	m_3
x_7	m_1
x_8	m_1
x_9	m_4
x_{10}	m_4
x_{11}	m_3
x_{12}	m_3

Données pour H_{c_2}

Instances	μ_{c_3}
x_1	m_3
x_2	m_3
x_3	m_2
x_4	m_2
x_5	m_1
x_6	m_1
x_7	m_1
x_8	m_1
x_9	m_2
x_{10}	m_2
x_{11}	m_4
x_{12}	m_4

Données pour H_{c_3}

Tableau 1.18: Décomposition verticale des données multi-labels graduées du Tableau 1.17

1.3.3.2 Décomposition horizontale

La décomposition horizontale du problème de la classification multi-labels graduée utilise la notion d'alpha-coupes des sous-ensembles flous (Zadeh,1965,[125]). Il s'agit de générer $s-1$ tâches de classification multi-labels, une pour chaque degré d'appartenance $m_g, g \in [2,s]$. Le classifieur multi-labels correspondant au degré d'appartenance m_1 n'a pas besoin d'une étape d'apprentissage et peut être construit à partir des autres classifieurs.

Soit $\lambda_{\geq m_g} : X \rightarrow \mathscr{P}(C)$ la fonction qui donne pour chaque instance $x_i \in X$ l'ensemble de labels associés avec un degré d'association $m_{g'}$ supérieur ou égal au degré m_g ($m_{g'} \geq m_g$).

La décomposition horizontale permet d'apprendre un classifieur multi-labels gradué H en se basant sur $s-1$ classifieurs multi-labels $\{H_{\geq m_g}\}_{2\leq g\leq s}$. Chaque classifieur $H_{\geq m_g} : a_1 \times \ldots \times a_p \rightarrow \mathscr{P}(C)$ apprend à partir de l'ensemble supervisé $(X, \lambda_{\geq m_g})$ à prédire pour toute instance $x \in a_1 \times \ldots \times a_p$ l'ensemble de labels ayant un degré d'association supérieur ou égal au degré m_g.

Soit $h_{c_l} : a_1 \times \ldots \times a_p \rightarrow M$ une fonction permettant d'agréger les prédictions des classifieurs $\{H_{\geq m_g}\}_{2\leq g\leq s}$ afin de prédire le degré d'association du label c_l. La fonction d'agrégation h_l permettant de prédire le plus grand degré d'association possible pour le label c_l est donnée pour toute instance $x \in a_1 \times \ldots \times a_p$ par : $h_{c_l}(x) = m_g$ avec:

- $c_l \in H_{\geq m_g}(x)$.
- $\forall m_{g'} > m_g$: $c_l \notin H_{\geq m_{g'}}(x)$

Le classifieur multi-labels gradué H est donné pour toute instance $x \in a_1 \times \ldots \times a_p$ par : $H(x) = \{c_l | h_{c_l}(x)\}_{1\leq l\leq k}$.

Le Tableau 1.19 illustre la décomposition horizontale appliquée au jeu de données du Tableau 1.17.

L'inconvénient de la décomposition horizontale est que la monotonie: $\mu_{c_l}(x) \geq m_g \Rightarrow \forall\, m_{g'} < m_g : \mu_{c_l}(x) \geq m_{g'}$ est triviale pour les données d'apprentissage mais n'est pas nécessairement respectée en prédiction.

En effet, les classifieurs peuvent se tromper en prédiction et donc on peut avoir $c_l \in H_{\geq m_g}(x)$, $c_l \notin H_{\geq m_{g'}}(x)$, et $m_{g'} < m_g$. Ceci est équivalent à prédire que le degré d'appartenance à c_l est à la fois supérieur ou égal à m_g, et en même temps inférieur strictement à $m_{g'}$ alors que $m_{g'} < m_g$.

Instances	$\lambda_{\geq m_2}$
x_1	$\{c_2, c_3\}$
x_2	$\{c_2, c_3\}$
x_3	$\{c_1, c_3\}$
x_4	$\{c_1, c_3\}$
x_5	$\{c_1, c_2\}$
x_6	$\{c_1, c_2\}$
x_7	$\{c_1\}$
x_8	$\{c_1\}$
x_9	$\{c_1, c_2, c_3\}$
x_{10}	$\{c_1, c_2, c_3\}$
x_{11}	$\{c_1, c_2, c_3\}$
x_{12}	$\{c_1, c_2, c_3\}$

Données pour $H_{\geq m_2}$

Instances	$\lambda_{\geq m_3}$
x_1	$\{c_3\}$
x_2	$\{c_3\}$
x_3	$\{c_1\}$
x_4	$\{c_1\}$
x_5	$\{c_2\}$
x_6	$\{c_2\}$
x_7	$\{c_1\}$
x_8	$\{c_1\}$
x_9	$\{c_2\}$
x_{10}	$\{c_2\}$
x_{11}	$\{c_1, c_2, c_3\}$
x_{12}	$\{c_1, c_2, c_3\}$

Données pour $H_{\geq m_3}$

Instances	$\lambda_{\geq m_4}$
x_1	$\{\}$
x_2	$\{\}$
x_3	$\{\}$
x_4	$\{\}$
x_5	$\{\}$
x_6	$\{\}$
x_7	$\{c_1\}$
x_8	$\{c_1\}$
x_9	$\{c_2\}$
x_{10}	$\{c_2\}$
x_{11}	$\{c_3\}$
x_{12}	$\{c_3\}$

Données pour $H_{\geq m_4}$

Tableau 1.19: Décomposition horizontale des données multi-labels graduées du Tableau 1.17

1.3.3.3 Décomposition Complète

La décomposition de la classification multi-labels graduée est dite décomposition verticale si elle est faite par rapport aux labels (Section 1.3.3.1), et elle est dite décomposition horizontale si elle faite par rapport aux degrés d'appartenance (Section 1.3.3.2). La décomposition est dite complète lorsqu'elle est faite à la fois par rapport aux labels et par rapport aux degrés d'association.

Soit $\eta_{\geq m_g, c_l} : X \rightarrow \{0,1\}$ la fonction qui associe à chaque instance $x_i \in X$ la valeur 1 si le label c_l est associé à l'instance x_i avec un degré d'association $m_{g'} \geq m_g$, et la valeur 0 sinon.

La décomposition complète permet de construire un classifieur multi-labels gradué H à partir de $k\ (s\ -\ 1)$ classifieurs binaires $\{H_{\geq m_g, c_l}\}_{\substack{2 \leq g \leq s \\ 1 \leq l \leq k}}$. Chaque classifieur $H_{\geq m_g, c_l} : a_1 \times \ldots \times a_p \rightarrow \{0,1\}$ apprend à partir de l'ensemble supervisé $(X, \eta_{\geq m_g, c_l})$ à prédire pour toute instance $x \in a_1 \times \ldots \times a_p$ si le label c_l est associé avec un degré supérieur ou égal à m_g ($H_{\geq m_g, c_l}(x) = 1$) ou non ($H_{\geq m_g, c_l}(x) = 0$).

Soit $h_{c_l} : a_1 \times \ldots \times a_p \rightarrow \{0,1\}$ une fonction permettant de prédire le degré d'association du label c_l pour toute instance x en agrégeant les prédictions des classifieurs $\{H_{\geq m_g, c_l}\}_{2 \leq g \leq s}$. La fonction h_{c_l} peut être donnée pour toute instance x par : $h_{c_l}(x) = m_g$ avec:

- $H_{\geq m_g,c_l}(x) = 1$.
- $\forall m_{g'} > m_g$: $H_{\geq m_{g'},c_l}(x) = 0$.

Le classifieur multi-labels gradué H peut être donné pour toute instance x par : $H(x) = \{c_l | h_{c_l}(x)\}_{1\leq l\leq k}$.

Le Tableau 1.20 illustre la décomposition complète appliquée au jeu de données du Tableau 1.17.

La décomposition complète peut être obtenue par deux autres méthodes:

- en appliquant la décomposition verticale, puis chaque classifieur ordinal H_{c_l} doit être construit à partir de $s-1$ classifieurs binaires $\{H_{\geq m_g,c_l}\}_{2\leq g\leq s}$ en appliquant l'approche décrite dans la Section 1.1.3.
- en appliquant la décomposition horizontale, puis chaque classifieurs multi-labels $H_{\geq m_g}$ doit être construit à partir de k classifieurs binaires $\{H_{\geq m_g,c_l}\}_{1\leq l\leq k}$ en appliquant l'approche BR (Section 1.2.4.5).

Instances	$\eta_{\geq m_2,c_1}$
x_1	0
x_2	0
x_3	1
x_4	1
x_5	1
x_6	1
x_7	1
x_8	1
x_9	1
x_{10}	1
x_{11}	1
x_{12}	1

Données pour $H_{\geq m_2,c_1}$

Instances	$\eta_{\geq m_2,c_2}$
x_1	1
x_2	1
x_3	0
x_4	0
x_5	1
x_6	1
x_7	0
x_8	0
x_9	1
x_{10}	1
x_{11}	1
x_{12}	1

Données pour $H_{\geq m_2,c_2}$

Instances	$\eta_{\geq m_2,c_3}$
x_1	1
x_2	1
x_3	1
x_4	1
x_5	0
x_6	0
x_7	0
x_8	0
x_9	1
x_{10}	1
x_{11}	1
x_{12}	1

Données pour $H_{\geq m_2,c_3}$

Instances	$\eta_{\geq m_3,c_1}$
x_1	0
x_2	0
x_3	1
x_4	1
x_5	0
x_6	0
x_7	1
x_8	1
x_9	0
x_{10}	0
x_{11}	1
x_{12}	1

Données pour $H_{\geq m_3,c_1}$

Instances	$\eta_{\geq m_3,c_2}$
x_1	0
x_2	0
x_3	0
x_4	0
x_5	1
x_6	1
x_7	0
x_8	0
x_9	1
x_{10}	1
x_{11}	1
x_{12}	1

Données pour $H_{\geq m_3,c_2}$

Instances	$\eta_{\geq m_3,c_3}$
x_1	1
x_2	1
x_3	0
x_4	0
x_5	0
x_6	0
x_7	0
x_8	0
x_9	0
x_{10}	0
x_{11}	1
x_{12}	1

Données pour $H_{\geq m_3,c_3}$

Instances	$\eta_{\geq m_4,c_1}$
x_1	0
x_2	0
x_3	0
x_4	0
x_5	0
x_6	0
x_7	1
x_8	1
x_9	0
x_{10}	0
x_{11}	0
x_{12}	0

Données pour $H_{\geq m_4,c_1}$

Instances	$\eta_{\geq m_4,c_2}$
x_1	0
x_2	0
x_3	0
x_4	0
x_5	0
x_6	0
x_7	0
x_8	0
x_9	1
x_{10}	1
x_{11}	0
x_{12}	0

Données pour $H_{\geq m_4,c_2}$

Instances	$\eta_{\geq m_4,c_3}$
x_1	0
x_2	0
x_3	0
x_4	0
x_5	0
x_6	0
x_7	0
x_8	0
x_9	0
x_{10}	0
x_{11}	1
x_{12}	1

Données pour $H_{\geq m_4,c_3}$

Tableau 1.20: Décomposition complète des données multi-labels graduées du Tableau 1.17

1.3.3.4 Décomposition basée sur l'apprentissage de préférences entre les labels

Dans l'approche CLR (Section 1.2.4.11), un label virtuel est introduit et un ensemble de classifieurs est appris tel que chaque classifieur permet la prédiction d'une préférence entre deux labels. A l'étape de la prédiction, les labels sont ordonnés selon le nombre de fois où ils ont été préférés. La position du label virtuel dans l'ensemble ordonné de labels permet de séparer les labels à prédire (préférés plus de fois que le label virtuel) des labels à ne pas prédire. Les trois approches de décomposition de la classification multi-labels graduée appelées 'Horizontal Calibrated Label Ranking' (Horizontal_CLR), 'Full Calibrated Label Ranking' (Full_CLR), et 'Joined Calibrated Label Ranking' (Joined_CLR) (Brinker et al.,2014,[21]) sont basées sur une extension de l'approche CLR qui consiste en l'utilisation d'un ensemble de labels virtuels $W = \{w_2, \ldots, w_s\}$ au lieu d'un seul. Un ensemble de degrés d'association $V = \{v_2, \ldots, v_s\}$ est également introduit tel que $m_1 < v_2 < m_2 < v_3 < \ldots < m_{s-1} < v_s < m_s$. Chaque label w_g, $g \in [\![2,s]\!]$ a un degré d'association fixe v_g.

Afin de simplifier la notation, les labels $\{w_2, \ldots, w_s\}$ sont notés $\{c_{k+1}, \ldots, c_{k+s-1}\}$, et les degrés d'association $\{v_2, \ldots, v_s\}$ sont notés $\{m_{1.5}, \ldots, m_{(s-1).5}\}$.

Le Tableau 1.21 illustre le jeu de données du Tableau 1.17 après l'ajout des labels virtuels $\{w_1, w_2, w_3\}$ et des degrés d'association correspondants $\{v_1, v_2, v_3\}$.

Instances	Ensemble gradué de labels
x_1	$\{c_1 \mid m_1, c_2 \mid m_2, c_3 \mid m_3, w_2 \mid v_2, w_3 \mid v_3, w_4 \mid v_4\}$
x_2	$\{c_1 \mid m_1, c_2 \mid m_2, c_3 \mid m_3, w_2 \mid v_2, w_3 \mid v_3, w_4 \mid v_4\}$
x_3	$\{c_1 \mid m_3, c_2 \mid m_1, c_3 \mid m_2, w_2 \mid v_2, w_3 \mid v_3, w_4 \mid v_4\}$
x_4	$\{c_1 \mid m_3, c_2 \mid m_1, c_3 \mid m_2, w_2 \mid v_2, w_3 \mid v_3, w_4 \mid v_4\}$
x_5	$\{c_1 \mid m_2, c_2 \mid m_3, c_3 \mid m_1, w_2 \mid v_2, w_3 \mid v_3, w_4 \mid v_4\}$
x_6	$\{c_1 \mid m_2, c_2 \mid m_3, c_3 \mid m_1, w_2 \mid v_2, w_3 \mid v_3, w_4 \mid v_4\}$
x_7	$\{c_1 \mid m_4, c_2 \mid m_1, c_3 \mid m_1, w_2 \mid v_2, w_3 \mid v_3, w_4 \mid v_4\}$
x_8	$\{c_1 \mid m_4, c_2 \mid m_1, c_3 \mid m_1, w_2 \mid v_2, w_3 \mid v_3, w_4 \mid v_4\}$
x_9	$\{c_1 \mid m_2, c_2 \mid m_4, c_3 \mid m_2, w_2 \mid v_2, w_3 \mid v_3, w_4 \mid v_4\}$
x_{10}	$\{c_1 \mid m_2, c_2 \mid m_4, c_3 \mid m_2, w_2 \mid v_2, w_3 \mid v_3, w_4 \mid v_4\}$
x_{11}	$\{c_1 \mid m_3, c_2 \mid m_3, c_3 \mid m_4, w_2 \mid v_2, w_3 \mid v_3, w_4 \mid v_4\}$
x_{12}	$\{c_1 \mid m_3, c_2 \mid m_3, c_3 \mid m_4, w_2 \mid v_2, w_3 \mid v_3, w_4 \mid v_4\}$

Tableau 1.21: Données multi-labels graduées avec les labels virtuels

Horizontal_CLR

L'ensemble d'instances dont le degré d'association au label c_l est supérieur ou égal à m_g est donné par : $X_{\geq m_g, c_l} = \{x_i \in X, \mu_{c_l}(x_i) \geq m_g\}$.

L'ensemble d'instances dont le degré d'association au label c_l est strictement inférieur à m_g est donné par : $X_{<m_g, c_l} = \{x_i \in X, \mu_{c_l}(x_i) < m_g\} = X - X_{\geq m_g, c_l}$.

L'ensemble d'instances telles que:

- soit le degré d'association au label c_l est supérieur ou égale à m_g, et le degré d'association au label $c_{l'}$ est strictement inférieur à m_g.
- soit le degré d'association au label c_l est inférieur strictement à m_g, et le degré d'association au label $c_{l'}$ est supérieur ou égal à m_g.

est donné par : $X_{\geq m_g,c_l,c_{l'}} = (X_{\geq m_g,c_l} \cap X_{<m_g,c_{l'}}) \cup (X_{<m_g,c_l} \cap X_{\geq m_g,c_{l'}})$.

Soit $\beta_{\geq m_g,c_l,c_{l'}} : X_{\geq m_g,c_l,c_{l'}} \rightarrow \{c_l, c_{l'}\}$ la fonction qui donne pour chaque instance $x_i \in X_{\geq m_g,c_l,c_{l'}}$:

- le label c_l si $x_i \in X_{\geq m_g,c_l} \cap X_{<m_g,c_{l'}}$.
- le label $c_{l'}$ sinon (si $x_i \in X_{<m_g,c_l} \cap X_{\geq m_g,c_{l'}}$).

Soit $\gamma_{\geq m_g,c_l,w_g} : X \rightarrow \{c_l, w_g\}$ la fonction qui donne pour chaque instance $x_i \in X$ le label c_l si $x_i \in X_{\geq m_g,c_l}$, et le label w_g sinon.

L'idée de l'approche Horizontal_CLR est de construire un classifieur multi-labels gradué H à partir de deux ensembles de classifieurs binaires $\{H_{\geq m_g,c_l,c_{l'}}\}_{\substack{2\leq g\leq s \\ 1\leq l<l'\leq k}}$ et $\{H_{\geq m_g,c_l,w_g}\}_{\substack{2\leq g\leq s \\ 1\leq l\leq k}}$ ($(s-1)\frac{k(k+1)}{2}$ classifieurs binaires au total).

Chaque classifieur $H_{\geq m_g,c_l,c_{l'}} : a_1 \times \ldots \times a_p \rightarrow \{c_l\ ,\ c_{l'}\}$ apprend à partir de l'ensemble supervisé $(X_{\geq m_g,c_l,c_{l'}}, \beta_{\geq m_g,c_l,c_{l'}})$ à prédire le label à préférer entre c_l et $c_{l'}$ pour lui affecter un degré d'association supérieur ou égal à m_g.

Chaque classifieur $H_{\geq m_g,c_l,w_g}$ apprend à partir de l'ensemble supervisé $(X, \gamma_{\geq m_g,c_l,w_g})$ à prédire le label à préférer entre c_l et w_g.

Le nombre de votes reçus par un label c_l pour une instance x en utilisant les classifieurs $\{H_{\geq m_g,c_{l'},c_{l''}}\}_{1\leq l'<l''\leq k}$ et les classifieurs $\{H_{\geq m_g,c_{l'},w_g}\}_{1\leq l'\leq k}$ est noté $V_{\geq m_g,c_l}(x)$.

Le nombre de votes reçus par un label virtuel w_g pour une instance x en utilisant les classifieurs $\{H_{\geq m_g,c_l,w_g}\}_{1\leq l\leq k}$ est noté $V_{\geq m_g,w_g}(x)$.

Soit $H_{\geq m_g} : a_1 \times \ldots \times a_p \rightarrow \mathscr{P}(C)$ la fonction qui prédit pour toute instance x l'ensemble des labels ayant un degré supérieur ou égal à m_g.

En se basant sur l'approche CLR (Section 1.2.4.11), la fonction $H_{\geq m_g}$ peut être donnée par : $H_{\geq m_g} = \{c_l \in C,\ V_{\geq m_g,c_l}(x) \geq V_{\geq m_g,w_g}(x)\}$.

En se basant sur la fonction d'agrégation $h_{c_l} : a_1 \times \ldots \times a_p \rightarrow M$ telle que définie pour la décomposition horizontale dans la Section 1.3.3.2, le classifieur multi-labels gradué H peut être donné pour toute instance $x \in a_1 \times \ldots \times a_p$ par : $H(x) = \{c_l | h_{c_l}(x)\}_{1\leq l\leq k}$.

Le Tableau 1.22 illustre les données d'apprentissage générées pour les classifieurs correspondant au degré d'association m_2:
$\{H_{\geq m_2,c_1,c_2,}, H_{\geq m_2,c_1,c_3,m_1}, H_{\geq m_2,c_2,c_3,m_1}, H_{\geq m_2,c_1,w_1,m_1}, H_{\geq m_2,c_2,w_1,m_1}, H_{\geq m_2,c_3,w_1,m_1}\}$ à partir des données du Tableau 1.21 en appliquant l'approche Horizontal_CLR.

L'inconvénient de l'approche Horizontal_CLR est que chaque label virtuel permet de générer k classifieurs binaires, alors que chacun des autres labels permet de générer $(k-1)(s-1)$ classifieurs binaires. Ceci rend la prédiction biaisée puisqu'elle est basée sur le nombre de fois où un label est préféré alors que les labels n'ont pas le même nombre de confrontations.

Instances	$\beta_{\geq m_2,c_1,c_2}$
x_1	c_2
x_2	c_2
x_3	c_1
x_4	c_1
x_7	c_1
x_8	c_1

Données pour $H_{\geq m_2,c_1,c_2}$

Instances	$\beta_{\geq m_2,c_1,c_3}$
x_1	c_3
x_2	c_3
x_5	c_1
x_6	c_1
x_7	c_1
x_8	c_1

Données pour $H_{\geq m_2,c_1,c_3}$

Instances	$\beta_{\geq m_2,c_2,c_3}$
x_3	c_3
x_4	c_3
x_5	c_2
x_6	c_2

Données pour $H_{\geq m_2,c_2,c_3}$

Instances	$\gamma_{\geq m_2,c_1,w_2}$
x_1	w_2
x_2	w_2
x_3	c_1
x_4	c_1
x_5	c_1
x_6	c_1
x_7	c_1
x_8	c_1
x_9	c_1
x_{10}	c_1
x_{11}	c_1
x_{12}	c_1

Données pour $H_{\geq m_2,c_1,w_2}$

Instances	$\gamma_{\geq m_2,c_2,w_2}$
x_1	c_2
x_2	c_2
x_3	w_2
x_4	w_2
x_5	c_2
x_6	c_2
x_7	w_2
x_8	w_2
x_9	c_2
x_{10}	c_2
x_{11}	c_2
x_{12}	c_2

Données pour $H_{\geq m_2,c_2,w_2}$

Instances	$\gamma_{\geq m_2,c_3,w_2}$
x_1	c_3
x_2	c_3
x_3	c_3
x_4	c_3
x_5	w_2
x_6	w_2
x_7	w_2
x_8	w_2
x_9	c_3
x_{10}	c_3
x_{11}	c_3
x_{12}	c_3

Données pour $H_{\geq m_2,c_3,w_2}$

Tableau 1.22: Décomposition par l'approche Horizontal_CLR pour le degré d'association m_2

Full_CLR

Soit $X_{c_l,c_{l'}} = \{x_i \in X, \mu_{c_l}(x_i) \neq \mu_{c_{l'}}(x_i)\}$ l'ensemble d'instances pour lesquelles les labels c_l et $c_{l'}$ ont des degrés d'association différents.

Soit $\beta_{c_l,c_{l'}} : X \to \{c_l, c_{l'}\}$ la fonction qui donne pour chaque instance $x_i \in X$ le label ayant le degré d'association le plus grand parmi les labels c_l et $c_{l'}$. L'approche Full_CLR permet de construire un classifieur multi-labels gradué H en se basant sur $\frac{(k+s-2)(k+s-1)}{2}$ classifieurs binaires $\{H_{c_l,c_{l'}}\}_{1\leq l<l'\leq k+s-1}$. Chaque classifieur $H_{c_l,c_{l'}} : a_1 \times \ldots \times a_p \to \{c_l, c_{l'}\}$ apprend à partir de l'ensemble supervisé $(X_{c_l,c_{l'}}, \beta_{c_l,c_{l'}})$ à prédire pour toute instance x le label à préféré entre c_l et $c_{l'}$ (le label ayant le plus grand degré d'association à l'instance x). A l'étape de la prédiction, les labels sont

ordonnés selon le nombre de fois où ils ont été préférés. Les labels virtuels $\{c_l\}_{k+1\leq l\leq k+s-1}$ agissent en tant que points de coupure pour déterminer les labels à prédire pour chaque degré d'association. Les labels préférés plus de fois que le label $c_{k+s-1} = w_{s-1}$ sont prédits avec un degré d'association m_s. Ensuite, parmi les labels non prédits, ceux qui sont préférés plus de fois que le label $c_{k+s-2} = w_{s-2}$ sont prédits avec un degré d'association m_{s-1}. Cette démarche est appliquée jusqu'au degré m_2, ensuite le reste des labels est prédit avec le degré d'association m_1. Le Tableau 1.23 illustre les données d'apprentissage générées pour les classifieurs $\{H_{c_1,c_2}, H_{c_1,c_3}, H_{c_2,c_3}, H_{c_1,w_1}, H_{c_1,w_2}, H_{c_1,w_3}\}$ à partir des données du Tableau 1.21 en appliquant l'approche Full_CLR.

L'approche Full_CLR a l'inconvénient de ne pas prendre en compte les degrés d'association en générant les données d'apprentissage. Par exemple, lorsque le label c_1 est confronté au label c_2 dans le Tableau 1.23, le label c_2 est préféré pour les instances x_1 et x_9. Cependant l'information sur la distance en terme de degré d'association ($|m_2 - m_1|$ pour x_1 et $|m_4 - m_2|$ pour x_9) est perdue.

Instances	β_{c_1,c_2}
x_1	c_2
x_2	c_2
x_3	c_1
x_4	c_1
x_5	c_2
x_6	c_2
x_7	c_1
x_8	c_1
x_9	c_2
x_{10}	c_2

Données pour H_{c_1,c_2}

Instances	β_{c_1,c_3}
x_1	c_3
x_2	c_3
x_3	c_1
x_4	c_1
x_5	c_1
x_6	c_1
x_7	c_1
x_8	c_1
x_{11}	c_3
x_{12}	c_3

Données pour H_{c_1,c_3}

Instances	β_{c_2,c_3}
x_1	c_3
x_2	c_3
x_3	c_3
x_4	c_3
x_5	c_2
x_6	c_2
x_9	c_2
x_{10}	c_2
x_{11}	c_3
x_{12}	c_3

Données pour H_{c_2,c_3}

Instances	β_{c_1,w_1}
x_1	w_1
x_2	w_1
x_3	c_1
x_4	c_1
x_5	c_1
x_6	c_1
x_7	c_1
x_8	c_1
x_9	c_1
x_{10}	c_1
x_{11}	c_1
x_{12}	c_1

Données pour H_{c_1,w_1}

Instances	β_{c_1,w_2}
x_1	w_2
x_2	w_2
x_3	c_1
x_4	c_1
x_5	w_2
x_6	w_2
x_7	c_1
x_8	c_1
x_9	w_2
x_{10}	w_2
x_{11}	c_1
x_{12}	c_1

Données pour H_{c_1,w_2}

Instances	β_{c_1,w_3}
x_1	w_3
x_2	w_3
x_3	w_3
x_4	w_3
x_5	w_3
x_6	w_3
x_7	c_1
x_8	c_1
x_9	w_3
x_{10}	w_3
x_{11}	w_3
x_{12}	w_3

Données pour H_{c_1,w_3}

Tableau 1.23: Vue sur la décomposition par l'approche Full-CLR

Joined_CLR

L'approche Joined_CLR est une extension de l'approche Horizontal_CLR permettant d'équilibrer le nombre de confrontations entre les labels $\{c_l\}_{1\leq l\leq k}$ et les labels virtuels $\{c_l\}_{k+1\leq l\leq k+s-1}$. Le principe est de produire un classifieur binaire $H_{\geq m_g,c_l,c_{l'}}$ pour chaque paire de labels $(c_l,c_{l'})$ et pour chaque degré d'association m_g comme dans l'approche Horizontal_CLR, mais en considérant aussi les labels virtuels. L'approche Joined-CLR génère donc un ensemble de classifieurs $\{H_{\geq m_g,c_l,c_{l'}}\}_{\substack{2\leq g\leq s\\ 1\leq l<l'\leq k+s-1}}$ contenant tous les classifieurs générés par l'approche Horizontal_CLR $\{H_{\geq m_g,c_l,c_{l'}}\}_{\substack{2\leq g\leq s\\ 1\leq l<l'\leq k}}$ en plus d'autres classifieurs $\{H_{c_l,c_{l'},m_g}\}_{\substack{2\leq g\leq s\\ l<l' \text{ et } (l>k \text{ ou } l'>k)}}$ pour équilibrer le nombre de confrontations entre les labels.

Le nombre de fois où chaque label a été préféré est calculé par rapport à tous les degrés d'association.

Ensuite, un ordre global de labels est établi selon le nombre de votes reçus par chaque label. Le degré d'association à prédire pour chaque label dépend de sa position par rapport aux labels virtuels. Par exemple, si w_g est le label virtuel ayant le plus grand degré d'association v_g tel que c_l est préféré plus de fois que w_g alors le degré d'association à prédire pour c_l est m_g. L'inconvénient de l'approche Joined_CLR est que les confrontations ajoutées sont triviales. Par exemple, pour le classifieur $H_{\geq m_g, w_g, w_{g+1}}$ le label w_{g+1} est toujours préféré par rapport au label w_{g-1} puisque les degrés d'association des labels virtuels sont fixes: $v_{g-1} < m_g < v_g$.

1.3.4 Conclusion

La tâche de la classification multi-labels graduée peut être décomposée soit verticalement par rapport à chaque label en plusieurs tâches de classification ordinales, soit horizontalement par rapport à chaque degré d'association en plusieurs tâches de classification multi-labels, soit complètement par rapport aux labels et par rapport aux degrés d'association en plusieurs tâches de classification binaires. La classification multi-labels graduée peut être décomposée aussi en utilisant une extension d'une approche d'apprentissage de préférences entre les labels. L'avantage des approches de décomposition est de pouvoir intervenir sur la stratégie d'apprentissage des relations entre les labels sans modifier le classifieur lui même. Par exemple, il est possible d'établir une décomposition horizontale, ensuite résoudre chaque sous-tâche de classification multi-labels par une approche tenant compte des relations entre les labels (Section 1.2). Cependant, l'état de l'art manque encore d'études concernant l'apprentissage des relations entre les labels dans le cas de la classification multi-labels graduée (Cheng et al.,2010,[25];Brinker et al.,2014,[21]), et concernant l'impacte des relations apprises sur la performance de prédiction.

1.4 Synthèse

La classification multi-labels graduée est décomposable en plusieurs tâches de classification multi-labels. Nous pouvons donc nous intéresser à l'apprentissage des relations entre les labels au niveau de la classification multi-labels. L'aspect de la gradualité de la classification multi-labels graduée peut être traité au niveau de la stratégie de décomposition.

Les approches de l'état de l'art ne permettent pas de lever tous les verrous mis en avant dans l'introduction. En effet, les approches de classification multi-labels existantes (Section 1.2.4) permettent soit l'apprentissage des préférences entre les labels (Les approches Ranking by Pairwise Comparisons et Calibrated Label Ranking), soit l'apprentissage des dépendances entre les labels. Certaines approches de classification multi-labels ne permettent pas du tout l'apprentissage des relations entre les labels (l'approche Binary Relevance). D'autres approches permettent l'apprentissage de dépendances uniquement par rapport à une structure de dépendance prédéfinie (l'approche Classifier Chains). D'autres approches permettent l'apprentissage de dépendances sans fixer une structure de dépendance au préalable, cependant elles nécessitent l'apprentissage d'une première couche de

classifieurs indépendants afin de pouvoir établir une prédiction en cas d'une dépendance cyclique (les approches Dependent Binary Relevance et Aggregating Independent and Dependent classifiers). Ce type d'approches combinant des classifieurs dépendants et des classifieurs indépendants produit en fin de compte des prédictions qui ne sont pas nécessairement cohérentes avec les dépendances apprises. Il est possible de remédier à ce problème par l'ajout d'une troisième étape permettant de sélectionner l'ordre de chaînage des classifieurs. Ceci assure le fait que chaque classifieur tient compte de la prédiction de ses antécédents afin que la prédiction finale soit cohérente par rapport aux dépendances entre les labels. Cependant, l'inconvénient majeur de ce type d'approches est leurs complexités.

Chapitre 2

Nouvelles approches de classification multi-labels graduée

Ce chapitre présente une partie de la contribution de cette thèse qui consiste en la proposition de nouvelles approches de classification multi-labels permettant de lever les verrous mis en avant dans l'introduction (Figure 2.1). La première approche permet l'apprentissage des dépendances entre les labels sans fixer une structure de dépendance au préalable. L'idée clé est d'apprendre un classifieur par label en considérant les autres labels en tant qu'attributs descriptifs supplémentaires. Les éventuelles dépendances cycliques sont ensuite supprimées par le remplacement de certains classifieurs en utilisant un sous-ensemble d'attributs descriptifs évitant les dépendances cycliques. L'ensemble de classifieurs candidats pour être remplacés est obtenu grâce à une mesure (heuristique) appelée mesure de Pré-sélection. Le classifieur à remplacer en premier parmi les classifieurs candidats est obtenu grâce à une mesure appelée mesure de Sélection. Le nouveau sous-ensemble d'attributs supplémentaires à considérer pour remplacer le classifieur sélectionné est obtenu grâce à une mesure appelée mesure d'Intérêt de chaînage. Cette nouvelle approche basée sur les trois mesures 'Pré-sélection', 'Sélection', et 'Intérêt de chaînage' est appelée 'PSI'.

La deuxième nouvelle approche proposée dans ce chapitre est appelée PSI2. Il s'agit d'une extension de l'approche PSI permettant la sélection des dépendances qu'un classifieur peut apprendre. En effet, l'approche PSI a l'avantage d'être indépendante du classifieur de base utilisé, par contre elle ne distingue pas les dépendances fortes et les dépendances faibles entre les labels. Le risque de la propagation d'erreur de la prédiction est plus grand pour les prédictions basées sur des dépendances faibles. L'approche

PSI2 a l'inconvénient d'être compatible uniquement avec les classifieurs de type arbre ou règles de décision. Cette restriction est utile pour évaluer l'intérêt d'apprendre une dépendance par rapport à la profondeur où apparaît le nœud label correspondant dans l'arbre de décision. En effet, les nœuds labels qui apparaissent proches de la racine de l'arbre reflètent les dépendances fortes, et les nœuds labels qui apparaissent proches des feuilles reflètent les dépendances faibles. L'approche PSI2 est basée sur deux paramètres: le premier permet de déterminer les profondeurs où l'apprentissage de dépendances est autorisé, et le deuxième paramètre permet de déterminer le nombre maximal de nœuds labels qui peuvent appartenir à la même branche. Ce paramètre permet de limiter le risque de propagation de l'erreur en prédiction en limitant le nombre de nœuds labels qui peuvent engendrer des erreurs de prédiction.

La troisième nouvelle approche proposée dans ce chapitre est appelée CLR_PSI. Il s'agit d'une combinaison entre l'approche Calibrated Label Ranking (CLR) permettant l'apprentissage des relations de préférence entre les labels, et l'approche PSI permettant l'apprentissage des relations de dépendance. L'intérêt de cette combinaison est d'améliorer les prédictions en exploitant les deux types de relations entre les labels. L'approche CLR_PSI consiste à apprendre certains classifieurs de l'approche CLR en utilisant l'approche PSI pour apprendre les relations de dépendance. La prédiction pour l'approche CLR_PSI est effectuée telle que définie par l'approche CLR.

La quatrième nouvelle approche proposée dans ce chapitre est appelée Stacked_RPC_PSI. Il s'agit d'une combinaison de l'approche Ranking by Pairwise Comparisons (RPC) permettant l'apprentissage des relations de préférence entre les labels, et l'approche PSI. L'idée de l'approche Stacked_RPC_PSI est d'apprendre les classifieurs de l'approche RPC, puis les utiliser sur l'ensemble d'apprentissage pour étendre les données d'apprentissage par les prédictions fournies. L'approche PSI est ensuite utilisée pour apprendre des classifieurs tenant compte des prédictions fournies par les classifieurs de l'approche RPC. La prédiction finale fournie par l'approche Stacked_RPC_PSI est celle donnée par les classifieurs de l'approche PSI.

Les nouvelles approches proposées PSI et PSI2 peuvent être combinées avec toutes les approches de décomposition de la classification multi-labels graduée. Les nouvelles approches CLR_PSI, et Stacked_RPC_PSI peuvent être combinées uniquement avec des approches de décomposition de la classification multi-labels graduée basées sur l'apprentissage de préférences entre les labels (Figure 2.2).

Ce chapitre présente une étude évaluant les approches PSI, PSI2, CLR_PSI, et Stacked_RPC_PSI sur des jeux de données multi-labels et multi-labels graduées en les comparant aux approches de l'état de l'art.

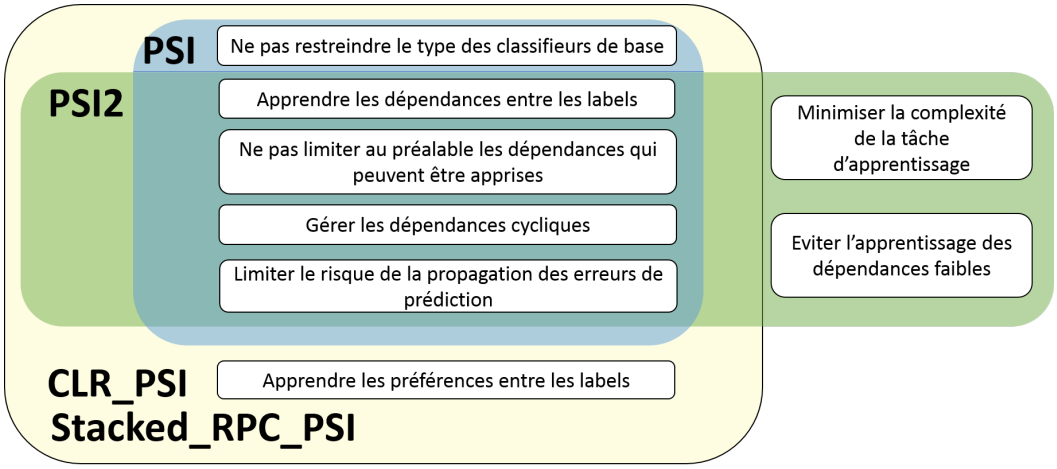


Figure 2.1: Verrous levés par les nouvelles approches proposées (PSI, PSI2, CLR_PSI, Stacked_RPC_PSI

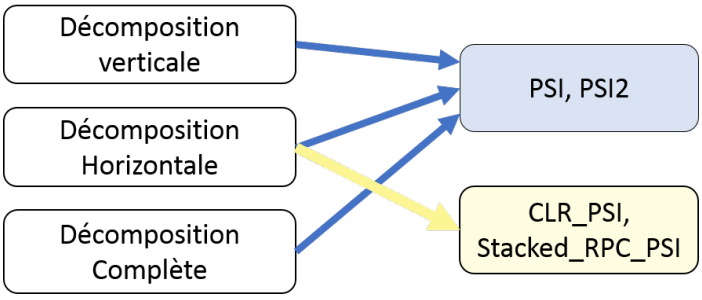


Figure 2.2: Compatibilité des approches proposées avec les stratégies de décomposition de la classification multi-labels graduée

2.1 L'approche PSI

L'idée de l'approche PSI est d'apprendre les dépendances entre les labels sans contraintes au préalable, et ensuite filtrer ces dépendances en supprimant les dépendances cycliques et en minimisant le risque de la propagation des erreurs de prédiction. Les mesures sur lesquelles l'approche PSI est basée doivent être construite dans cette perspective. Nous détaillons dans la suite l'algorithme de l'approche PSI, et nous illustrons l'impact de certains exemples de mesures que nous avons construit en se basant sur des heuristiques.

2.1.1 Algorithme PSI

Nous reprenons dans la suite les notations de la Section 1.2.1.

L'approche PSI (Algorithme 1) consiste à construire un classifieur $H : a_1 \times \ldots \times a_p \rightarrow \mathscr{P}(C)$ à partir d'un ensemble de k classifieurs $\{H_l\}_{1\leq l\leq k}$. Chaque classifieur $H_l : a_1 \times \ldots \times a_p \rightarrow \{0,1\}$ permet de prédire si une instance x est associée au label c_l ($H_l(x) = 1$) ou non ($H_l(x) = 0$). Le classifieur H est construit tel que $H(x) = \{c_l \in C, H_l(x) = 1\}$. Les classifieurs $\{H_l\}_{1\leq l\leq k}$ sont construits en considérant les labels en tant qu'attributs descriptifs supplémentaires. En effet, chaque instance x_i de l'ensemble d'apprentissage est étendue de k valeurs binaires $(x_{i,1},\ldots,x_{i,p},b_{i,1},\ldots,b_{i,k})$. L'ensemble d'instances étendues $\{(x_{i,1},\ldots,x_{i,p},b_{i,1},\ldots,b_{i,k})\}_{1\leq i\leq n}$ est noté $X_{A\cup C}$. Chaque valeur binaire $b_{i,l}$ traduit la présence ($b_{i,l} = 1$) ou l'absence ($b_{i,l} = 0$) du label c_l parmi l'ensemble de labels associés à x_i. Pour chaque instance x_i, le classifieur binaire H_l considère la valeur $b_{i,l}$ en tant que valeur de la classe à prédire, et le vecteur de valeurs $(x_{i,1},\ldots,x_{i,p},b_{i,1},\ldots,b_{i,l-1},b_{i,l+1},\ldots,b_{i,k})$ en tant que vecteur de valeurs des attributs descriptifs (Algorithme 1 lines 16-23).

L'ensemble initial de classifieurs $\mathscr{H} = \{H_l\}_{1\leq l\leq k}$ ne peut pas être utilisé directement en prédiction en cas d'une dépendance cyclique entre les classifieurs. Un ensemble final de classifieurs $\mathbb{H} = \{\mathbb{H}_l\}_{1\leq l\leq k}$ est construit à partir de l'ensemble $\mathscr{H}$ d'une façon itérative pour établir une chaîne de dépendances entre les classifieurs et éviter les dépendances cycliques. En chaque itération, **une mesure de pré-sélection** $\mathbb{P} : \mathscr{H} \rightarrow \{0,1\}$ est utilisée pour distinguer les classifieurs non pré-sélectionnés $\{H_l \in \mathscr{H}, \mathbb{P}(H_l) = 0\}$ qui peuvent être déplacés directement de $\mathscr{H}$ dans $\mathbb{H}$, (Algorithme 1 lines 6-7) et les classifieurs pré-sélectionnés $\{H_l \in \mathscr{H}, \mathbb{P}(H_l) = 1\}$ qui doivent être remplacés avant de les ajouter dans $\mathbb{H}$.

Une mesure de sélection $\mathbb{S} : \mathscr{P}(\mathscr{H}) \rightarrow \mathscr{H}$ est utilisée pour sélectionner un classifieur à remplacer H_r parmi l'ensemble de classifieurs pré-sélectionnés (Algorithme 1, line 9). Les classifieurs de l'ensemble final $\mathbb{H}$ ne dépendent d'aucun classifieur dans $\mathscr{H}$ y compris le classifieur H_r. Le nouveau classifieur $\mathbb{H}_r$ qui remplace H_r peut donc dépendre des classifieurs dans $\mathbb{H}$ sans risque d'une dépendance cyclique. **Une mesure d'intérêt de chaînage** $\mathbb{I} : \mathbb{H} \rightarrow \{0,1\}$ permet de sélectionner les labels que le classifieur $\mathbb{H}_r$ peut considérer en tant qu'attributs additionnels $\mathscr{C} = \{c_l, \mathbb{H}_l \in \mathbb{H}\ et\ \mathbb{I}(\mathbb{H}_l) = 1\}$ (Algorithme 1, line 10). Le classifieur $\mathbb{H}_r$ est construit en considérant les attributs additionnels $\mathscr{C}$ puis il est ajouté dans $\mathbb{H}$. L'ancien classifieur H_r est ensuite retiré de l'ensemble initial de classifieurs $\mathscr{H}$

(Algorithme 1, lines 11-13).

Certains classifieurs dans $\mathscr{H}$ peuvent perdre des dépendances et devenir indépendants lorsqu'un ou plusieurs classifieurs sont retirés de $\mathscr{H}$. La mesure de pré-sélection est évaluée chaque fois qu'un ensemble de classifieurs est retiré de $\mathscr{H}$ afin de déplacer le reste des classifieurs non pré-sélectionnés de $\mathscr{H}$ dans $\mathbb{H}$ (Algorithme 1, line 5).

2.1.2 Exemple d'exécution de l'approche PSI

Soit $X = \{x_i\}_{1 \leq i \leq 10}$ l'ensemble d'apprentissage du Tableau **??**. Chaque instance x_i est décrite par deux attributs descriptifs a_1 et a_2, et peut être associée à un ou plusieurs labels de l'ensemble $C = \{c_1, c_2, c_3, c_4\}$.

	a_1	a_2	c_1	c_2	c_3	c_4
x_1	20	30	0	0	0	0
x_2	35	35	0	0	0	0
x_3	15	40	0	1	0	1
x_4	20	50	0	1	0	1
x_5	30	45	1	1	0	1
x_6	35	30	1	1	0	1
x_7	10	40	1	0	1	1
x_8	15	45	1	0	1	1
x_9	25	55	1	0	1	0
x_{10}	30	60	1	0	1	0

Tableau 2.1: Ensemble de données multi-labels

Le classifieur de base considéré est un arbre de décision basé sur l'entropie sans élagage (Quinlan,1993,[89]). L'ensemble initial de classifieurs obtenu $\mathscr{H}$ est illustré dans la Figure 2.3. La structure de dépendance de $\mathscr{H}$ est illustrée dans la Figure 2.4.

Soit $D^{\rightarrow} : \mathscr{H} \rightarrow \mathscr{P}(\mathscr{H})$ la fonction qui donne pour chaque classifieur $H_l \in \mathscr{H}$ l'ensemble de classifieurs dont il dépend.

Soit $\mathbb{P}$ la mesure de pré-sélection donnée par $\mathbb{P}(H_l) = 1$ *si* $D^{\rightarrow}(H_l) \neq \emptyset$. Cette mesure permet de pré-sélectionner uniquement les classifieurs dépendants sans que la dépendance soit nécessairement cyclique. Comme la mesure de pré-sélection est évaluée itérativement après chaque déplacement des classifieurs non pré-sélectionnés de $\mathscr{H}$ dans $\mathbb{H}$ (Algorithme 1, lignes 5-8), les classifieurs impliqués dans une dépendance chaînée (non cyclique) seront tous retirés itérativement de $\mathscr{H}$. Les classifieurs restants dans $\mathscr{H}$ sont nécessairement impliqués dans une dépendance cyclique. La mesure de pré-sélection des classifieurs dépendants permet en effet de pré-sélectionner les classifieurs impliqués dans une dépendance cyclique.

L'approche PSI fait évoluer l'ensemble $\mathscr{H}$ en retirant itérativement des classifieurs jusqu'à ce que l'ensemble $\mathscr{H}$ devienne vide. L'état de départ de l'ensemble $\mathscr{H}$ est noté $\mathscr{H}^0$.

Algorithm 1 PSI

Input:

$A = \{a_j\}_{1 \leq j \leq p}$ */* ensemble d'attributs */*

$X_A = \{(x_{ij})_{a_j \in A}\}_{1 \leq i \leq n}$ */* ensemble d'instances */*

$C = \{c_l\}_{1 \leq l \leq k}$ */* ensemble de labels */*

$X_{A \cup C} = \{(x_{i,1}, \ldots, x_{i,p}, b_{i,1}, \ldots, b_{i,k})\}_{1 \leq i \leq n}$ */* ensemble d'instances étendues en considérant les labels comme attributs additionnels */*

H */* ensemble de tous les classifieurs possibles */*

$\mathscr{H} \subseteq H$ */* ensemble initial des classifieurs */*

$\mathbb{H} \subseteq H$ */* ensemble final des classifieurs */*

$\mathbb{P} : \mathscr{H} \rightarrow \{0,1\}$ */* mesure de pré-sélection */*

$\mathbb{S} : \mathscr{P}(\mathscr{H}) \rightarrow \mathscr{H}$ */* mesure de sélection */*

$\mathbb{I} : \mathbb{H} \rightarrow \{0,1\}$ */* mesure d'intérêt de chaînage */*

Output:

$\mathbb{H} = \{\mathbb{H}_l\}_{1 \leq l \leq k}$ */* ensemble final de classifieurs */*

1: **procedure** PSI($X_{A \cup C}, \mathbb{P}, \mathbb{S}, \mathbb{I}$)

2: $\mathbb{H} \leftarrow \emptyset$ */* l'ensemble final de classifieurs est initialement vide */*

3: $\mathscr{H} \leftarrow BUILD\_INITIAL\_CLASSIFIERS(X_{A \cup C})$ */* l'ensemble initial de classifieurs peut contenir des dépendances cycliques */*

4: **repeat** */* répéter les étapes suivantes tant que l'ensemble final de classifieurs n'est pas complet $|\mathbb{H}| \neq k$ */*

5: **repeat** */* vérifier après chaque déplacement des classifieurs non pré-sélectionnés de $\mathscr{H}$ dans $\mathbb{H}$ si ceci a permis de rendre d'autres classifieurs non pré-sélectionnés dans $\mathscr{H}$ */*

6: $\mathbb{H} \leftarrow \mathbb{H} \cup \{H_l \in \mathscr{H}, \mathbb{P}(H_l) = 0\}$ */* ajouter les classifieurs non pré-sélectionnés $\mathbb{P}(H_l) = 0$ dans $\mathbb{H}$ */*

7: $\mathscr{H} \leftarrow \mathscr{H} - \{H_l \in \mathscr{H}, \mathbb{P}(H_l) = 0\}$ */* retirer les classifieurs non pré-sélectionnés de $\mathscr{H}$ */*

8: **until** $\{H_l \in \mathscr{H}, \mathbb{P}(H_l) = 0\} = \emptyset$ */* ne pas s'arrêter avant de retirer tous les classifieurs non pré-sélectionnés de $\mathscr{H}$ */*

9: $H_r \leftarrow \mathbb{S}(\mathscr{H})$ */* sélectionner parmi les classifieurs pré-sélectionnés dans $\mathscr{H}$ celui à remplacer en premier $H_r \in \mathscr{H}$ */*

10: $\mathscr{C} \leftarrow \{c_l \in C, \mathbb{H}_l \in \mathbb{H}$ *and* $\mathbb{I}(\mathbb{H}_l) = 1\}$ */* sélectionner les labels à considérer comme attributs additionnels en utilisant la mesure d'intérêt de chaînage $\mathbb{I}$ */*

11: $\mathbb{H}_r \leftarrow BUILD\_BASE\_CLASSIFIER(X_{A \cup \mathscr{C}}, c_r)$ */* apprendre un classifieur binaire permettant de prédire la présence ou l'absence du label c_r en considérant les attributs descriptifs $A \cup \mathscr{C}$ */*

12: $\mathbb{H} \leftarrow \mathbb{H} \cup \{\mathbb{H}_r\}$ */* ajouter le nouveau classifieur $\mathbb{H}_r$ dans l'ensemble final de classifieurs */*

13: $\mathscr{H} \leftarrow \mathscr{H} - \{H_r\}$ */* retirer l'ancien classifieur H_r de l'ensemble initial de classifieurs */*

14: **until** $|\mathbb{H}| = k$ */* s'arrêter quand l'ensemble final de classifieurs est complet */*

15: **return** $\mathbb{H}$

16: **procedure** BUILD_INITIAL_CLASSIFIERS($X_{A\cup C}$) */* apprendre l'ensemble initial de classifieurs pouvant présenter des dépendances cycliques */*
17: $\mathscr{H} \leftarrow \emptyset$
18: **for each** $c_l \in C$ **do**
19: $H_l \leftarrow BUILD\_BASE\_CLASSIFIER(X_{A\cup C}, c_l)$
20: $\mathscr{H} \leftarrow \mathscr{H} \cup \{H_l\}$
21: **return** $\mathscr{H}$

22: **procedure** BUILD_BASE_CLASSIFIER($X_{A\cup C}, c_l$) */* apprendre un classifieur binaire sans contrainte sur le type du classifieur */*
23: **return** un classifieur binaire permettant de prédire l'absence ou la présence du label c_l en considérant un ensemble étendu d'attributs descriptifs $A \cup C - \{c_l\}$

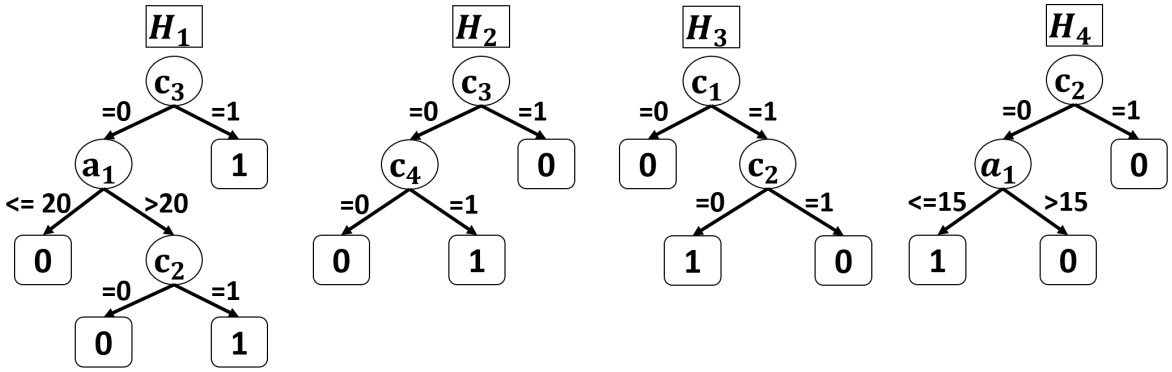


Figure 2.3: Arbres de décision dans $\mathscr{H}$

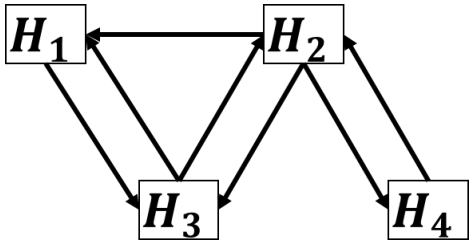


Figure 2.4: Structure de dépendance initiale des classifieurs de $\mathscr{H}$

Soit $D^{\rightarrow 0}: \mathscr{H}^0 \rightarrow \mathscr{P}(\mathscr{H}^0)$ la fonction qui donne pour chaque classifieur $H_l \in \mathscr{H}^0$ l'ensemble de classifieurs dont il dépend. La différence entre $D^{\rightarrow}$ et $D^{\rightarrow 0}$ est que certains classifieurs peuvent apparaître indépendants dans $\mathscr{H}$ à cause des classifieurs retirés, mais ils sont toujours dépendants dans $\mathscr{H}^0$.

Soit s un entier, et soit $\mathbb{P}$ la mesure de pré-sélection donnée par $\mathbb{P}(H_l) = 1 \ si \ |D^{\rightarrow 0}(H_l)| > s$. Cette mesure $\mathbb{P}$ permet de pré-sélectionner en plus des classifieurs impliqués dans une dépendance cyclique, les classifieurs qui dépendent d'un nombre de classifieurs supérieur à s. En effet, puisque $\mathscr{H} \subseteq \mathscr{H}^0$ alors $\forall H_l \in \mathscr{H} : D^{\rightarrow}(H_l) \subseteq D^{\rightarrow 0}(H_l)$. L'avantage de cette mesure $\mathbb{P}$ est qu'elle ne laisse pas passer dans l'ensemble $\mathbb{H}$ les classifieurs qui dépendent d'un grand nombre de classifieurs afin de réduire le risque de la propagation d'erreur de prédiction. Cependant, afin de simplifier l'exemple nous considérons dans la suite la première mesure de pré-sélection donnée par $\mathbb{P}(H_l) = 1 \ si \ D^{\rightarrow}(H_l) \neq \emptyset$.

Soit $D^{\leftarrow} : \mathscr{H} \rightarrow \mathscr{P}(\mathscr{H})$ la fonction qui donne pour chaque classifieur $H_l \in \mathscr{H}$ l'ensemble de classifieurs qui en dépendent.

Soit $MaxD^{\leftarrow} : \mathscr{P}(\mathscr{H}) \rightarrow \mathscr{P}(\mathscr{H})$ la fonction qui donne l'ensemble de classifieurs dont dépend le maximum nombre de classifieurs donnée par:
$MaxD^{\leftarrow}(\mathscr{H}) = \{H_l \in \mathscr{H}, |D^{\leftarrow}(H_l)| = \max_{H_{l'} \in \mathscr{H}} (|D^{\leftarrow}(H_{l'})|)\}$.

Soit $MaxD^{\rightarrow} : \mathscr{P}(\mathscr{H}) \rightarrow \mathscr{P}(\mathscr{H})$ la fonction qui donne l'ensemble de classifieurs qui dépendent du maximum nombre de classifieurs donnée par:
$MaxD^{\rightarrow}(\mathscr{H}) = \{H_l \in \mathscr{H}, |D^{\rightarrow}(H_l)| = \max_{H_{l'} \in \mathscr{H}} (|D^{\rightarrow}(H_{l'})|)\}$.

Soit $\mathbb{S} : \mathscr{P}(\mathscr{H}) \rightarrow \mathscr{H}$ la mesure de sélection donnée par
$\mathbb{S}(\mathscr{H}) = \underset{1 \leq l \leq k}{argmin}(H_l \in MaxD^{\rightarrow}(MaxD^{\leftarrow}(\mathscr{H})))$. Cette mesure trouve les classifieurs qui dépendent de plus de classifieurs $MaxD^{\leftarrow}(\mathscr{H})$ et récupère le classifieur $H_r \in \mathscr{H}$ du plus petit indice $r \in [\![1,k]\!]$ dont dépend le plus de classifieurs. L'objectif de cette mesure est de sélectionner le classifieur le plus chargé de dépendances afin de le remplacer pour réduire le risque de la propagation d'erreur de prédiction.

Comme il n'y a aucun classifieur indépendant dans $\mathscr{H}$ (Figure 2.4), tous les classifieurs sont pré-sélectionnés. Le classifieur H_2 est sélectionné puisqu'il est le classifieur dont dépend le plus de classifieurs (3 classifieurs). La mesure d'intérêt de chaînage n'a pas d'effet à cette étape puisque l'ensemble final de classifieurs est encore vide $\mathbb{H} = \emptyset$. Le nouveau classifieur $\mathbb{H}_2$ est donc construit sans attributs supplémentaires (Figure 2.5). Les nouvelles structures de dépendances de $\mathscr{H}$ et $\mathbb{H}$ après avoir retiré H_2 de $\mathscr{H}$ et après avoir ajouté $\mathbb{H}_2$ dans $\mathbb{H}$ sont illustrées dans la Figure 2.6.

Le classifieur H_4 n'est pas pré-sélectionné par la mesure $\mathbb{P}$. En effet, même s'il dépend toujours du classifieur H_2 (Figure 2.3), le fait d'avoir retiré H_2 de $\mathscr{H}$ fait apparaître H_2 comme un classifieur indépendant. H_2 est donc déplacé de $\mathscr{H}$ dans $\mathbb{H}$. Les structures de dépendances résultant sont illustrées dans la Figure 2.7.

Les classifieurs restant H_1 et H_3 sont pré-sélectionnés tous les deux par la mesure $\mathbb{P}$. Le nombre de classifieurs qui dépendent de H_1 (1 classifieur qui est H_3) est le même nombre de classifieurs qui

dépendent de H_3 (1 classifieur qui est H_1). Cependant le classifieur H_1 est sélectionné selon la mesure de sélection S puisqu'il a le plus petit indice.

Soit $\mathbb{I} : \mathbb{H} \rightarrow \{0,1\}$ la mesure d'intérêt de chaînage donnée par $\mathbb{I}(\mathbb{H}_l) = 1 \; \forall \mathbb{H}_l \in \mathbb{H}$. Cette mesure permet de chaîner le nouveau classifieur à apprendre avec tous les classifieurs présents dans l'ensemble final de classifieurs. Ceci permet d'apprendre plus de dépendances non cycliques entre les classifieurs mais risque d'augmenter le risque de la propagation d'erreur. La mesure d'intérêt de chaînage donnée par $\mathbb{I}(\mathbb{H}_l) = 0 \; \forall \mathbb{H}_l \in \mathbb{H}$ empêche les nouveaux classifieurs à construire de dépendre des classifieurs de l'ensemble final. Les seuls dépendances que l'approche PSI permet d'apprendre dans ce cas sont celles que laisse passer la mesure de pré-sélection comme le cas du classifieur H_4. Il est possible de construire la mesure d'intérêt de chaînage en n'autorisant par exemple que le chaînage avec les classifieurs indépendants de $\mathbb{H}$ afin de limiter le risque de propagation de l'erreur de prédiction. Cette mesure est donnée par $\forall \mathbb{H}_l \in \mathbb{H}$:
$\mathbb{I}(\mathbb{H}_l) = 1$ *si* $|D^{\rightarrow}(\mathbb{H}_l)| = 0$, *et* $\mathbb{I}(\mathbb{H}_l) = 0$ *sinon*. Cependant, afin de simplifier l'exemple dans la suite, nous considérons la première mesure d'intérêt de chaînage donnée par $\mathbb{I}(\mathbb{H}_l) = 1 \; \forall \mathbb{H}_l \in \mathbb{H}$.

Le nouveau classifieur $\mathbb{H}_1$ est construit selon la mesure d'intérêt de chaînage $\mathbb{I}$ en considérant les labels c_2 et c_4 en tant qu'attributs additionnels (Figure 2.8). Les structures de dépendances résultant sont illustrées dans la Figure 2.9.

H_3 est le seul classifieur restant dans $\mathscr{H}$ et il n'est pas pré-sélectionné selon la mesure $\mathbb{P}$. H_3 est donc déplacé de $\mathscr{H}$ dans $\mathbb{H}$. Les structures de dépendances résultant sont illustrées dans la Figure 2.10.

La mesure de pré-sélection, la mesure de sélection, et la mesure d'intérêt de chaînage sont appelées les mesures PSI. Le Tableau **??** illustre des structures de dépendances obtenues en exécutant l'approche PSI avec différentes mesures PSI.

2.2 L'approche PSI2

2.2.1 Description de l'approche PSI2

L'approche PSI2 est une amélioration de l'approche PSI dont l'objectif est de remédier aux points suivants:

- lorsqu'un classifieur est sélectionné, il faut le remplacer en le construisant à nouveau ;
- certaines dépendances apprises sont faibles (nœuds labels proches des feuilles dans un arbre de décion par exemple) et ne font qu'engendrer des dépendances cycliques inutiles.

Contrairement à l'approche PSI qui est compatible avec n'importe quel type de classifieur binaire, l'approche PSI2 est adaptée aux règles de décision et aux arbres de décision seulement. Les deux principaux avantages de cette restriction sont:

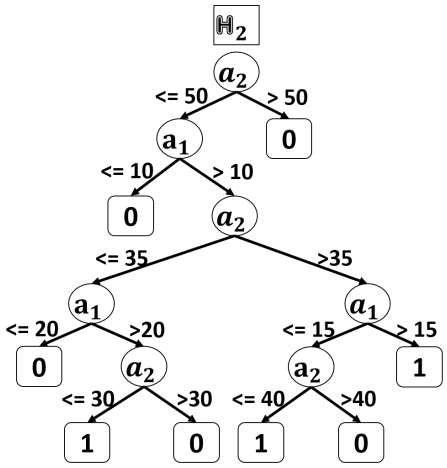


Figure 2.5: Le nouveau classifieur $\mathbb{H}_2$

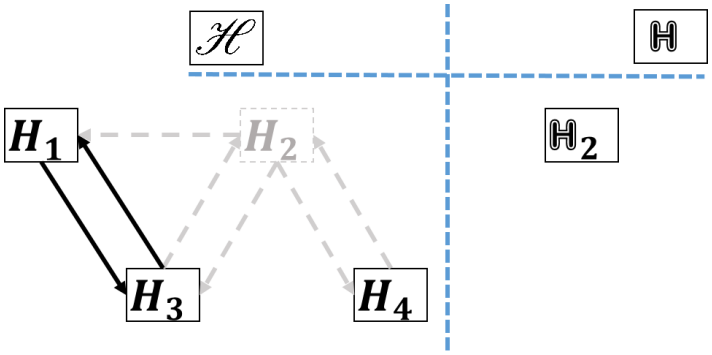


Figure 2.6: Structures de dépendances après le retrait de H_2

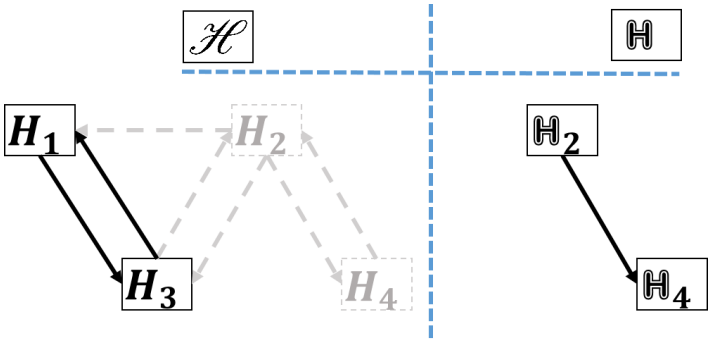


Figure 2.7: Structures de dépendances après l'ajout de H_4 dans $\mathbb{H}$

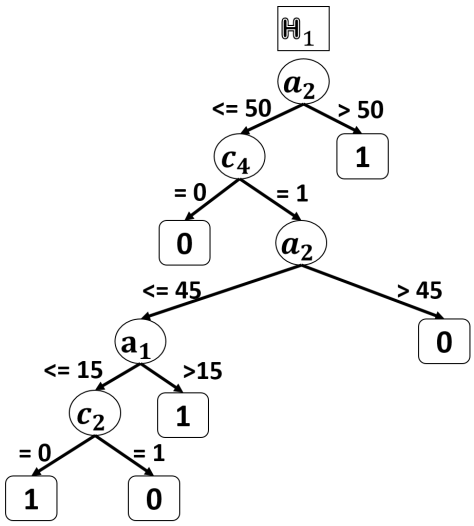


Figure 2.8: Le nouveau classifieur $\mathbb{H}_1$

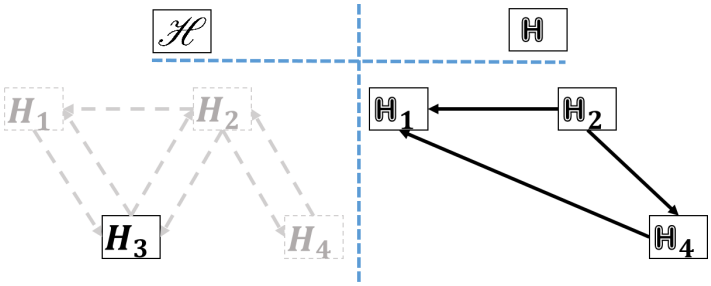


Figure 2.9: Structures de dépendances après le retrait de H_1

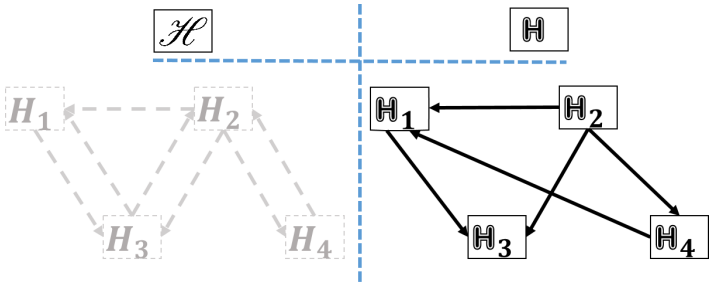


Figure 2.10: Structures de dépendances après l'ajout de H_3 dans $\mathbb{H}$

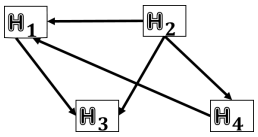
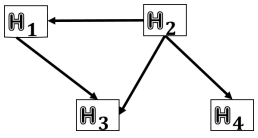
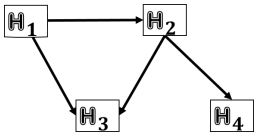
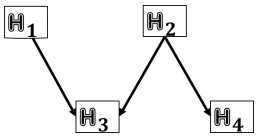
$\mathbb{P}$	$\mathbb{S}$	$\mathbb{I}$	Structure de dépendance finale
$\forall\ H_l \in \mathscr{H}$: $\mathbb{P}(H_l) = 1$ *if* $D^{\rightarrow}(H_l) \neq \emptyset$	$\mathbb{S}(\mathscr{H}) = \underset{1 \leq l \leq k}{argmin}(H_l \in MaxD^{\rightarrow}(MaxD^{\leftarrow}(\mathscr{H})))$	$\forall\ \mathbb{H}_l \in \mathbb{H}$: $\mathbb{I}(\mathbb{H}_l) = 1$	
$\forall\ H_l \in \mathscr{H}$: $\mathbb{P}(H_l) = 1$ *if* $D^{0\rightarrow}(H_l) \neq \emptyset$	$\mathbb{S}(\mathscr{H}) = \underset{1 \leq l \leq k}{argmin}(H_l \in MaxD^{\rightarrow}(MaxD^{\leftarrow}(\mathscr{H})))$	$\forall\ \mathbb{H}_l \in \mathbb{H}$: $\mathbb{I}(\mathbb{H}_l) = 1$	
$\forall\ H_l \in \mathscr{H}$: $\mathbb{P}(H_l) = 1$ *if* $D^{\rightarrow}(H_l) \neq \emptyset$	$\mathbb{S}(\mathscr{H}) = \underset{1 \leq l \leq k}{argmin}(H_l \in MaxD^{\rightarrow}(\mathscr{H}))$	$\forall\ \mathbb{H}_l \in \mathbb{H}$: $\mathbb{I}(\mathbb{H}_l) = 1$	
$\forall\ H_l \in \mathscr{H}$: $\mathbb{P}(H_l) = 1$ *if* $D^{\rightarrow}(H_l) \neq \emptyset$	$\mathbb{S}(\mathscr{H}) = \underset{1 \leq l \leq k}{argmin}(H_l \in MaxD^{\rightarrow}(MaxD^{\leftarrow}(\mathscr{H})))$	$\forall\ \mathbb{H}_l \in \mathbb{H}$: $\mathbb{I}(\mathbb{H}_l) = 0$	

Tableau 2.2: Impact des mesures PSI sur la structure de dépendance obtenue

- il suffit de reconstruire le sous-arbre à partir du nœud label qui cause la dépendance cyclique au lieu de reconstruire le classifieur sélectionné en entier ;

- il est possible d'empêcher l'apprentissage des dépendance faibles en ne considérant plus les attributs supplémentaires à partir d'une profondeur seuil de l'arbre de décision ;

- il est possible de minimiser le risque de propagation de l'erreur de prédiction en minimisant le nombre de nœuds labels dans la même branche de l'arbre de décision.

L'approche PSI2 est basée sur les mêmes mesures de l'approche PSI et reste toujours une approche de transformation qui ne modifie pas l'algorithme de base. L'approche PSI2 est décrite dans l'Algorithme 2. Seule la procédure 'BUILD_BASE_CLASSIFIER' nécessite d'être modifiée par rapport à l'approche PSI (Algorithme 1). L'apport de l'approche PSI2 est que les attributs descriptifs considérés par les classifieurs de base sont dynamiques et varient selon deux paramètres:

- L'intervalle de profondeur dans lequel les labels sont considérés en tant qu'attributs descriptifs supplémentaires (Depth Range: DR) ;

- Le nombre autorisé d'attributs label dans la même règle de décision (branche dans l'arbre de décision) (chain length: CL).

2.2.2 Exemple d'exécution de l'approche PSI2

L'ensemble initial de classifieurs construits par l'approche PSI2 en considérant les parmètres $DR = [0,1]$ et $CL = 1$ pour l'exemple du Tableau **??** est illustré dans la Figure 2.11.

L'arbre de décision H_1 construit par l'approche PSI (Fig. 2.3) contient le nœud label c_2 à la profondeur 3. Cependant, l'attribut label c_2 ne peut pas être utilisé dans l'approche PSI2 à la profondeur 3 à cause du paramètre $DR = [0,1]$. Le nœud label c_2 dans l'approche PSI2 est remplacé par les nœuds attributs a_1 et a_2.

L'arbre de décision H_3 construit par l'approche PSI (Fig. 2.3) contient le nœud label c_2 à la profondeur 1. Cependant, l'attribut label c_2 ne peut pas être utilisé dans l'approche PSI2 après avoir utilisé l'attribut label c_1 à la profondeur 0 à cause du paramètre $CL = 1$.

La structure de dépendance obtenue pour l'ensemble initial de classifieurs est illustrée dans la Figure 2.12. La structure de dépendance finale obtenue en utilisant les mesures PSI données par la première ligne du Tableau **??** est illustrée dans la Figure 2.13.

Les paramètres de l'approche PSI2 ont un impact sur la structure de dépendance initiale qui peut être cyclique. Le Tableau **??** illustre des structures initiales de dépendance obtenues en utilisant différentes valeurs pour les paramètres *DR* et *CL*.

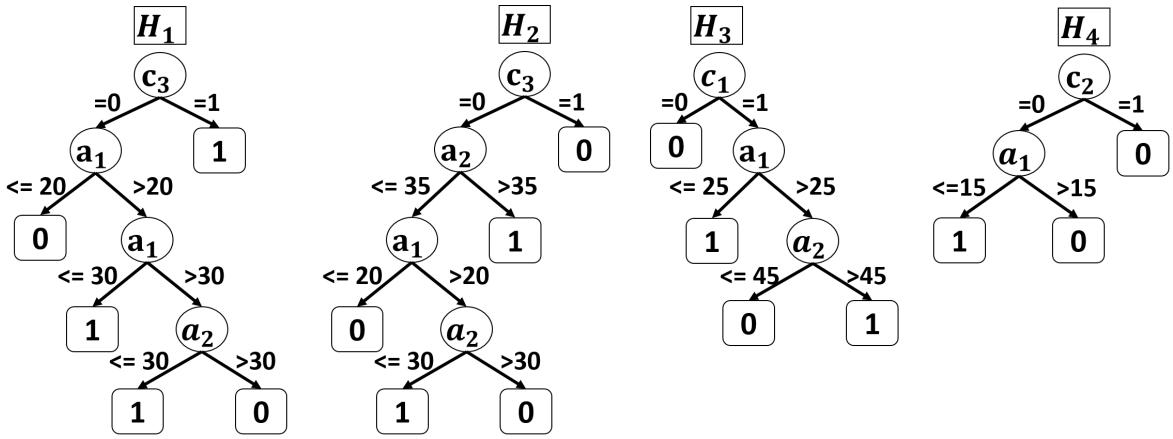


Figure 2.11: L'approche PSI2: arbres de décision à l'étape initiale

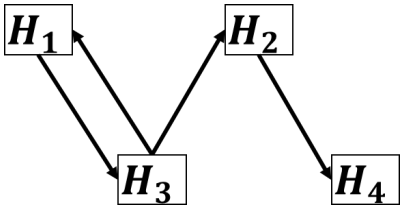


Figure 2.12: L'approche PSI2: structure de dépendance des classifieurs initiaux

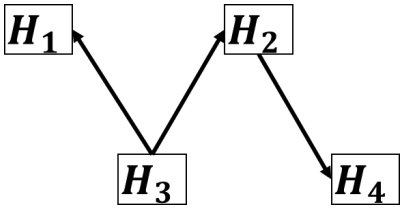


Figure 2.13: L'approche PSI2: structure de dépendance finale

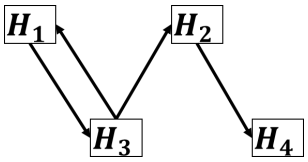
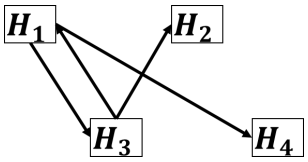
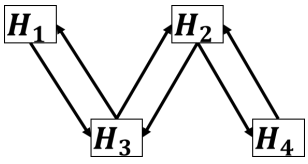
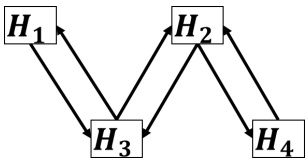
DR	*CL*	Initial dependency structure
[0, 1]	1	
[1, 3]	1	
[0, 1]	2	
[1, 3]	2	

Tableau 2.3: Structures de dépendances pour différents paramètres de l'approche PSI2

Algorithm 2 PSI2

Input:

$A = \{a_j\}_{1 \leq j \leq p}$ */* ensemble d'attributs */*

$X_A = \{(x_{ij})_{a_j \in A}\}_{1 \leq i \leq n}$ */* ensemble d'instances */*

$C = \{c_l\}_{1 \leq l \leq k}$ */* ensemble de labels */*

$X_{A \cup C} = \{(x_{i,1}, \ldots, x_{i,p}, b_{i,1}, \ldots, b_{i,k})\}_{1 \leq i \leq n}$ */* ensemble d'instances étendues en considérant les labels comme attributs additionnels */*

$\mathbb{P} : \mathscr{H} \to \{0,1\}$ */* mesure de pré-sélection */*

$\mathbb{S} : \mathscr{P}(\mathscr{H}) \to \mathscr{H}$ */* mesure de sélection */*

$\mathbb{I} : \mathbb{H} \to \{0,1\}$ */* mesure d'intérêt de chaînage */*

DR / l'intervalle de profondeur où les nœuds labels sont autorisés */*

CL / le nombre seuil de nœuds labels dans une même branche de l'arbre de décsion */*

Output:

$\mathbb{H} = \{\mathbb{H}_l\}_{1 \leq l \leq k}$ */* ensemble final de classifieurs */*

1: **procedure** PSI2($X_{A \cup C}, \mathbb{P}, \mathbb{S}, \mathbb{I}$)

2: $\mathbb{H} \leftarrow \emptyset$ */* l'ensemble final de classifieurs est initialement vide */*

3: $\mathscr{H} \leftarrow BUILD\_INITIAL\_CLASSIFIERS(X_{A \cup C})$ */* l'ensemble initial de classifieurs peut contenir des dépendances cycliques */*

4: **repeat** */* répéter les étapes suivantes tant que l'ensemble final de classifieurs n'est pas complet* $|\mathbb{H}| \neq k$ **/*

5: **repeat** */* vérifier après chaque déplacement des classifieurs non pré-sélectionnés de* $\mathscr{H}$ *dans* $\mathbb{H}$ *si ceci a permis de rendre d'autres classifieurs non pré-sélectionnés dans* $\mathscr{H}$ **/*

6: $\mathbb{H} \leftarrow \mathbb{H} \cup \{H_l \in \mathscr{H}, \mathbb{P}(H_l) = 0\}$ */* ajouter les classifieurs non pré-sélectionnés* $\mathbb{P}(H_l) = 0$ *dans* $\mathbb{H}$ **/*

7: $\mathscr{H} \leftarrow \mathscr{H} - \{H_l \in \mathscr{H}, \mathbb{P}(H_l) = 0\}$ */* retirer les classifieurs non pré-sélectionnés de* $\mathscr{H}$ **/*

8: **until** $\{H_l \in \mathscr{H}, \mathbb{P}(H_l) = 0\} = \emptyset$ */* ne pas s'arrêter avant de retirer tous les classifieurs non pré-sélectionnés de* $\mathscr{H}$ **/*

9: $H_r \leftarrow \mathbb{S}(\mathscr{H})$ */* sélectionner parmi les classifieurs pré-sélectionnés dans* $\mathscr{H}$ *celui à remplacer en premier* $H_r \in \mathscr{H}$ **/*

10: $\mathscr{C} \leftarrow \{c_l \in C, \mathbb{H}_l \in \mathbb{H}$ *and* $\mathbb{I}(\mathbb{H}_l) = 1\}$ */* sélectionner les labels à considérer comme attributs additionnels en utilisant la mesure d'intérêt de chaînage* $\mathbb{I}$ **/*

11: $\mathbb{H}_r \leftarrow BUILD\_BASE\_CLASSIFIER(X_{A \cup \mathscr{C}}, c_r)$ */* apprendre un classifieur binaire permettant de prédire la présence ou l'absence du label* c_r *en considérant les attributs descriptifs* $A \cup \mathscr{C}$ **/*

12: $\mathbb{H} \leftarrow \mathbb{H} \cup \{\mathbb{H}_r\}$ */* ajouter le nouveau classifieur* $\mathbb{H}_r$ *dans l'ensemble final de classifieurs */*

13: $\mathscr{H} \leftarrow \mathscr{H} - \{H_r\}$ */* retirer l'ancien classifieur* H_r *de l'ensemble initial de classifieurs */*

14: **until** $|\mathbb{H}| = k$ */* s'arrêter quand l'ensemble final de classifieurs est complet */*

15: **return** $\mathbb{H}$

Algorithm 2 PSI2 - suite

procedure BUILD_INITIAL_CLASSIFIERS($X_{A\cup C}$) */* apprendre l'ensemble initial de classifieurs pouvant présenter des dépendances cycliques */*

$\mathscr{H} \leftarrow \emptyset$

for each $c_l \in C$ **do**

$H_l \leftarrow BUILD\_BASE\_CLASSIFIER(X_{A\cup C}, c_l)$

$\mathscr{H} \leftarrow \mathscr{H} \cup \{H_l\}$

return $\mathscr{H}$

---

procedure BUILD_BASE_CLASSIFIER($X_{A\cup\mathscr{C}}, c_l$) */* $\mathscr{C} \subseteq C$ */*

$dep \leftarrow 0$ */* Le nœud racine de l'arbre de décision correspond à la profondeur* 0 *et au nombre de nœuds labels précédents* 0 **/*

$pre \leftarrow 0$

RECURSIVE_BUILD_TREE($N_{X_{A\cup\mathscr{C}},dep,pre}, c_l$) */* construire un arbre de décision pour prédire l'absence ou la présence du label c_l d'une façon récursive à partir du nœud $N_{X_{A\cup\mathscr{C}},dep,pre}$ correspondant à l'ensemble d'apprentissage $X_{A\cup\mathscr{C}}$, à la profondeur $dep = 0$ et au nombre de nœuds labels précédents dans la même branche $pre = 0$ */*

---

procedure RECURSIVE_BUILD_TREE($N_{X'_{A\cup\mathscr{C}},dep,pre}, c_l$) */* $\mathscr{C} \subseteq C, X'_{A\cup\mathscr{C}} \subseteq X_{A\cup\mathscr{C}}$ */*

for each $N_{X''_{A\cup\mathscr{C}},dep',pre'} \in SPLIT\_NODE(N_{X'_{A\cup\mathscr{C}},dep,pre}, c_l)$ **do** */* Tant que le nœud $N_{X'_{A\cup\mathscr{C}},dep,pre}$ est divisé en plusieurs nœuds par la procédure SPLIT_NODE il faut appliquer la procédure RECURSIVE_BUILD_TREE sur chaque nœud généré $N_{X''_{A\cup\mathscr{C}},dep',pre'}$ avec $dep' = dep+1$, et $pre' = pre+1$ si N est un nœud label, sinon $pre' = pre$ */*
RECURSIVE_BUILD_TREE($N_{X''_{A\cup\mathscr{C}},dep',pre'}, c_l$)

---

procedure SPLIT_NODE($N_{X'_{A\cup\mathscr{C}},dep,pre}, c_l$)

if $((dep \notin DR)$ *or* $(pre > CL))$ **then** */* si la profondeur 'dep' n'est pas dans l'intervalle de profondeur où l'apprentissage de dépendances est autorisé, ou le nombre maximal de nœuds labels dans une même branche de l'arbre de décision est atteint, alors le nœud $N_{X'_{A\cup\mathscr{C}},dep,pre}$ ne doit pas considérer des attributs supplémentaires */*

$a \leftarrow SELECT\_ATTRIBUTE(N_{X'_A,dep,pre}, c_l)$ */* sélectionner un attribut parmi l'ensemble A pour construire un nouveau nœud */*

else $a \leftarrow SELECT\_ATTRIBUTE(N_{X'_{A\cup\mathscr{C}},dep,pre}, c_l)$ */* sinon, sélectionner un attribut pour construire un nouveau nœud en considérant les attributs supplémentaires $A\cup\mathscr{C}$ */*

return un nœud pour chaque sous-ensemble d'instances correspondant à une valeur de l'attribut sélectionné a, ou un ensemble vide de nœuds si le nœud $N_{X'_{A\cup\mathscr{C}},dep,pre}$ doit être une feuille.

---

procedure SELECT_ATTRIBUTE($N_{X'_A,Y'_\mathscr{C},dep,pre}, \{y_{il}\}_{x_i\in X'_A}$)

return le meilleur attribut dans $A\cup\mathscr{C}$ qui discrimine au mieux les instances associées au label c_l

2.3 L'approche CLR_PSI

2.3.1 Description de l'approche CLR_PSI

L'approche CLR_PSI consiste à construire un classifieur multi-labels H à partir de $\frac{k(k+1)}{2}$ classifieurs binaires $\{H_{c_l,c_{l'}}\}_{1\leq l<l'\leq k} \cup \{H_{c_l,c_0}\}_{1\leq l\leq k}$. Les classifieurs $\{H_{c_l,c_{l'}}\}_{1\leq l<l'\leq k}$ sont construits exactement comme dans l'approche CLR (Section 1.2.4.11). Par contre, les classifieurs $\{H_{c_l,c_0}\}_{1\leq l\leq k}$ sont construits selon l'approche PSI (Section 2.1) pour permettre l'apprentissage des dépendances entre les labels. Le classifieur multi-labels H prédit pour une instance x tous les labels qui ont été préférés plus de fois que le label virtuel c_0 (comme dans l'approche CLR).

2.3.2 Exemple d'exécution de l'approche CLR_PSI

Nous reprenons l'exemple du Tableau **??** en se limitant à l'ensemble de labels $C = \{c_1, c_2, c_3\}$ afin de simplifier l'exemple. Le Tableau 2.4 présente les valeurs correspondant à la préférence entre les labels. La valeur $\emptyset$ indique que la donnée correspondante n'est pas incluse dans l'apprentissage de préférences. Par exemple, l'ensemble d'apprentissage du classifieur H_{c_1,c_2} qui prédit la préférence entre les labels c_1 et c_2 est réduit à $X_{c_1,c_2} = \{x_3, x_4, x_7, x_8, x_9, x_{10}\}$. Les classifieurs $\{H_{c_1,c_2}, H_{c_1,c_3}, H_{c_2,c_3}\}$ ne partagent pas le même ensemble de données d'apprentissage. Ceci explique le fait que seuls les classifieurs apprenant les relations de préférence par rapport au label virtuel c_0 sont construits selon l'approche PSI pour apprendre les relations de dépendance.

La Figure 2.14 illustre les arbres de décision $\{H_{c_1,c_2}, H_{c_1,c_3}, H_{c_2,c_3}\}$. La Figure 2.15 illustre les arbres de décision $\{H_{c_1,c_0}, H_{c_2,c_0}, H_{c_3,c_0}\}$. La Figure 2.16 illustre la structure de dépendances des classifieurs construits par l'approche CLR_PSI.

L'avantage de l'approche CLR_PSI est qu'elle utilise le même nombre de classifieurs que l'approche CLR. L'approche CLR_PSI permet d'apprendre simultanément les relations de préférence et de dépendance entre les labels. Cependant, l'inconvénient de l'approche CLR_PSI est qu'elle ne permet pas d'apprendre une relation de dépendance impliquant une relation de préférence. L'apprentissage des relations de dépendance est en fait limité aux classifieurs impliquant le label virtuel $(H_{c_1,c_0}, H_{c_2,c_0}, H_{c_3,c_0})$.

Instances	a_1	a_2	$c_1\#c_2$	$c_1\#c_3$	$c_2\#c_3$	$c_1\#c_0$	$c_2\#c_0$	$c_3\#c_0$
x_1	20	30	$\emptyset$	$\emptyset$	$\emptyset$	c_0	c_0	c_0
x_2	35	35	$\emptyset$	$\emptyset$	$\emptyset$	c_0	c_0	c_0
x_3	15	40	c_2	$\emptyset$	c_2	c_0	c_2	c_0
x_4	20	50	c_2	$\emptyset$	c_2	c_0	c_2	c_0
x_5	30	45	$\emptyset$	c_1	c_2	c_1	c_2	c_0
x_6	35	30	$\emptyset$	c_1	c_2	c_1	c_2	c_0
x_7	10	40	c_1	$\emptyset$	c_3	c_1	c_0	c_3
x_8	15	45	c_1	$\emptyset$	c_3	c_1	c_0	c_3
x_9	25	55	c_1	$\emptyset$	c_3	c_1	c_0	c_3
x_{10}	30	60	c_1	$\emptyset$	c_3	c_1	c_0	c_3

Tableau 2.4: Exemple de données d'apprentissage pour l'approche CLR_PSI

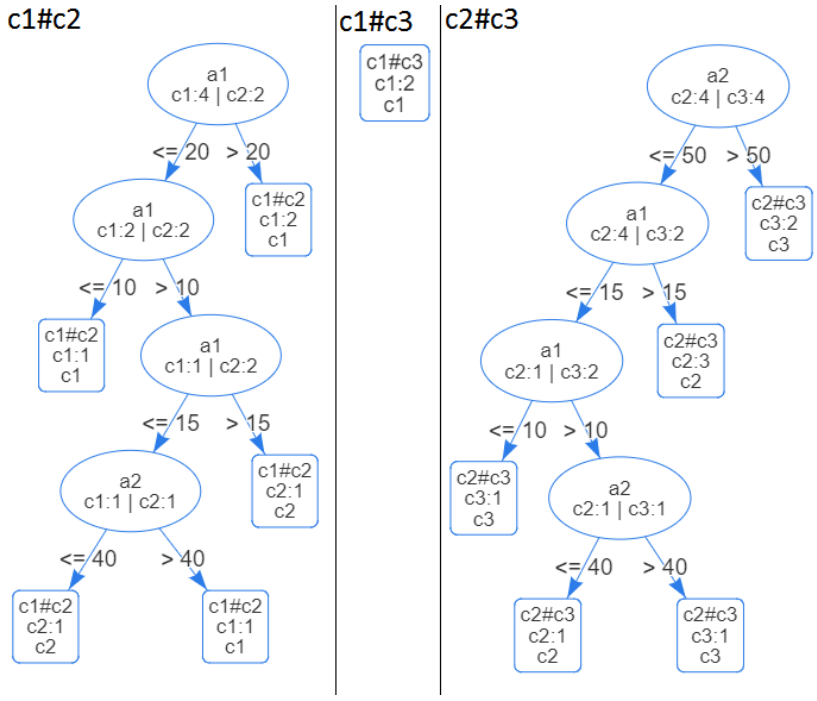


Figure 2.14: L'ensemble d'arbres de décision construits par l'approche CLR_PSI selon l'approche CLR à partir des données d'apprentissage du Tableau 2.4

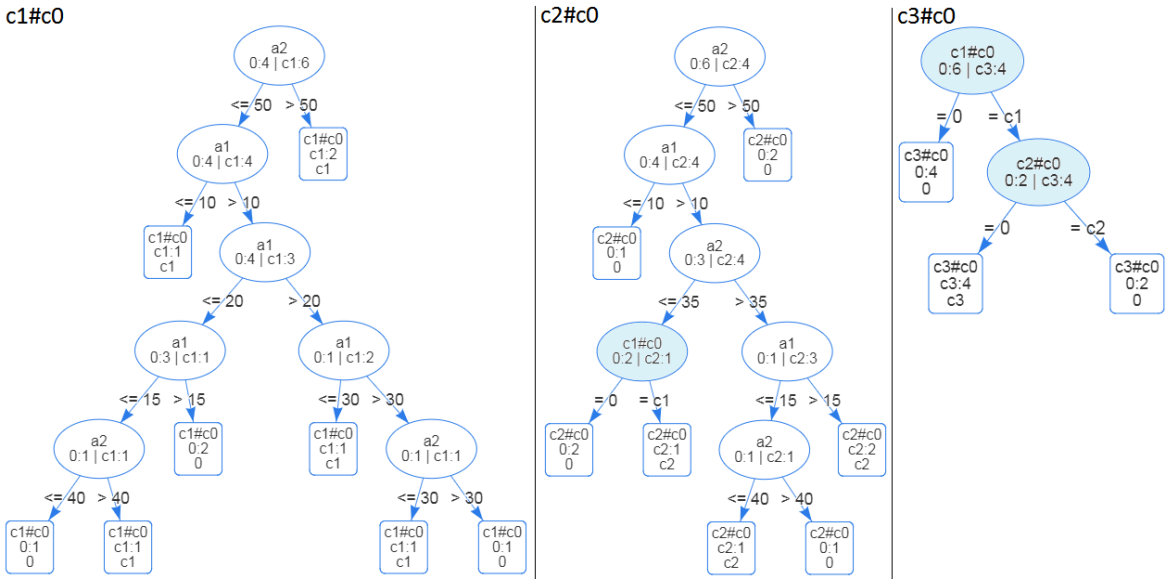


Figure 2.15: L'ensemble d'arbres de décision construits par l'approche CLR_PSI selon l'approche PSI à partir des données d'apprentissage du Tableau 2.4

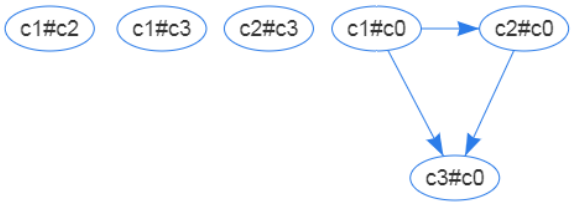


Figure 2.16: Structure de dépendances correspondant aux classifieurs construits par l'approche CLR_PSI à partir des données d'apprentissage du Tableau 2.4

2.4 L'approche Stacked_RPC_PSI

2.4.1 Description de l'approche Stacked_RPC_PSI

D'un côté, l'approche PSI (Section 2.1) est basée sur k classifieurs mono-labels, et permet l'apprentissage des relations de dépendance entre les labels sans imposer une restriction au préalable sur les relations à apprendre. D'un autre côté, l'approche RPC (Section 1.2.4.10) permet l'apprentissage des relations de préférence entre les labels en utilisant $\frac{k(k-1)}{2}$ classifieurs mono-labels. L'approche Stacked_RPC_PSI combine les approches RPC et PSI selon une stratégie différente de l'approche CLR_PSI afin de permettre l'apprentissage des relations de dépendance impliquant des relations de préférence entre les labels. L'idée de l'approche Stacked_RPC_PSI est d'utiliser les classifieurs de préférence pour prédire les valeurs manquantes. Ceci permet d'inclure les classfieurs de préférence dans la structure de dépendances apprise par l'approche PSI.

L'ensemble des classifieurs $\{h_{c_l,c_{l'}}\}_{1\leq l<l'\leq k}$ est construit exactement comme dans l'approche RPC. Ensuite, l'ensemble d'apprentissage est étendu par les prédictions des classifieurs de l'approche RPC: $\forall x_i \in X$: $x_i^e = (x_{i,a_1},\ldots,x_{i,a_p},h_{c_1,c_2}(x_i),h_{c_1,c_3}(x_i),\ldots,h_{c_{k-1},c_k}(x_i))$.
Un ensemble de k classifieurs $\{H_{c_l}\}_{1\leq l\leq k}$ est construit comme dans l'approche PSI mais en considérant l'ensemble d'apprentissage étendu $X^e = \{x_i^e\}_{1\leq i\leq n}$. Le classifieur multi-labels H est donné par: $H(x) = \{c_l \in C, H_{c_l}(x) = 1\}$.

2.4.2 Exemple d'exécution de l'approche Stacked_RPC_PSI

Nous reprenons l'exemple du Tableau **??** en se limitant à l'ensemble de labels $C = \{c_1,c_2,c_3\}$ afin de simplifier l'exemple. Le Tableau 2.5 présente les données d'apprentissage de l'approche Stacked_RPC_PSI.

L'ensemble des classifieurs $\{H_{c_1,c_2},H_{c_1,c_3},H_{c_2,c_3}\}$ est construit exactement comme dans l'approche CLR_PSI (Figure 2.14). Le Tableau 2.6 illustre les données d'apprentissage mises à jour après l'introduction des prédictions des classifieurs H_{c_1,c_2}, H_{c_1,c_3}, et H_{c_2,c_3}. La Figure 2.17 illustre les arbres de décision construits selon l'approche PSI en considérant les prédictions fournies par les classifieurs H_{c_1,c_2}, H_{c_1,c_3}, et H_{c_2,c_3}. La Figure 2.18 illustre la structure de dépendances correspondante.

L'avantage de l'approche Stacked_RPC_PSI est qu'elle permet d'apprendre des règles de classification impliquant des relations de dépendance et de préférence entre les labels. Par exemple, la règle de classification extraite de l'arbre de décision H_2 dans la Figure 2.17: "si le label c_2 est préféré au label c_1, et si la valeur de l'attribute a_2 est ≤ 35, et si le label c_1 est prédit, alors prédire le le label c_2". L'approche Stacked_RPC_PSI utilise le même nombre de classifieurs que l'approche CLR_PSI. Cependant, pour l'approche Stacked_RPC_PSI, les classifieurs construits selon l'approche PSI dépendent en apprentissage et en prédiction de la sortie des classifieurs construits selon l'approche RPC.

Instances	a_1	a_2	$c_1\#c_2$	$c_1\#c_3$	$c_2\#c_3$	c_1	c_2	c_3
x_1	20	30	$\emptyset$	$\emptyset$	$\emptyset$	0	0	0
x_2	35	35	$\emptyset$	$\emptyset$	$\emptyset$	0	0	0
x_3	15	40	c_2	$\emptyset$	c_2	0	1	0
x_4	20	50	c_2	$\emptyset$	c_2	0	1	0
x_5	30	45	$\emptyset$	c_1	c_2	1	1	0
x_6	35	30	$\emptyset$	c_1	c_2	1	1	0
x_7	10	40	c_1	$\emptyset$	c_3	1	0	1
x_8	15	45	c_1	$\emptyset$	c_3	1	0	1
x_9	25	55	c_1	$\emptyset$	c_3	1	0	1
x_{10}	30	60	c_1	$\emptyset$	c_3	1	0	1

Tableau 2.5: Exemple de données d'apprentissage pour l'approche Stacked_RPC_PSI

Instances	a_1	a_2	$c_1\#c_2$	$c_1\#c_3$	$c_2\#c_3$	c_1	c_2	c_3
x_1	20	30	c_2	c_1	c_2	0	0	0
x_2	35	35	c_1	c_1	c_2	0	0	0
x_3	15	40	c_2	c_1	c_2	0	1	0
x_4	20	50	c_2	c_1	c_2	0	1	0
x_5	30	45	c_1	c_1	c_2	1	1	0
x_6	35	30	c_1	c_1	c_2	1	1	0
x_7	10	40	c_1	c_1	c_3	1	0	1
x_8	15	45	c_1	c_1	c_3	1	0	1
x_9	25	55	c_1	c_1	c_3	1	0	1
x_{10}	30	60	c_1	c_1	c_3	1	0	1

Tableau 2.6: Données d'apprentissage mises à jour pour l'approche Stacked_RPC_PSI après l'introduction des prédictions des classifieurs H_{c_1,c_2}, H_{c_1,c_3}, et H_{c_2,c_3}

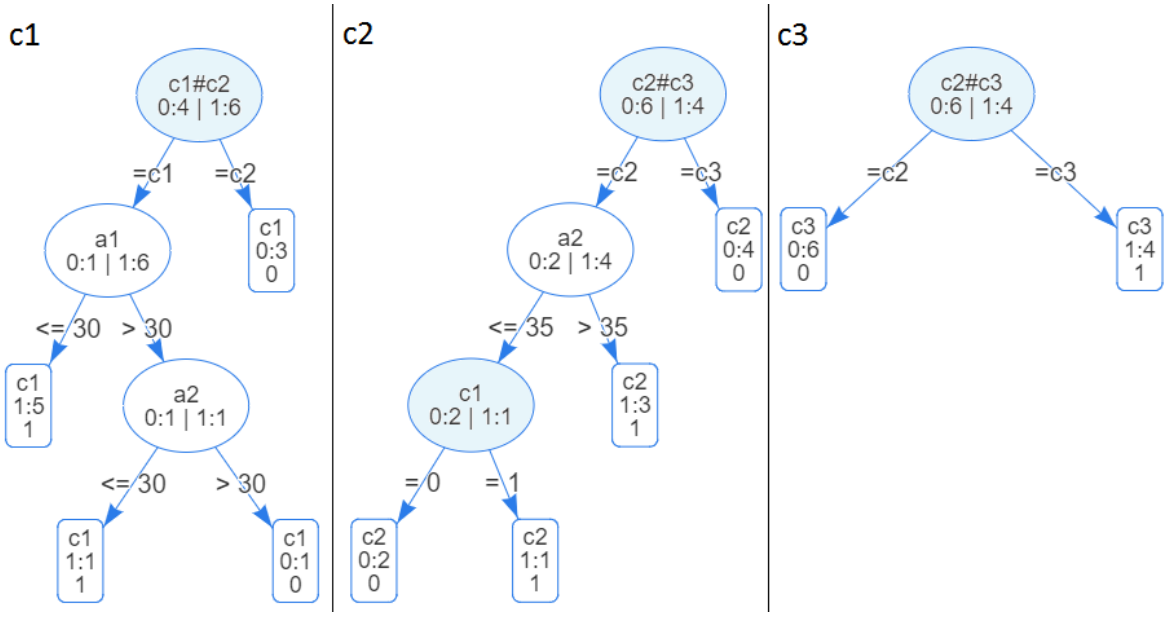


Figure 2.17: L'ensemble d'arbres de décision construits par l'approche Stacked_RPC_PSI selon l'approche PSI à partir des données d'apprentissage du Tableau 2.6

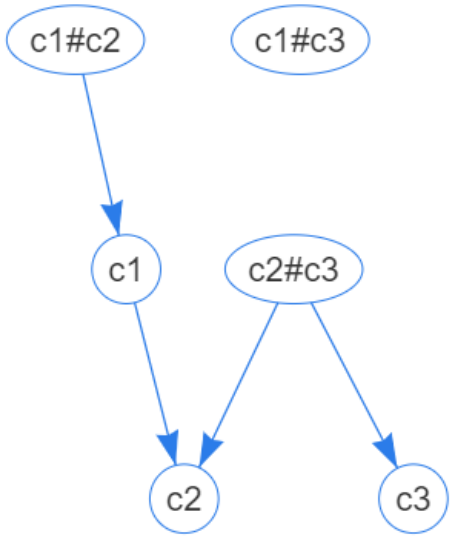


Figure 2.18: Structure de dépendances correspondant aux classifieurs construits par l'approche Stacked_RPC_PSI à partir des données d'apprentissage du Tableau 2.6

2.5 Expérimentation

L'étude expérimentale est menée sur trois jeux de données multi-labels appartenant à des domaines différents. Le Tableau 2.7 présente une description des données utilisées en utilisant les mesures définies dans la Section 1.2.2.

Chaque ensemble de données est ordonné aléatoirement puis il est réparti en 10 sous-ensembles pour faire une validation croisée. L'ordonnancement est effectué 10 fois pour faire 10 validations croisées de type 10 plis (10 times 10 fold cross validation).

Afin de comparer l'approche PSI2 aux autres approches de classification multi-labels, il faut d'abord sélectionner les bonnes valeurs des paramètres *DR* et *CL* de l'approche PSI2.

Donnée	Domaine	Instances	Attribues	Labels	$\mathbb{LC}$	$\mathbb{LD}$	$\mathbb{DLC}$
emotions	music	593	72	6	1.869	0.311	27
scenes	image	2407	294	6	1.074	0.179	15
yeast	biology	2417	103	14	4.237	0.303	198

Tableau 2.7: Données multi-labels réelles

2.5.1 Sélection des paramètres des approches PSI et PSI2

Le classifieur de base utilisé est un arbre de décision basé sur la mesure de l'entropie pour la sélection des attributs (Quinlan,1993,[89]). L'élagage de l'arbre de décision est effectué par rapport à trois paramètres adaptés aux jeux de données:

- la profondeur maximale de l'arbre de décision est 20
- la proportion des instances associées à la classe majoritaire est supérieure à 0.9
- le nombre minimal d'instances par nœud est 10

Les mesures utilisées pour les approches PSI et PSI2 sont les mêmes et sont données par la première ligne du Tableau **??**. L'évaluation d'autres mesures PSI est planifiée pour les travaux à venir. La sélection des paramètres *DR* et *CL* pour l'approche PSI2 est faite en évaluant l'impact de différentes valeurs sur la performance en prédiction, le risque de la propagation d'erreur de prédiction, et la complexité du modèle appris.

La première configuration des paramètres de l'approche PSI2 utilisée pour les trois jeux de données est $(DR = [0,19], CL = 20)$. $DR = [0,19]$ correspond au plus grand intervalle de profondeur où les labels peuvent être considérés en tant qu'attributs supplémentaires. $CL = 20$ correspond au nombre maximal de nœuds labels qui peuvent apparaitre dans la même branche d'un arbre de décision. Cette configuration de paramètres $(DR = [0,19], CL = 20)$ correspond à l'approche PSI qui n'impose aucune restriction sur l'apprentissage des dépendances entre les labels. Le nombre maximal de nœuds

labels observés dans une même branche d'un arbre de décision est $CL_{obs} = 4$ pour le jeu de données 'emotions', $CL_{obs} = 5$ pour le jeu de données 'scenes', et $CL_{obs} = 7$ pour le jeu de données 'yeast'. L'approche PSI2 est ensuite appliquée sur les trois jeux de données en combinant toutes les valeurs $CL \in [\![0, CL_{obs}]\!]$ avec les valeurs $DR \in \{[0,4],[0,9],[0,19],[5,9],[5,14],[10,19]\}$. L'objectif est d'étudier à la fois l'impact du nombre de nœuds labels dans une même branche de l'arbre de décision, et l'impact de la profondeur des nœuds labels dans l'arbre de décision (proches de la racine, au milieu, ou proches des feuilles).

L'approche PSI2 est évaluée pour chaque configuration de paramètres (DR, CL) par rapport à la performance en prédiction, par rapport au risque de la propagation d'erreur de prédiction, par rapport au nombre de dépendances apprises, et par rapport à la complexité du modèle appris. Les résultats des différents paramètres sont classés par rapport à chaque mesure d'évaluation, puis un classement moyen est calculé afin de sélectionner les meilleurs paramètres.

Les valeurs $\boldsymbol{DR = [0,9]}$ et $\boldsymbol{DR = [0,19]}$ combinées avec les valeurs $\boldsymbol{CL = 2}$ ou $\boldsymbol{CL = CL_{obs}}$ constituent les meilleurs configurations selon l'évaluation décrite dans le protocole ci-dessus. En effet, l'apparition d'un nœud label dans un arbre de décision indique que son pouvoir discriminant est plus grand que celui des attributs d'origine. Il faut donc autoriser les nœuds labels à partir de la racine $DR \in \{[0,4],[0,9],[0,19]\}$ pour discriminer les données plus rapidement et construire des arbres de décision avec moins de nœuds. L'utilité de l'intervalle $DR = [0,4]$ est réduite pour les trois jeux de données utilisées parce que la plupart des nœuds labels apparaissent dans l'intervalle $DR = [0-9]$. D'une part l'autorisation de plusieurs nœuds labels dans la même branche $CL = CL_{obs}$ permet de mieux discriminer les données et réduire le nombre de nœuds de l'arbre de décision. D'une autre part l'autorisation d'un maximum de deux nœuds labels dans la même branche $CL = 2$ permet de réduire le risque de la propagation de l'erreur en prédiction.

2.5.2 Évaluation des approches PSI, PSI2, CLR_PSI, et Stacked_RPC_PSI sur des données multi-labels

Évaluation de la prédiction

Les nouvelles approches proposées PSI, PSI2, CLR_PSI et Stacked_RPC_PSI sont comparées aux approches AID (Montañés et al.,2011,[79]), BR (Tsoumakas and Katakis,2007,[107]), CC (Read et al.,2011,[94]), et CLR (Fürnkranz et al.,2008,[37]) (Section 1.2.4) en utilisant plusieurs mesures d'évaluation (Section 1.2.2).

Le Tableau 2.8 présente les résultats de l'évaluation des prédictions pour les données 'emotions', 'scenes', et 'yeast'. Une flèche dirigée vers le bas $\downarrow$ indique que la mesure d'évaluation correspondante doit être minimisée. Une flèche dirigée vers le haut $\uparrow$ indique que la mesure d'évaluation correspondante doit être maximisée.

L'approche AID permet d'apprendre des relations de dépendance entre les labels, cependant les prédictions ne sont pas nécessairement cohérentes avec les dépendances apprises. L'approche AID four-

nit de meilleurs résultats que l'approche BR pour les données 'emotions' et 'scenes' (Tableau 2.8). Cependant, pour les données 'yeast' l'approche BR qui ne permet pas d'apprendre les relations entre les labels fournit de meilleurs résultats que l'approche AID. Ceci peut être expliqué par le fait que le nombre de labels disponible dans les données 'yeast' est plus grand que pour les données 'emotions' et 'scenes'. L'approche AID apprend donc plus de dépendance pour les données 'yeast' qui ne sont pas nécessairement prises en compte en prédiction, ce qui cause la dégradation de la performance en prédiction.

Selon les résultats collectés à l'étape de la sélection des valeurs pour les paramètres de l'approche PSI2 (Section 2.5.1), nous pouvons conclure que l'approche PSI correspond à l'approche PSI2 avec les paramètres $(DR = [0,19], CL = 4)$ pour les données 'emotions', $(DR = [0,19], CL = 5)$ pour les données 'scenes', et $(DR = [0,19], CL = 7)$ pour les données 'yeast'.

Le Tableau 2.8 montre que les approches PSI et PSI2 sont légèrement plus performantes que l'approche CC pour les données 'emotions', et elles sont aussi performantes que l'approche CC pour les données 'scenes', et 'yeast'.
Les paramètres $DR = [0,9]$ et $DR = [0,19]$ pour l'approche PSI2 fournissent des résultats similaires parce que la plupart des nœuds labels apparaissent dans l'intervalle $[0,9]$. Les paramètres $CL = 2$ et $CL = 7$ donnent des résultats similaires pour l'approche PSI2 parce que le nombre de nœuds labels qui apparaissent dans une même branche dépasse rarement 2 pour les données 'emotions', 'scenes', et 'yeast'.

L'approche CLR est basée sur l'apprentissage de préférences et sur un mécanisme de votes pour la sélection des labels à prédire. Plus le nombre de labels est grand plus le nombre de préférences apprises est grand, et plus la prédiction par vote majoritaire est fiable. Ceci explique le fait que l'approche CLR fournit les meilleurs résultats pour les données 'yeast'.

Les approches CLR_PSI et Stacked_RPC_PSI combinent l'apprentissage des relations de dépendance et des relations de préférence. Les résultats du Tableau 2.8 montrent que cette combinaison permet effectivement d'améliorer les résultats de prédiction pour les données 'emotions' et 'yeast'.

L'approche Stacked_RPC_PSI fournit de meilleurs résultats que les approches AID, BR, CC, PSI, et PSI2 pour les données 'yeast' parce qu'elle exploite les relations de préférence entre les labels. Cependant l'approche Stacked_RPC_PSI n'utilise pas le mécanisme de votes pour la prédiction. Ceci explique le fait que l'approche RPC_PSI qui utilise ce mécanisme de vote est plus performante pour les données 'yeast'. Les résultats de prédiction montrent que l'approche RPC_PSI est moins performante par rapport à l'approche CLR pour les données 'yeast' à cause de la propagation d'erreur de prédiction due aux relations de dépendance entre les labels.

Complexité des classifieurs et relations entre les labels

La complexité des classifieurs pour l'approche AID n'est pas évaluée parce qu'elle est basée sur deux couches de classifieurs. La complexité pour les approches CLR, CLR_PSI et Stacked_RPS_PSI n'est pas évaluée car ces approches sont basées sur $\frac{(k-1)k}{2}$ classifieurs. Elles construisent naturelle-

ment plus de classifieurs que les approches BR,CC, et PSI, et PSI2 basées sur seulement k classifieurs.

Le Tableau 2.9 montre que les approches PSI et PSI2 permettent d'apprendre plus de relations entre les labels que l'approche CC. Cependant, l'approche CC construit moins de nœuds pour les données 'yeast' que les approches PSI et PSI2. En effet, dans certains cas l'approche CC est obligée de faire d'un nœud une feuille de décision alors que les approches PSI et PSI2 ont la possibilité d'ajouter des nœuds permettant de mieux identifier les labels. Ceci est dû au fait que les approches PSI et PSI2 utilisent un ensemble d'attributs supplémentaires plus large que celui de l'approche CC.

Le Tableau 2.10 montre que les dépendances entre les labels augmentent le nombre d'erreurs de prédiction propagées (err-label $\Rightarrow$ err-decision). Cependant, le nombre des erreurs de prédiction non propagées augmente aussi (err-label $\Rightarrow$ not-err-decision).

Approche	CRHS	FMEASURE	GMEAN	EM	HL	PRECISION	RECALL	ACC−
AID	0.46	0.55	0.60	0.19	0.24	0.58	0.58	0.84
BR	0.45	0.54	0.60	0.18	0.24	0.59	0.56	0.85
CC	0.46	0.55	0.60	0.21	0.25	0.59	0.56	0.85
PSI	0.48	0.56	0.61	**0.25**	0.24	0.59	0.58	0.84
PSI2 [0,19] 2	0.47	0.55	0.59	0.23	0.25	0.58	0.56	0.84
PSI2 [0,9] 2	0.47	0.55	0.59	0.23	0.25	0.58	0.56	0.84
PSI2 [0,19] 4	0.48	0.56	0.61	**0.25**	0.24	0.59	0.58	0.84
PSI2 [0,9] 4	0.48	0.56	0.60	0.24	0.25	0.59	0.57	0.83
CLR	0.45	0.54	0.59	0.17	0.24	0.55	0.60	0.83
CLR_PSI	**0.49**	**0.58**	**0.64**	0.20	0.24	0.59	**0.64**	0.83
Stacked_RPC_PSI	0.48	0.57	0.62	0.24	**0.23**	**0.60**	0.59	**0.86**

Evaluation de la prédiction pour les données 'emotions'

Approach	CRHS	FMEASURE	GMEAN	EM	HL	PRECISION	RECALL	ACC−
AID	0.55	0.58	0.62	0.46	0.15	0.57	0.63	0.91
BR	0.51	0.54	0.56	0.44	**0.12**	0.53	0.57	**0.95**
CC	0.57	0.59	0.60	0.54	0.13	0.60	0.59	0.93
PSI	0.57	0.59	0.60	0.54	0.13	0.59	0.59	0.93
PSI2 [0,19] 2	0.56	0.58	0.60	0.53	0.13	0.58	0.58	0.93
PSI2 [0,9] 2	0.56	0.58	0.60	0.53	0.13	0.58	0.58	0.93
PSI2 [0,19] 5	0.57	0.59	0.60	0.54	0.13	0.59	0.59	0.93
PSI2 [0,9] 5	0.57	0.59	0.60	0.54	0.13	0.59	0.59	0.93
CLR	0.50	0.53	0.58	0.40	0.13	0.51	0.59	0.94
CLR_PSI	**0.60**	**0.63**	**0.67**	0.50	**0.12**	0.62	**0.68**	0.93
Stacked_RPC_PSI	**0.60**	0.62	0.63	**0.56**	**0.12**	**0.63**	0.62	0.94

Evaluation de la prédiction pour les données 'scenes'

Approach	CRHS	FMEASURE	GMEAN	EM	HL	PRECISION	RECALL	ACC−
AID	0.40	0.53	0.63	0.06	0.27	0.56	0.55	0.81
BR	0.42	0.54	0.63	0.06	0.25	0.60	0.55	0.85
CC	0.43	0.52	0.59	**0.16**	0.24	0.60	0.52	0.87
PSI	0.43	0.53	0.60	0.15	0.26	0.58	0.53	0.84
PSI2 [0,19] 2	0.43	0.53	0.60	0.15	0.25	0.58	0.54	0.85
PSI2 [0,9] 2	0.42	0.52	0.59	0.13	0.25	0.58	0.53	0.85
PSI2 [0,19] 7	0.43	0.53	0.60	0.15	0.26	0.58	0.53	0.84
PSI2 [0,9] 7	0.43	0.53	0.61	0.15	0.25	0.58	0.54	0.84
CLR	**0.47**	**0.59**	**0.66**	0.10	**0.21**	**0.67**	**0.59**	**0.88**
CLR_PSI	**0.47**	0.58	0.65	0.12	0.22	**0.67**	0.58	**0.88**
Stacked_RPC_PSI	0.44	0.54	0.61	0.13	0.23	0.62	0.53	**0.88**

Evaluation de la prédiction pour les données 'yeast'

Tableau 2.8: Évaluation de la prédiction pour les données 'emotions', 'scenes', et 'yeast'

Données	Approche	nœuds ↓	Feuilles ↓	Profondeur ↓	Dépendances ↑	Dépendances par branche ↑
emotions	BR	48.59	24.80	6.92	0.00	0.00
	CC	44.49	22.75	**6.49**	1.18	0.91
	PSI2 $[0,19]$ 2	44.24	22.62	6.92	1.12	0.97
	PSI2 $[0,9]$ 2	44.24	22.62	6.92	1.12	0.97
	PSI2 $[0,19]$ 4	**43.70**	**22.35**	6.94	**1.29**	**1.07**
	PSI2 $[0,9]$ 4	43.71	22.36	6.94	1.28	**1.07**
scenes	BR	59.60	30.30	6.63	0.00	0.00
	CC	**40.54**	**20.77**	5.93	1.68	1.08
	PSI2 $[0,19]$ 2	43.23	22.11	5.96	1.20	0.90
	PSI2 $[0,9]$ 2	43.23	22.11	5.96	1.20	0.90
	PSI2 $[0,19]$ 5	40.69	20.85	**5.71**	**1.92**	**1.19**
	PSI2 $[0,9]$ 5	40.70	20.85	**5.71**	**1.92**	**1.19**
yeast	BR	180.13	90.57	8.78	0.00	0.00
	CC	**85.86**	**43.43**	**6.26**	2.18	**1.71**
	PSI2 $[0,19]$ 2	99.54	50.27	6.50	1.67	1.31
	PSI2 $[0,9]$ 2	99.54	50.27	6.50	1.67	1.31
	PSI2 $[0,19]$ 7	94.17	47.59	6.35	**2.62**	1.64
	PSI2 $[0,9]$ 7	94.26	47.63	6.35	2.58	1.64

Tableau 2.9: Evaluation de la complexité du modèle appris et des dépendances entre les labels pour les données 'emotions', 'scenes', et 'yeast'

Données	Approche	REP ↓	ULR ↑	err-label ⇒ err-decision ↓	err-label ⇒ not-err-decision ↑	not-err-label ⇒ err-decision ↓	not-err-label ⇒ not-err-decision ↑
emotions	CC	0.38	**0.76**	26.70	43.69	**49.00**	151.20
	PSI2 $[0,19]$ 2	**0.36**	0.75	**26.51**	**46.24**	55.67	**163.70**
	PSI2 $[0,9]$ 2	**0.36**	0.75	**26.51**	**46.24**	55.67	**163.70**
	PSI2 $[0,19]$ 4	0.40	0.72	29.35	43.38	62.52	160.10
	PSI2 $[0,9]$ 4	0.40	0.72	29.32	43.31	62.46	160.03
scenes	CC	**0.57**	0.72	65.73	**49.26**	211.85	**536.06**
	PSI2 $[0,19]$ 2	0.58	**0.78**	**53.01**	39.51	**116.42**	399.14
	PSI2 $[0,9]$ 2	0.58	**0.78**	**53.01**	39.51	**116.42**	399.14
	PSI2 $[0,19]$ 5	0.59	0.72	66.30	46.41	203.16	528.25
	PSI2 $[0,9]$ 5	0.59	0.72	66.29	46.41	203.05	528.28
yeast	CC	0.54	**0.79**	612.46	530.61	614.57	2247.40
	PSI2 $[0,19]$ 2	**0.53**	**0.79**	**576.80**	515.19	**537.87**	2038.52
	PSI2 $[0,9]$ 2	**0.53**	**0.79**	**576.80**	515.19	**537.87**	2038.52
	PSI2 $[0,19]$ 7	**0.53**	0.76	664.74	605.06	707.44	**2267.61**
	PSI2 $[0,9]$ 7	**0.53**	0.76	666.31	**606.59**	706.40	2264.15

Tableau 2.10: Evaluation de la propagation de l'erreur de prédiction pour les données 'emotions', 'scenes' et 'yeast'

2.5.3 Évaluation des approches PSI, CLR_PSI, et Stacked_RPC_PSI sur des données multi-labels graduées

Les approches Vertical_BR, et Compltete_BR correspondent aux décompositions verticale et complète où les classifieurs générés sont appris de façon indépendante en utilisant l'approche BR. Les approches Vertical_PSI et Compltete_PSI correspondent au cas où les classifieurs générés sont appris selon l'approche PSI. Les approches Horizontal_CLR, Full_CLR, et Joined_CLR correspondent au cas où l'approche CLR est utilisée pour construire les classifieurs permettant l'apprentissage de préférence entre les labels. Les approches Horizontal_CLR_PSI, Full_CLR_PSI, et Joined_CLR_PSI et les approches Horizontal_Stacked_RPC_PSI, Full_Stakced_RPC_PSI, et Joined_Stacked_RPC_PSI correspondent respectivement au cas où l'approche CLR_PSI et au cas où l'approche Stacked_RPC_PSI est utilisée pour apprendre les classifieurs générés par la décomposition.

Description de données

L'expérimentation est menée sur un ensemble de 1930 instances multi-labels graduées appelé 'BelaE' (Abele-Brehm and Stief,2004,[1]). Chaque instance représente un étudiant décrit par deux attribut: l'âge et le genre. Chaque étudiant affecte des degrés d'importance à 48 propriétés d'emploi. Le degré d'importance de chaque propriété d'emploi varie selon la vision de chaque étudiant pour son futur emploi. Seules les dix dernières propriétés sont considérées en tant que labels. Le reste des propriétés d'emploi sont considérées en tant qu'attributs descriptifs des étudiants (Cheng et al.,2010,[25]). Les degrés d'importance disponibles correspondent aux valeurs $\{0,1,2,3,4\}$ où la valeur 0 représente une propriété d'emploi complètement insignifiante pour l'étudiant. Le degré moyen (Average Grade) associé à une propriété d'emploi dans le jeu de données 'BelaE' est $\mathbb{AG} = 2.66$. La distribution de chaque degré d'association m_g (nombre de labels associés aux instance avec ce degré) notée $DG(m_g)$ est illustrée dans le Tableau 2.11.

Résultats et discusion

La tâche correspondante à chaque degré d'association est une tâche de classification multi-labels. Les approches de classification multi-labels graduée peuvent être évaluées en utilisant les mesures d'évaluation de la classification multi-labels pour chaque degré d'association (Tableau 2.12).

Les résultats du Tableau 2.12 montrent que les approches Vertical_PSI et Complete_PSI ne sont pas assez performantes par rapport aux autres approches. Ceci est dû au problème de la propagation d'erreur de prédiction. En effet, la décomposition de la classification multi-labels graduée génère plusieurs classifieurs ce qui augmente le risque de la propagation d'erreur entre les classifieurs dépendants.

Les approches de décomposition complète et horizontale permettent d'obtenir plus de prédictions exactes que les autres décompositions (la mesure $\mathbb{EM}$ dans le Tableau 2.12). Ceci est expliqué par la stratégie de prédiction utilisée. Par exemple, même si un label n'est pas prédit dans la tâche de

classification multi-labels pour les degrés ≥ 1 il aura des chances d'être prédit dans les autres tâches de classification correspondant aux degrés $\geq 2, \geq 3 \geq 4$. Lorsque les prédictions sont évaluées, un label associé à l'instance avec un degré ≥ 2 ou ≥ 3 ou ≥ 4 est considéré comme associé à l'instance avec un degré ≥ 1. Ceci explique la dégradation des résultats pour le degré d'association ≥ 4 parce qu'il n'y a pas une autre tâche de classification qui compense les erreurs de prédiction à ce niveau.

Les approches Full_CLR_PSI, Full_Stacked_RPC_PSI et les approches Joined_CLR_PSI et Joined_Stacked_RPC_PSI fournissent des résultats compétitives avec les approches de base correspondantes Full_CLR et Joined_CLR. Ceci confirme l'hypothèse que la combinaison des relations de préférence et de dépendance peut améliorer les prédictions dans le cas de la classification multi-labels graduée.

Instances	Attributs	Labels	Degrés	$\mathbb{AG}$	$\mathbb{DG}(0)$	$\mathbb{DG}(1)$	$\mathbb{DG}(2)$	$\mathbb{DG}(3)$	$\mathbb{DG}(4)$
1930	40	10	5	2.66	0.0379	0.1044	0.2490	0.3786	0.2302

Tableau 2.11: Description des données multi-labels graduées 'BelaE'

Degré	Approache	CRHS	FMEASURE	GMEAN	EM	HL	PRECISION	RECALL	ACC−
	Vertical_BR	0.95	0.97	0.98	0.66	0.05	0.97	0.98	0.98
	Vertical_PSI	0.95	0.97	0.98	0.66	0.05	0.97	0.98	0.98
	Complete_BR	0.97	0.98	0.99	0.74	0.03	0.97	0.99	0.98
	Complete_PSI	0.96	0.98	0.98	0.70	0.04	0.97	0.98	0.98
	Full_CLR	0.94	0.96	0.96	0.53	0.06	0.98	0.96	0.97
	Full_CLR_PSI	0.94	0.96	0.96	0.54	0.06	0.98	0.96	0.97
≥ 1	Full_Stacked_RPC_PSI	0.95	0.97	0.97	0.66	0.05	0.97	0.98	0.98
	Horizontal_CLR	0.96	0.98	0.99	0.74	0.03	0.97	0.99	0.98
	Horizontal_CLR_PSI	0.96	0.98	0.99	0.74	0.03	0.97	0.99	0.98
	Horizontal_Stacked_RPC_PSI	0.96	0.98	0.99	0.74	0.03	0.97	0.99	0.99
	Joined_CLR	0.89	0.94	0.94	0.29	0.10	0.98	0.91	0.97
	Joined_CLR_PSI	0.90	0.95	0.94	0.34	0.09	0.98	0.92	0.98
	Joined_Stacked_RPC_PSI	0.95	0.97	0.98	0.65	0.05	0.97	0.98	0.98
	Vertical_BR	0.84	0.90	0.87	0.24	0.15	0.90	0.91	0.86
	Vertical_PSI	0.84	0.90	0.87	0.24	0.15	0.90	0.91	0.86
	Complete_BR	0.87	0.93	0.89	0.33	0.12	0.90	0.96	0.86
	Complete_PSI	0.85	0.91	0.89	0.28	0.14	0.91	0.93	0.87
	Full_CLR	0.75	0.84	0.82	0.13	0.22	0.94	0.80	0.88
	Full_CLR_PSI	0.75	0.84	0.82	0.14	0.22	0.94	0.80	0.88
≥ 2	Full_Stacked_RPC_PSI	0.84	0.91	0.88	0.23	0.14	0.91	0.92	0.86
	Horizontal_CLR	0.87	0.93	0.90	0.33	0.12	0.91	0.96	0.86
	Horizontal_CLR_PSI	0.87	0.93	0.90	0.32	0.12	0.91	0.96	0.87
	Horizontal_Stacked_RPC_PSI	0.87	0.92	0.89	0.31	0.12	0.90	0.96	0.86
	Joined_CLR	0.84	0.91	0.87	0.21	0.14	0.93	0.89	0.87
	Joined_CLR_PSI	0.85	0.91	0.87	0.24	0.14	0.92	0.91	0.86
	Joined_Stacked_RPC_PSI	0.84	0.91	0.88	0.23	0.15	0.91	0.92	0.86
	Vertical_BR	0.61	0.74	0.70	0.04	0.29	0.75	0.76	0.68
	Vertical_PSI	0.61	0.73	0.70	0.04	0.29	0.75	0.76	0.69
	Complete_BR	0.65	0.77	0.72	0.05	0.26	0.76	0.80	0.68
	Complete_PSI	0.62	0.75	0.71	0.04	0.27	0.76	0.77	0.70
	Full_CLR	0.31	0.41	0.44	0.02	0.43	0.66	0.34	0.93
	Full_CLR_PSI	0.34	0.44	0.46	0.02	0.42	0.67	0.39	0.90
≥ 3	Full_Stacked_RPC_PSI	0.61	0.73	0.70	0.04	0.29	0.74	0.76	0.68
	Horizontal_CLR	0.63	0.75	0.72	0.05	0.26	0.78	0.78	0.71
	Horizontal_CLR_PSI	0.63	0.75	0.72	0.06	0.26	0.78	0.78	0.71
	Horizontal_Stacked_RPC_PSI	0.64	0.76	0.72	0.05	0.27	0.77	0.79	0.69
	Joined_CLR	0.68	0.80	0.72	0.07	0.25	0.74	0.90	0.60
	Joined_CLR_PSI	0.68	0.80	0.71	0.07	0.25	0.72	0.92	0.59
	Joined_Stacked_RPC_PSI	0.61	0.74	0.70	0.03	0.29	0.75	0.77	0.68
	Vertical_BR	0.26	0.34	0.40	0.12	0.23	0.37	0.37	0.84
	Vertical_PSI	0.26	0.33	0.39	0.11	0.23	0.37	0.36	0.84
	Complete_BR	0.27	0.34	0.39	0.15	0.21	0.39	0.35	0.87
	Complete_PSI	0.26	0.34	0.39	0.13	0.22	0.38	0.35	0.86
	Full_CLR	0.04	0.05	0.05	0.18	0.23	0.06	0.05	0.98
	Full_CLR_PSI	0.06	0.07	0.08	0.19	0.22	0.08	0.07	0.98
≥ 4	Horizontal_CLR	0.24	0.31	0.35	0.18	0.19	0.39	0.30	0.92
	Horizontal_CLR_PSI	0.24	0.31	0.35	0.18	0.19	0.37	0.30	0.92
	Horizontal_Stacked_RPC_PSI	0.24	0.31	0.35	0.15	0.20	0.38	0.30	0.91
	Joined_CLR	0.30	0.41	0.50	0.00	0.45	0.31	0.73	0.46
	Joined_CLR_PSI	0.29	0.40	0.48	0.00	0.48	0.30	0.75	0.42
	Joined_Stacked_RPC_PSI	0.27	0.35	0.40	0.11	0.23	0.38	0.37	0.85

Tableau 2.12: Evaluation de la prédiction pour les données 'BelaE' pour chaque degré d'association

2.6 Conslusion

Apprendre les dépendances entre les labels en limitant le risque de la propagation de l'erreur de prédiction est une tâche intéressante de la classification multi-labels. Les approches existantes permettant l'apprentissage des dépendances entre les labels gèrent le problème des dépendances cycliques soit en imposant une structure de dépendance au préalable, soit en combinant plusieurs couches de classifieurs ce qui risque selon le cas, de produire des prédictions incohérentes avec les dépendances apprises (Section 1.2.4).

Dans ce chapitre nous proposons l'approche PSI qui permet d'apprendre les dépendances entre les labels en utilisant un seul ensemble de classifieurs et sans restriction au préalable sur la structure de dépendance. Les éventuelles dépendances cycliques sont ensuite éliminées en remplaçant certains classifieurs en se basant sur trois mesures appelées pré-sélection, sélection, et intérêt de chaînage. L'approche PSI a deux inconvénients: le premier est que certains classifieurs impliqués dans une dépendance cyclique sont reconstruits. Le deuxième inconvénient est que l'approche PSI permet d'apprendre certaines dépendances faibles qui ne font que créer des dépendances cycliques et augmenter le risque de la propagation d'erreur de prédiction.

Nous avons introduit l'approche PSI2 qui permet de remédier aux inconvénients de l'approche PSI mais elle est compatible uniquement avec les classifieurs binaires de type arbre de décision ou règles de décision. L'avantage d'utiliser un arbre de décision (qui peut être convertit en règles de décision) est qu'il suffit de reconstruire le sous-arbre à partir du nœud label qui cause la dépendance cyclique. Un autre avantage de l'arbre de décision est qu'il est possible de distinguer les dépendances fortes (nœuds labels proches de la racine) et les dépendances faibles (nœuds labels proches des feuilles). L'approche PSI2 est basée sur le paramètre *DR* permettant de fixer l'intervalle de profondeur où les dépendances peuvent être apprises, et le paramètre *CL* permettant de limiter le nombre de nœuds labels maximal dans une même branche de l'arbre de décision afin de réduire le risque de la propagation d'erreur de la prédiction.

La comparaison des approches PSI et PSI2 avec des approches existantes montre qu'elles donnent des résultats compétitifs en terme de performance de prédiction et de dépendances apprises. L'avantage des approches PSI et PSI2 est qu'elle permettent d'apprendre des ensembles différents de dépendances entre les labels. En effet, le fait de varier les heuristiques des trois mesures PSI ou les valeurs des deux paramètres *DR* et *CL* de l'approche PSI2 permet de varier les dépendances que l'approche peut apprendre.

Nous avons introduit les deux approches CLR_PSI et Stacked_RPC_PSI qui permettent d'apprendre à la fois les relations de dépendance et de préférence entre les labels. Les résultats d'expérimentation sur trois jeux de données réelles montrent que ces deux combinaisons permettent effectivement d'améliorer les résultats de prédiction.

D'une part, l'expérimentation sur des données multi-labels graduées montre que la propagation de l'erreur de prédiction pour les approches basées sur les dépendances entre les labels est un incon-

vénient majeur. En effet la classification multi-labels graduée est décomposée en plusieurs sous-tâches de classification mono-label ou multi-labels. Ceci augmente le risque de la propagation de l'erreur de prédiction entre les classifieurs dépendants qui peuvent être nombreux, et par la suite ceci risque d'aboutir à une croissance des erreurs de prédiction.

D'autre part, l'expérimentation sur des données multi-labels graduées montre l'intérêt de la combinaison des relations de préférence et de dépendance pour améliorer les résultats de prédiction. Les approches de classification multi-labels graduée basées sur les deux approches CLR_PSI et Stacked_RPC_PSI permettent d'obtenir des résultats très compétitifs par rapport aux approches de l'état de l'art.

Chapitre 3

Systèmes de recommandation et classification multi-labels graduée

Les données multi-labels graduées sont caractérisées par le fait que chaque instance peut être associée à un ou plusieurs labels moyennant des degrés d'association graduels. Par exemple, un internaute peut exprimer son degré de satisfaction par rapport aux films qu'il a regardé moyennant l'échelle de degrés de satisfaction des cinqs étoiles: {Avatar ⋆⋆⋆⋆, Star Wars ⋆⋆⋆, Titanic ⋆⋆}. Les plateformes qui collectent ce type de données intègrent un système de recommandation permettant d'apprendre les préférences des utilisateurs pour leur recommander les produits ou les services qui peuvent les intéresser. Les systèmes de recommandation doivent relever plusieurs défis pour assurer de bonnes recommandations. En effet, les utilisateurs ne s'enregistrent pas tous en même temps, et ils ne peuvent pas exprimer leurs degrés de satisfaction en même temps. Le système de recommandation doit se mettre à jour continuellement pour tenir en compte les nouvelles données collectées. Un utilisateur ne peut pas exprimer son degré de satisfaction par rapport à tous les produits ou les services disponibles. Le système de recommandation doit gérer ce problème de faible densité des données. Un utilisateur peut, dans certains cas, changer d'avis: par exemple, un utilisateur qui a regardé seulement le film 'Titanic' lui affecte une évaluation de quatre étoiles, mais après avoir regardé les films 'Star Wars' et 'Avatar' son évaluation du film 'Titanic' peut l'amener à changer en deux étoiles. Le système de recommandation doit être sensible aux changements de préférence des utilisateurs pour améliorer ses recommandations. Les approches de classification multi-labels graduée existantes ne peuvent pas être utilisées directement pour élaborer un système de recommandation. En effet, elles ne sont pas adaptées aux défis des données arrivant continuellement, des données à faible densité, et des préférences

évolutives des utilisateurs. Ce travail propose une nouvelle approche permettant de relever ces défis afin d'élaborer un système de recommandation basé sur la classification multi-labels graduée. L'approche proposée est évaluée sur deux jeux de données servant à recommander des films et des anecdotes. Les résultats d'expérimentation montrent que l'approche proposée est compétitive avec les approches existantes des systèmes de recommandation.

3.1 Introduction

Les données multi-labels graduées sont caractérisées par le fait que chaque instance est décrite par des attributs descriptifs, et peut être associée à un ou plusieurs labels moyennant des degrés d'association graduels.

Par exemple, un internaute (instance) est décrit par son âge et sa fonction (attributs descriptifs) peut exprimer son degré de satisfaction par rapport aux films (labels) qu'il a regardé moyennant l'échelle de degrés de satisfaction d'une à cinq étoiles. Par exemple, l'ensemble multi-labels gradué {Avatar ⋆⋆⋆⋆, Star Wars ⋆⋆⋆, Titanic ⋆⋆} correspond à l'évaluation de plusieurs film par le même utilisateur. Les données multi-labels graduées peuvent être traitées dans le sens opposé: par exemple, un film (instance) décrit par sa durée et son genre, peut être évalué par plusieurs internautes (labels).
Par exemple, l'ensemble multi-labels gradué {Alice ⋆⋆, Bob ⋆⋆⋆⋆, Carol ⋆⋆⋆} correspond à l'évaluation par plusieurs utilisateurs (Alice, Bob, et Carol) d'un même film. D'une façon générale, chaque utilisateur peut évaluer un ensemble d'items (produis, services, films, jeux, etc) moyennant une échelle graduelle de degrés de satisfaction. La relation entre les utilisateurs et les items peut être représentée par une matrice dite matrice d'évaluation (Figure 3.1).

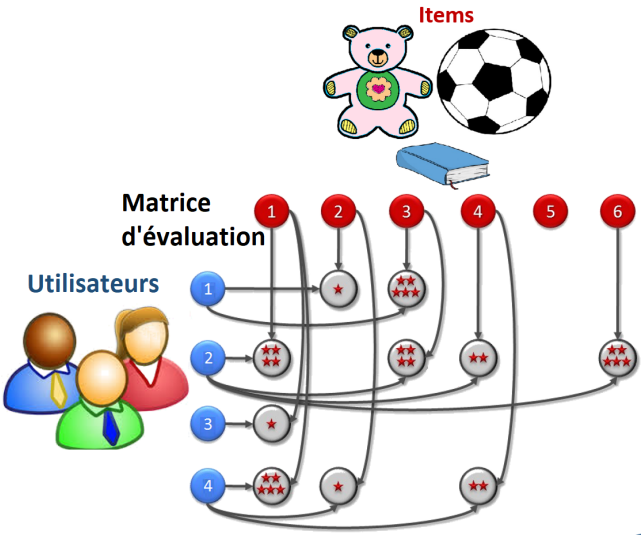


Figure 3.1: Exemple d'une matrice d'évaluation

La classification multi-labels graduée est la tâche d'apprendre un classifieur à partir des données

multi-labels graduées permettant de prédire, pour chaque instance, les labels qui lui sont associés avec les degrés d'association correspondants (Cheng et al.,2010,[25]). La source la plus riche en données multi-labels graduées est le web. En effet, plusieurs plateformes web offrent aux internautes la possibilité d'évaluer des produits ou des services (items) en utilisant une échelle graduelle de degrés de satisfaction. Ainsi, les sites de streaming permettent aux utilisateurs d'évaluer les films selon l'échelle de satisfaction d'une à cinq étoiles. Les évaluations des utilisateurs constituent une base d'apprentissage pouvant être exploitée pour prédire les items qui intéressent chaque utilisateur. Cette base peut être enrichie par des informations sur les utilisateurs et des informations sur les items. La tâche d'apprentissage et de recommandation des items est effectuée par un système de recommandation (Bobadilla et al.,2013,[15]).

D'un côté, un classifieur multi-labels gradué permet de prédire des labels avec leurs degrés d'association correspondants. D'un autre côté, un système de recommandation permet de recommander une liste des items (labels) ayant la meilleure évaluation prédite (plus grand degré d'association). Un système de recommandation effectue donc une tâche supplémentaire par rapport à un classifieur multi-labels gradué consistant à ordonner les items selon l'évaluation prédite. L'objectif est de recommander à un utilisateur uniquement les items ayant de bonnes évaluations prédites. Cependant, un système de recommandation ne peut pas être basé directement sur un classifieur multi-labels gradué. En effet, les données collectées par un système de recommandation présentent plusieurs défis non traités par les classifieurs multi-labels gradués existants:

- le grand volume de données: le nombre des utilisateurs, le nombre des items, et le nombre des évaluations des items peuvent être très grands ;
- l'incrémentalité des données: des nouveaux utilisateurs, des nouveaux items, et des nouvelles évaluations d'items peuvent être collectées à tout moment ;
- le changement de concept (concept drift): les préférences des utilisateurs ne sont pas nécessairement définitives, et peuvent changer au cours du temps (Haque et al.,2016,[45]) ;
- la faible densité des données: chaque utilisateur ne peut pas évaluer qu'un sous ensemble d'items.

Un classifieur multi-labels gradué effectue l'étape d'apprentissage sur un ensemble d'instances décrites par des attributs descriptifs et associées à des labels avec des degrés d'association. Une importante particularité des données traitées par les systèmes de recommandation est que même les labels (utilisateurs ou items) peuvent avoir des attributs descriptifs. Deux approches intuitives permettent de construire un classifieur multi-labels gradué en apprenant sur des données ayant des attributs descriptifs pour les labels:

- considérer les utilisateurs en tant qu'instances, et exploiter les attributs descriptifs des utilisateurs pour prédire les degrés d'association aux items considérés en tant que labels ;

- considérer les items en tant qu'instances, et exploiter les attributs descriptifs des items pour prédire les degrés d'association aux utilisateurs considérés en tant que labels.

Les deux approches précédentes peuvent être combinées en apprenant deux classifieurs multi-labels gradués et en agrégeant leurs prédictions pour fournir une prédiction finale.

3.2 Systèmes de recommandation

3.2.1 Description formelle de la tâche de recommandation

Soit $U = \{u_e\}_{1 \leq e \leq n}$ un ensemble d'utilisateurs. Soit $A = \{a_j\}_{1 \leq j \leq p}$ un ensemble d'attributs descriptifs des utilisateurs. Chaque utilisateur u_e est un vecteur de valeurs d'attributs descriptifs:
$u_e = (u_{e,a_1}, \ldots, u_{e,a_p}) = (u_{e,a_j})_{1 \leq j \leq p}$.
Soit $I = \{i_f\}_{1 \leq f \leq m}$ un ensemble d'items. Soit $B = \{b_j\}_{1 \leq j \leq q}$ un ensemble d'attributs descriptifs des items. Chaque item i_f est un vecteur de valeurs d'attributs descriptifs:
$i_f = (i_{f,b_1}, \ldots, i_{f,b_q}) = (i_{f,b_j})_{1 \leq j \leq q}$.
Soit $D = \{d_g\}_{1 \leq g \leq s}$ l'ensemble des évaluations disponibles (ratings) tel que $d_1 < \ldots < d_s$.
Soit $R = (r_{u_e,i_f})_{\substack{1 \leq e \leq n \\ 1 \leq f \leq m}}$ la matrice d'évaluation (rating matrix) telle que la valeur se trouvant à l'intersection de la ligne e et la colonne f représente l'évaluation r_{u_e,i_f} de l'item i_f par l'utilisateur u_e. l'évaluation r_{u_e,i_f} est aussi appelée appréciation ou degré de satisfaction de l'utilisateur u_e par rapport à l'item i_f. Soit η le nombre souhaité d'items que le système de recommandation doit proposer à chaque utilisateur. Soit $\mathscr{P}(I)$ l'ensemble de sous ensembles d'items.
Un système de recommandation $\mathscr{R} : U \rightarrow \mathscr{P}(I)$ doit apprendre à partir de l'ensemble d'utilisateurs, de l'ensemble d'items, et de la matrice d'évaluation (U, I, R) à recommander pour chaque utilisateur $u_e \in U$ un ensemble d'au plus η items $(\mathscr{R}(u_e) \subseteq I, |\mathscr{R}(u_e)| \leq \eta)$ ayant le plus de chance d'intéresser l'utilisateur u_e (être bien évalués par l'utilisateur u_e). La tâche de recommander pour un utilisateur un sous ensemble d'items est souvent appelée filtrage dans l'état de l'art.

3.2.2 Approches d'élaboration des systèmes de recommandation

Les systèmes de recommandation sont utilisés pour exploiter les données collectées par les plateformes web afin de proposer à chaque utilisateur une liste d'items ayant le plus de chance de l'intéresser (Bobadilla et al.,2013,[15]). Les systèmes de recommandation peuvent être élaborés en se basant sur trois approches principales (Pazzani,1999,[85]):

- filtrage basé sur les attributs descriptifs des utilisateurs (demographic based filtering): le système de recommandation apprend à partir des attributs descriptifs des utilisateurs (age, genre, ville, ...) à prédire l'évaluation des items pour chaque utilisateur (Wang et al.,2012,[118]) ;

- filtrage basé sur les attributs descriptifs des items (content based filtering): le système de recommandation apprend à partir des attributs descriptifs des items (prix, catégorie, année de production, ...) à prédire l'évaluation des utilisateurs pour chaque item (Lops et al.,2011,[71]) ;

- filtrage collaboratif (collaborative filtering): le système de recommandation prédit l'évaluation d'un item par un utilisateur en se basant sur les évaluations d'autres utilisateurs qui ont évalué des items en commun (Yang et al.,2014,[123]).

3.3 Systèmes de recommandations basés sur la classification

3.3.1 Modes d'apprentissage des classifieurs

L'apprentissage par lot (batch learning) est le mode d'apprentissage usuel pour les classifieurs. Il consiste à apprendre un classifieur H à partir des données d'un ensemble d'apprentissage X. L'avantage de ce mode d'apprentissage est que le classifieur tient en compte toutes les données d'apprentissage. Cependant, ce mode d'apprentissage présente des inconvénients majeurs pour le cas des systèmes de recommandation:

- il faut reconstruire le classifieur régulièrement par ce qu'il ne peut pas être mis à jour lorsque l'ensemble d'apprentissage est enrichi par de nouvelles données ;

- les limites des ressources matérielles (notamment en terme de mémoire et de performance du processeur) rendent ce mode d'apprentissage inefficace en cas de données de grand volume (big data).

L'apprentissage séquentiel (sequential learning) est le cas où le classifieur reçoit les données une par une ou morceau par morceau (chunk by chunk) jusqu'à ce qu'un critère d'arrêt est rencontré (Betancourt,2014,[12]). Une étape d'apprentissage supplémentaire est nécessaire pour rendre le classifieur opérationnel pour fournir des prédictions. L'apprentissage séquentiel nécessite moins de ressources que l'apprentissage par lot. Cependant, le classifieur n'est pas utilisable en prédiction avant l'étape supplémentaire d'apprentissage.

L'apprentissage incrémental (incremental learning) est un apprentissage séquentiel permettant d'utiliser le classifieur en prédiction à tout moment (Gepperth and Hammer,2016,[41]). Le classifieur est mis à jour et devient prêt pour fournir des prédiction à chaque fois qu'une donnée ou un ensemble de données est reçu. L'apprentissage incrémental est donc adapté au cas des systèmes de recommandation où les données arrivent continuellement, et où les prédictions doivent être fournies en temps réel.

3.3.2 Arbres de décision incrémentales

Les arbres de décision font partie des algorithmes d'apprentissage les plus utilisés parce qu'ils sont faciles à implémenter et à interpréter (Quinlan,1993,[89];Breiman et al.,1984,[19]). Plusieurs extensions ont été proposées pour permettre la construction des arbres de décision de façon incrémentale.

L'algorithme ID4 (Schlimmer and Fisher,1986,[97]) permet de construire un arbre de décision incrémentale en adaptant l'algorithme de base ID3 (Quinlan,1986,[88]). Un arbre de décision est construit initialement. Chaque nouvelle instance injectée dans l'arbre de décision est gardée dans la feuille correspondante. L'entropie est recalculée à partir des feuilles mises à jour en remontant vers la racine. Lorsque le gain d'entropie au niveau d'un nœud n'est pas suffisant, le sous-arbre de décision correspondant est reconstruit. L'inconvénient de l'algorithme ID4 est qu'il risque de reconstruire plusieurs fois les sous-arbres de décision en cas de changements de concept récurrents.

L'algorithme ID5 (Utgoff,1988,[112]) est une amélioration de ID4 qui consiste à fusionner les sous-arbres de décision ou à ajouter de nouveaux nœuds au lieu de les reconstruire. Cependant l'arbre de décision appris par l'algorithme incrémental ID4 ou ID5 n'est pas nécessairement le même appris par l'algorithme d'apprentissage par lot ID3.

L'algorithme ID5R (Utgoff,1989,[113]) permet de construire les mêmes arbres de décision que l'algorithme ID3 en se basant sur des opérations récursives de mise à jour de l'arbre de décision. L'inconvénient de l'algorithme ID5R est qu'il peut être beaucoup plus lent que l'algorithme ID3.

L'algorithme ITI (Utgoff,1994,[114]) est une extension de ID5R intégrant la capactité de gérer les attributs numériques et les valeurs manquantes.

Les arbres de décision très rapides (very fast decision trees: VFDT) ne gardent aucune données en mémoire. Seules des informations statistiques sur les données sont gardées (Domingos and Hulten,2000,[31]). Chaque instance injectée dans l'arbre de décision participe à la mise à jour des statistiques de la feuille correspondante. Lorsqu'une feuille accumule suffisamment d'informations statistiques elle devient un nœud interne et de nouvelles feuilles sont créées (Hoeffding,1963,[49]). L'inconvénient de la méthode VFDT est qu'elle ne peut pas détécter un changement de concept nécessitant une reconstruction de l'arbre de décision. En effet, en cas de changement de concept la méthode VFDT ne peut que continuer à générer des nœuds dans l'arbre de décision ce qui peut causer un problème de sur-apprentissage.

La méthode CVFDT est une extension de la méthode VFDT permettant de détecter et de gérer les changements de concept (Hulten et al.,2001,[51]). L'idée est de commencer à construire un nouveau sous-arbre de décision dès que les statistiques au niveau d'un nœud rencontrent certains critères. Le nouveau sous-arbre de décision remplace le nœud correspondant lorsqu'il devient plus précis en terme de prédiction.

3.3.3 Gestion des changements de concept

Un changement de concept se produit lorsque la précision en prédiction du classifieur décroit à cause d'un changement dans les propriétés statistiques de l'attribut à prédire. Pour les systèmes de recommandation ceci se traduit par un changement dans les préférences des utilisateurs: par exemple, au fil du temps, les films fantastiques bien évalués par un utilisateur reçoivent de moins en moins de bonnes évaluations en faveur des films d'action. La préférence de cet utilisateur pour les films fantastiques s'est réorientée vers les films d'action en fonction de son âge. Le système de recommandation ne sera pas performant s'il ne détecte pas ce changement de préférence et continue à recommander des films fantastiques pour cet utilisateur.

Les fenêtre mobiles (sliding windows) font partie des approches les plus utilisées pour la gestion des changements de concept. L'idée est de garder uniquement un certain nombre de données récentes en oubliant la donnée la moins récente à l'arrivée d'une nouvelle donnée. Le classifieur peut détecter un changement de concept en utilisant deux fenêtres mobiles consécutives par rapport aux données les plus récentes. Un changement de concept est détecté lorsque les distributions des valeurs de l'attribut à prédire (évaluation d'un item) sont significativement différentes par rapport aux deux fenêtres mobiles. Les fenêtres mobiles utilisées peuvent avoir soit une taille fixe (Xioufis et al.,2011,[120]) soit une taille variable (Bifet and Gavaldà,2007,[14]) pour le nombre maximal de données dans la fenêtre.

Les approches de détection des changements de concept basées sur les fenêtres mobiles supposent que les données récentes sont plus pertinentes pour décrire la distribution actuelle des valeurs de l'attribut à prédire. Cependant, les données n'ont pas toutes la même importance pour un classifieur. Les données peuvent être pondérées selon l'impact de leurs suppression sur la précision en prédiction. Ensuite, les données à faible poids sont éliminées au lieu d'éliminer les données les moins récentes (Loeffel et al.,2016,[70]).

Les changements de concept peuvent être traités par des approches dites aveugles qui n'essayent pas de détecter un changement de concept ou de vérifier qu'il s'est produit. L'idée est d'utiliser une fonction affectant des poids aux données de façon à pénaliser les données moins récentes au lieu de les oublier brusquement (Cohen and Strauss,2006,[26]).

3.4 Une nouvelle approche pour les systèmes de recommandation

Dans cette thèse nous proposons de construire un système de recommandation basé sur la classification multi-labels graduée en utilisant une approche incrémentale d'apprentissage (Laghmari et al.,2017,[62]). Deux classifieurs multi-labels gradués sont construits afin de pouvoir exploiter à la fois les attributs descriptifs des utilisateurs et des items dans l'objectif de produire de bonnes recommandations:

- un classifieur noté H^U considérant les utilisateurs en tant qu'instances, et les items en tant que labels dont il faut prédire les degrés d'association ;

- un classifieur noté H^I considérant les items en tant qu'instances, et les utilisateurs en tant que labels dont il faut prédire les degrés d'association.

Dans la suite de cette section, nous détaillons notre stratégie de représentation des données permettant de traiter le fait que les données arrivent en flux (Section 3.4.1). Nous détaillons ensuite notre stratégie d'optimisation de l'accès au disque afin de gérer le fait que les données collectées sont larges et doivent être stockées dans le disque pour libérer la mémoire (Section 3.4.2). Nous discutons ensuite les paramètres concernant l'apprentissage incrémental que nous utilisons dans notre modèle prédictif (Section 3.4.3). Nous détaillons en fin la stratégie de décomposition de la classification multi-labels graduée que nous utilisons dans notre approche (Section 3.4.4).

3.4.1 Représentation des données

Les évaluations fournies par les utilisateurs, et les valeurs d'attributs descriptifs des utilisateurs et des items ne sont pas nécessairement disponibles au démarrage du système de recommandation. Elles peuvent être collectées plus tard de façon incrémentale. La représentation vectorielle des valeurs d'attributs descriptifs et la représentation matricielle des évaluations des utilisateurs ne sont pas efficaces en terme de ressources de mémoire nécessaires à cause des valeurs manquantes (faible densité des données). Une nouvelle représentation est nécessaire pour permettre au classifieur d'apprendre à partir de toutes les données collectées en optimisant les ressources nécessaires. L'idée est de regrouper les valeurs d'attributs descriptifs ainsi que les évaluations dans un même ensemble de triplets (e, f, v) tels que:

- un triplet composé d'une valeur positive e et d'une valeur positive f correspond à l'évaluation v donnée par l'utilisateur u_e à l'item i_f ;

- un triplet composé d'une valeur positive e et d'une valeur négative f correspond à la valeur v de l'attribut descriptif a_{-f} de l'utilisateur u_e ;

- un triplet composé d'une valeur négative e et d'une valeur positive f correspond à la valeur v de l'attribut descriptif b_{-e} de l'item i_f.

Chaque instance d'un utilisateur u_e n'est plus considérée comme un vecteur de valeurs d'attributs descriptifs, mais plutôt comme un ensemble de paires (f, v) correspondant aux valeurs d'attributs descriptifs de l'utilisateur ($f < 0$) et aux évaluations qu'il a fourni aux items ($f > 0$).

Chaque instance d'un item u_e n'est plus considérée comme un vecteur de valeurs d'attributs descriptifs, mais plutôt comme un ensemble de paires (e, v) correspondant aux valeurs d'attributs descriptifs de l'item ($e < 0$) et aux évaluations qu'il a reçu par les utilisateurs ($e > 0$).

3.4.2 Optimisation de l'accès au disque

De nouveaux utilisateurs, de nouveaux items, et de nouvelles évaluations sont collectées de façon incrémentale, cependant leurs nombre total peut être très grand pour les garder en mémoire. Un stockage sur le disque est donc nécessaire mais doit être optimisé pour la mise à jour ou l'ajout de données.

Par exemple l'ensemble des utilisateurs les plus actifs (qui évaluent souvent des items) et des items les plus actifs (qui sont souvent évalués par des utilisateurs) doit être gardé en mémoire pour gagner en temps de recherche et de mise à jour des données correspondantes.

Soit α le nombre des derniers utilisateurs actifs à garder en mémoire. L'ensemble de ces utilisateurs est noté $U_{[\alpha]}$.
Soit β le nombre des derniers items actifs à garder en mémoire. L'ensemble de ces items est noté $I_{[\beta]}$.

Les paramètres α and β doivent être maximisés selon la capacité de mémoire disponible afin de minimiser le nombre d'accès au disque.

Lorsqu'un utilisateur u_e donne pour un item i_f une évaluation r_{u_e,i_f}, un triplet (e,f,r_{u_e,i_f}) est généré puis il est sauvegardé sur le disque. Ensuite l'instance de l'utilisateur u_e et l'instance de l'item i_f doivent être mises à jour pour inclure l'évaluation ajoutée:

- $u_e \leftarrow u_e \cup \{(f, r_{u_e,i_f})\}$.

- $i_f \leftarrow i_f \cup \{(e, r_{u_e,i_f})\}$.

Les instances u_e et i_f sont d'abord cherchées en mémoire pour être mises à jour. Elles sont collectées à partir du disque dans le cas où elles ne font pas partie des instances actives gardées en mémoire (les derniers α utilisateurs actifs et les derniers β items actifs). Dans ce cas elles seront déjà à jour puisque l'information est sauvegardée d'abord dans le disque pour ne pas être perdue.

L'instance u_e correspond au dernier utilisateur actif et doit donc être placée en tête de la liste $U_{[\alpha]}$. Dans le cas où l'instance u_e se trouve déjà dans $U_{[\alpha]}$, elle est d'abord retirée de $U_{[\alpha]}$ puis elle est ajoutée en tête de la liste $U_{[\alpha]}$: $U_{[\alpha]} \leftarrow \{u_e\} \cup U_{[\alpha]}$.

Dans le cas où la liste $U_{[\alpha]}$ contient exactement α utilisateurs, celui en fin de la liste $U_{[\alpha]}$ est d'abord retiré. L'instance de l'utilisateur ayant la plus ancienne activité est donc oubliée au niveau de la mémoire mais elle est toujours gardée au niveau du disque.

Ce même traitement est appliqué à l'instance item i_f pour mettre à jour l'ensemble des derniers items actifs: $I_{[\beta]} \leftarrow \{i_f\} \cup I_{[\beta]}$.

L'instance de l'utilisateur u_e est injectée au classifieur H^U pour qu'il tienne en compte la nouvelle information reçue $\{f, r_{u_e,i_f}\}$. L'instance de l'item i_f est aussi injectée au classifieur H^I pour qu'il tienne compte de la nouvelle information reçue $\{e, r_{u_e,i_f}\}$.

3.4.3 Paramètres de l'apprentissage incrémental

Chaque instance injectée à un classifieur permet de tenir compte d'une nouvelle information ajoutée. En effet chaque instance est injectée à un classifieur autant de fois que de valeurs associées à l'instance. Cependant, les instances ne sont pas dupliquées au niveau du classifieur et seulement des mesures statistiques sur les données sont gardées en mémoire (par exemple, le nombre de données associées à chaque valeur d'appréciation disponible). Lorsque les mesures statistiques arrivent à un seuil spécifié, le classifieur est reconstruit en considérant les dernières instances actives gardées en mémoire. Trois seuils sont utilisés dans l'approche que nous proposons pour reconstruire un classifieur:

- le nombre minimal d'instances distinctes à recevoir Δ_{inst} ;
- la complexité maximale du classifieur Δ_{comp} (par exemple, le nombre de nœuds dans le cas d'un arbre de décision) ;
- l'évaluation maximale des erreurs de prédictions Δ_{pred}.

3.4.4 Spécification du classifieur multi-labels gradué

Le classifieur multi-labels gradué H^U est construit en utilisant une décomposition complète (Section 1.3.3.3) en $(|D| - 1) \times |I|$ classifieurs binaires $\{H^U_{\geq d_g, i_f}\}_{\substack{2 \leq g \leq |D| \\ 1 \leq f \leq |I|}}$ par rapport aux degrés d'association et par rapport aux labels (items).

Le classifieur multi-labels gradué H^I est construit en utilisant une décomposition complète en $(|D| - 1) \times |U|$ classifieurs binaires $\{H^I_{\geq d_g, u_e}\}_{\substack{2 \leq g \leq |D| \\ 1 \leq e \leq |U|}}$ par rapport aux degrés d'association et par rapport aux labels (utilisateurs).

Chaque classifieur binaire permet de prédire si le degré d'association à un label est supérieur ou égal à une valeur d_g ou non.

La prédiction de l'évaluation d'un utilisateur u_e pour un item i_f est obtenue en agrégeant les prédictions des classifieurs H^U et H^I. L'évaluation à prédire dans l'approche que nous proposons est la moyenne des évaluations prédites par H^U et H^I: $H^{U.I}(u_e\ ,\ i_f) = \dfrac{H^U(u_e) + H^I(i_f)}{2}$.

L'avantage de construire chacun des classifieurs H^U et H^I à partir d'un ensemble de classifieurs binaires est de pouvoir distribuer facilement sur un ensemble de serveurs. Chaque serveur peut s'occuper de l'apprentissage d'un sous ensemble de classifieurs binaires (Laghmari et al.,2017,[61]).

Le système de recommandation utilise les deux classifieurs H^U et H^I pour prédire les évaluations manquantes pour un utilisateur u_e, ensuite les η items ayant les meilleures évaluations prédites sont recommandés à l'utilisateur u_e.

3.5 Expérimentation

3.5.1 Données utilisées

Le jeu de données 'MovieLens' (Harper and Konstan,2015,[46]) est donné sous forme de trois fichiers: le premier contient 6040 utilisateurs avec leurs attributs descriptifs: genre, âge, occupation, et code postale. Le second fichier contient les informations de 3883 films: titre, année, et catégories du film. Le troisième fichier contient 1,000,209 évaluations numériques dans l'ensemble $\{1,2,3,4,5\}$.

Le jeu de données 'Jester' (Goldberg et al.,2001,[43]) est donné sous forme de $1,761,440$ évaluations dans l'intervalle $[-10,+10]$ de 150 anecdotes par $59,132$ utilisateurs. Aucune information sur les utilisateurs n'est disponible. Le texte des anecdotes n'a pas été converti en attributs descriptifs. Seules les évaluations sont donc disponibles pour ce jeux de données.

Les données multi-labels sont généralement décrites en utilisant des mesures telles que le nombre moyen de labels associés à une instance. L'approche que nous proposons est basée sur deux modèles de classification multi-labels graduée. Les mesures de description des données multi-labels sont illustrées pour chaque modèle dans la Tableau **??**

Données	Modèle	Instances	Attributs	Labels	Cardinalité de labels	densité de labels	combinaisons distinctes de labels
MovieLens	H^U	6040	4	3883	165.60	0.04	6040
MovieLens	H^I	3883	3	6040	269.89	0.04	3883
Jester	H^U	59132	0	150	29.79	0.20	24362
Jester	H^I	150	0	59132	12581.70	0.21	140

Tableau 3.1: Description des données

3.5.2 Méthodes d'évaluation de la classification incrémentale

Deux méthodes sont généralement utilisées dans la littérature pour évaluer les approches incrémentales (Gama et al.,2009,[39];Gama et al.,2013,[40]) en utilisant une fonction d'erreur (loss) $L : D \times D \rightarrow [0,1]$ permettant d'évaluer pour une instance x l'erreur entre la prédiction $H(x)$ et la valeur correcte y.

- la méthode 'holdout error' consiste à évaluer le même classifieur H sur un ensemble d'instances X qui ne font pas partie de l'ensemble d'apprentissage. Elle est donnée par $E_{hol}(X) = \frac{\sum_{x \in X} L(H(x),y)}{|X|}$;
- la méthode pré-séquentielle E_{pre} consiste à évaluer la prédiction du classifieur H pour toute instance avant de l'utiliser pour mettre à jour le classifieur H. En notant $H_{[n]}$ le classifieur H après avoir reçu n instances d'apprentissage $X = \{x_i\}_{1 \leq i \leq n}$. L'évaluation pré-séquentielle du classifieur $H_{[n]}$ est donnée par: $E_{pre}(X) = \frac{\sum_{i=1}^{n} L(H_{[i-1]}(x_i), y_i)}{n}$.

Les mesures d'évaluation multi-labels telles que la précision, le rappel et la FMesure ne sont pas adaptées aux données des systèmes de recommandation. En effet, elles permettent d'évaluer seulement la capacité du classifieur à prédire les bons labels. Cependant, pour un système de recommandation ce n'est pas intéressant de prédire si un item sera évalué par un utilisateur ou non, mais plutôt l'important est de prédire l'évaluation d'un item s'il est évalué par l'utilisateur. Les mesures d'évaluation mesurant la distance entre les prédictions et les vraies évaluations sont donc plus adaptées pour les systèmes de recommandation:

- l'erreur absolue (AE) est une mesure utilisée souvent dans la littérature pour évaluer les systèmes de recommandation (Ekstrand et al.,2011,[32]). Elle est donnée par $AE(H_{[i-1]}(x_i)\ ,\ y_i)\ =\ |H_{[i-1]}(x_i)\ -\ y_i|$;
- l'erreur de Hamming (HL) est une mesure utilisée aussi pour l'évaluation des systèmes de recommandation. Elle est donnée par $HL(H_{[i-1]}(x_i)\ ,\ y_i)\ =\ \frac{|H_{[i-1]}(x_i)\ -\ y_i|}{|m_s - m_1|}$.

3.5.3 Paramètres de l'approche proposée

L'approche proposée pour un système de recommandation basé sur la classification multi-labels graduée est implémentée en utilisant le langage C#. L'expérimentation est menée sur des machines de 2.2 GHz pour le CPU et de 4Go de mémoire RAM. Les paramètres $\alpha = 2000$, $\beta = 2000$, et $\eta = 1$ sont considérés.

La mesure MDL est utilisée pour décider si une feuille doit être convertie en un nœud interne ou non. La mesure MDL est évaluée dès qu'un nombre suffisant d'instance δ est reçu par le classifieur. Les expérimentations menées dans l'état de l'art ont montré que les valeurs $\delta = 100$ à $\delta = 1000$ instances permettent d'obtenir de bons résultats (Salperwyck and Lemaire,2013,[96]). Le seuil $\delta = 100$ est utilisé dans notre expérimentation.

Chaque arbre de décision binaire concerne un label et reçoit uniquement les instances associées à ce label. Les arbres de décision n'ont pas le même ensemble d'apprentissage et ne sont donc pas nécessairement construits au même moment. Cependant, par précaution une barrière de temps (walltime) de $0.5s$ est appliquée. Notre modèle sera donc forcé de fournir une prédiction même s'il y a quelques arbres de décision qui sont en cours de construction et qui n'ont pas fourni leurs prédictions.

3.5.3.1 Comparaison avec les approches de l'état de l'art

Notre approche est comparée à trois approches de base pour les systèmes de recommandation (Zaier et al.,2008,[126]):

- 'collaborative filtering': basée sur les évaluations faites par les utilisateurs similaires ;
- 'hybrid collaborative filtering': basée sur les attributs descriptifs des items, ainsi que les évaluations par les utilisateurs similaires ;

- 'demographic hybrid collaborative filtering': basée sur les attributs descriptifs utilisateurs, les attributs descriptifs des items, ainsi que les évaluations par les utilisateurs similaires.

Les trois approches ci-dessus sont basées sur la méthode des k plus proches voisins. La mesure de similarité utilisée pour identifier les voisins est le coefficient de corrélation de Pearson. Deux tailles de voisinage sont considérées $k = 50$ et $k = 950$ correspondant aux meilleurs et aux moins bons résultats reportés (Zaier et al.,2008,[126]).

Notre approche est comparée aussi à une approche récente de filtrage collaboratif appelée UO-CRBMF (Xie et al.,2016,[119]). Cette approche est basée sur les réseaux de neurones et tient compte des informations sur les utilisateurs. Les trois variantes disponibles pour cette approche sont appelées 'UO-CRBMF UserBased', 'UO-CRBMF ItemBased', et 'Hybrid UO-CRBMF'.

3.5.4 Résultats et discussion

La méthode pré-séquentielle est utilisée pour évaluer notre approche. L'évaluation de la mesure de Hamming est collectée pour chaque 1000 évaluations d'items. Les résultats pour les premiers $50,000$ évaluations d'items sont illustrés pour les deux jeux de données (Figure 3.2 et Figure 3.3). Les courbes d'évolution de la mesure de Hamming pour les deux jeux de données sont similaires. Ceci peut être expliqué par le fait que les mêmes paramètres ont été utilisés pour les deux jeux de données.

Le Tableau 3.2 et le Tableau 3.3 montrent que l'approche que nous proposons prédit rarement les évaluations minimale et maximale par rapport aux autres évaluations. Ceci s'explique par la méthode d'agrégation $H^{U.I}$ des classifieurs H^U et H^I qui consiste à prédire l'évaluation moyenne. Les évaluations minimale et maximale ne peuvent donc être prédites par $H^{U.I}$ que si elles sont prédites à la fois par H^U et H^I.

Les données 'MoviesLens' sont utilisées pour comparer notre approche aux approches existantes puisque les données 'Jester' ne disposent pas des informations sur les utilisateurs et les items.

Le Tableau 3.4 montre que notre approche notée 'GMLC based RS' est plus performante que les approches 'collaborative filtering', 'hybrid collaborative filtering' et 'demographic hybrid collaborative filtering'. Le Tableau 3.4 montre aussi que notre approche est très compétitive avec les trois variantes de l'approche 'UO-CRBMF' (une différence de 0.01 par rapport à la mesure de l'erreur de Hamming).

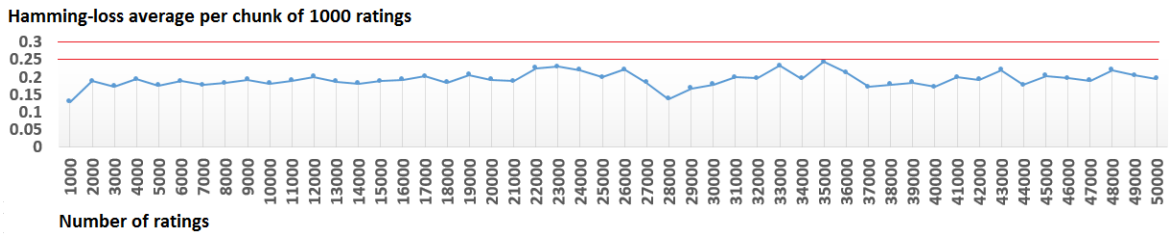


Figure 3.2: Evaluation de l'approche proposée sur les données 'MovieLens'

Evaluation prédite	nombre de fois	Hamming-loss
1	28	0.0179
1.5	109	0.1456
2	835	0.2030
2.5	2690	0.2465
3	5988	0.2059
3.5	13724	0.2160
4	18283	0.1717
4.5	7287	0.1852
5	1057	0.1055

Tableau 3.2: Évaluation détaillée de l'approche proposée sur les données 'MovieLens'

Evaluation prédite	nombre de fois	Hamming-loss
-10	15	0.0488
-9	15	0.0881
-8	27	0.1592
-7	385	0.1458
-6	798	0.1653
-5	1205	0.1832
-4	1864	0.2135
-3	2259	0.2131
-2	2893	0.2033
-1	4058	0.1972
0	5705	0.1844
1	7868	0.1746
2	8327	0.1795
3	6697	0.1960
4	4411	0.2127
5	2685	0.2230
6	696	0.1860
7	63	0.1986
8	30	0.1351

Tableau 3.3: Evaluation détaillée de l'approche proposée sur les données 'Jester'

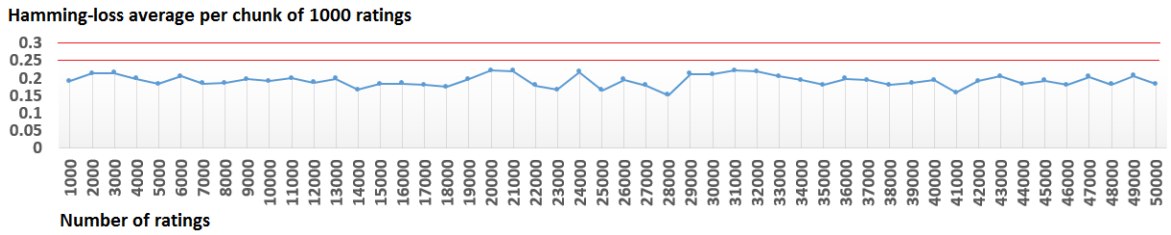


Figure 3.3: Evaluation de l'approche proposée sur les données 'Jester'

Approche	Hamming loss
GMLC based RS	0.19
Collaborative filtering - 50 Neighbours	0.23
Collaborative filtering - 950 Neighbours	0.20
Hybrid Collaborative filtering - 50 Neighbours	0.25
Hybrid Collaborative filtering - 950 Neighbours	0.20
Demographic hybrid Collaborative filtering - 50 Neighbours	0.24
Demographic hybrid Collaborative filtering - 950 Neighbours	0.20
UO-CRBMF UserBased	0.19
UO-CRBMF ItemBased	0.18
Hybrid CRBMF	0.18

Tableau 3.4: Comparaison de l'approche proposée avec des approches existantes sur les données 'MovieLens'

3.6 Conclusion

Ce travail introduit une nouvelle approche pour les systèmes de recommandation basée sur deux classifieurs multi-labels gradué. Le premier classifieur considère les utilisateurs en tant qu'instances et les items en tant que labels, et le deuxième classifieur considère les items en tant qu'instances et les utilisateurs en tant que labels. Chacun des classifieurs multi-labels gradués est construit à partir d'un ensemble de classifieurs binaires incrémentaux de type arbre de décision. L'approche que nous proposons peut être facilement distribuée en affectant à chaque serveur la tâche d'apprendre un sous ensemble de classifieurs binaires.

La mesure 'Minimum description length' (MDL) est évaluée régulièrement pour chaque arbre de décision afin de décider s'il faut étendre l'arbre de décision en ajoutant des nœuds ou non. Une fenêtre mobile des instances est utilisée pour reconstruire régulièrement l'arbre de décision afin de gérer les changements de concept.

Lorsqu'une nouvelle information (évaluation d'un item par un utilisateur, ou valeur d'un attribut descriptif pour un utilisateur ou un item) est collectée, l'instance de l'utilisateur ou de l'item correspondant est mise à jour en intégrant l'information reçue. Ensuite, l'instance est donnée en entrée au classifieur incrémental pour le mettre à jour. La partie qui consomme le plus de temps dans cette procédure est la recherche de l'instance correspondante sur le disque pour la mettre à jour. L'approche que nous

proposons essaye de minimiser le nombre d'accès au disque en gardant en mémoire l'ensemble des derniers utilisateurs actifs et des derniers items actifs.

L'approche que nous proposons est testée sur deux jeux de données. Les résultats de l'expérimentation montrent que l'approche que nous proposons permet de maintenir la mesure de l'erreur de Hamming à une valeur inférieure à 0.25 pour les deux jeux de données.

Chapitre 4

Reconnaissance des odeurs et classification multi-labels graduée

Les données multi-labels graduées peuvent être associées à plusieurs labels avec des degrés d'association appartenant à une échelle graduelle. Par exemple, une molécule odorante peut émettre à la fois une odeur forte herbacée, une odeur modérée de la rose, et une odeur faible de la menthe. La classification multi-labels graduée consiste à apprendre un modèle permettant de prédire pour chaque instance les labels qui lui sont associés avec leurs degrés d'association correspondants. Le modèle de classification établit un lien entre les variables descriptives d'une instance et l'ensemble gradué de labels qui lui sont associés. La particularité des données de molécules odorantes est qu'on sait déjà que les propriétés physico-chimiques des molécules ne permettent pas de bien discriminer les odeurs. Les approches de classification de base ne sont pas adaptées à ce défi présenté par les molécules odorantes. Ce travail introduit une nouvelle approche de classification basée sur l'apprentissage de relations entre les labels dans l'objectif de compenser le faible pouvoir discriminant des variables descriptives.

4.1 Introduction

Les molécules odorantes s'échappent dans l'air et sont captées par le nez. Des capteurs olfactifs réagissent aux molécules odorantes et émettent un signal au cerveau. Le cerveau reçoit le signal et l'interprète en une sensation d'odeur. Les capteurs associés à chaque odeur, et les propriétés des molécules odorantes qui activent chaque capteur ne sont pas complètement identifiés. En effet l'idée même que les propriétés des molécules puissent entièrement caractériser l'odeur reste une question ouverte. Des recherches récentes indiquent que les propriétés des molécules ne représentent que 30%

de l'identité de l'odeur (Bosc et al.,2016,[17]). Certaines molécules peuvent même émettre des odeurs totalement différentes selon leur concentration (de March et al.,2015,[28]).

L'une des motivations des travaux de recherche liées à l'olfaction est l'élaboration de nez électroniques (e-noses) pour reconnaître les odeurs. Un nez électronique est constitué d'un ensemble de capteurs et d'un système intelligent permettant de reconnaître certaines odeurs à partir des signaux émis par les capteurs (Patil and Kulkarni,2015,[84]). L'utilité des nez électroniques apparaît dans des domaines où l'utilisation du nez humain peut être inefficiente ou dangereuse: par exemple, l'identification du pain moisie par l'odeur (Estakhroueiyeh and Rashedi,2015,[34]), ou l'autodépistage du diabète à partir de l'odeur d'urine (Seesaard et al.,2016,[98]).

La partie intelligente du nez électronique est basée sur des approches d'apprentissage artificiel et statistique telles que les réseaux de neurones (Omatu et al.,2014,[83]) et l'analyse de composantes principales (Ansari et al.,2015,[7]). Le Tableau 4.1 présente une liste non exhaustive des approches appliquées à l'apprentissage de la reconnaissance d'odeurs ainsi que des informations sur les données d'apprentissage utilisées.

Ce travail s'intéresse au jeu de données marqué en gras dans le Tableau 4.1. Chaque instance de ces données représente une molécule qui peut avoir plusieurs odeurs graduées: par exemple, une molécule odorante qui émet à la fois une odeur forte herbacée, une odeur modérée de la rose, et une odeur faible de la menthe. Les odeurs 'herbacée', 'rose', et 'menthe' sont appelées qualités olfactives. Ce jeu de données est intéressant parce qu'il peut être exploité à la fois pour l'élaboration des nez électroniques, et pour la synthèse de nouveaux parfums sur mesure avec les nuances d'odeurs souhaitées. Ce jeu de données est dit multi-labels gradué parce que chaque instance peut être associée à plusieurs labels (qualités olfactives) avec des degrés d'association gradués (faible < modérée < forte).

La classification multi-labels graduée (classification MLG) consiste à apprendre à partir d'un jeu de données multi-labels gradué un modèle prédictif dit classifieur (Cheng et al.,2010,[25]). Le classifieur se base sur les variables descriptives des instances (propriétés physico-chimiques des molécules) pour prédire les labels (qualités olfactives) avec leurs degrés d'association correspondants. Cependant, les variables descriptives des molécules ne permettent pas de discriminer complètement les odeurs. Ceci réduit la fiabilité de la prédiction des qualités olfactives par les approches de base de la classification multi-labels graduée (Cheng et al.,2010,[25];Brinker et al.,2014,[21]).

Ce travail présente l'étude de la performance des approches de classification MLG sur le jeu de données des molécules odorantes, et des adaptations possibles qui peuvent améliorer les résultats des prédictions.

Références	Approches d'apprentissage	Données	Applications
Jha and Hayashi,2017,[52]	analyse de composantes principales, sélection d'attributs basée la corrélation	60 sujets	classification des odeurs du corps humain
Seesaard et al.,2016,[98]	analyse de composantes principales	7 sujets	santé et dépistage précoce, e-nose
Lanata et al.,2016,[67]	analyse de composantes principales, analyse discriminante linéaire	32 sujets	classification des odeurs agréables et néfastes
Rahman et al.,2015,[90]	analyse de composantes principales, analyse discriminante linéaire, k plus proches voisins, réseaux de neurones, machines à vecteur de support	90 échantillons, 3 classes	e-nose
Mahlke et al.,2007,[74]	analyse de composantes principales, k moyennes, clustering hiérarchique	155 composants de gaz résiduaires provenant des installations d'élevage, 25 composants de gaz provenant des grosses raffineries, 26 plantes de production du cacao et du café	traitement des gaz d'échappement
Chen et al.,2013,[24]	analyse discriminante linéaire, analyse multi-variées	51 échantillons, 3 types de thé: vert, noir, et Oolong	surveillance de la fermentation du thé, e-nose
Kroupi et al.,2016,[57]	analyse discriminante linéaire	25 sujets	introduction de l'olfaction aux multi-media
Daud et al.,2016,[27]	raisonnement par traitement des cas		identification du niveau de dégradation de l'huile de lubrification
Uçar and Özalp,2017,[110]	réseaux de neurones, k plus proches voisins, machines à vecteur de support	176 données, 6 odeurs de fruit	reconnaissance des odeurs fruitées, e-nose
Zhang and Deng,2017,[127]	réseaux de neurones	482 échantillons (Zhang and Zhang,2017,[128];Zhang et al.,2013,[129];Tian et al.,2016,[106])	e-nose
Bachtiar et al.,2011,[9] Bachtiar et al.,2014,[10] Bachtiar et al.,2016,[11]	réseaux de neurones	108 odeurs (Hallem and Carlson,2006,[44];Carey et al.,2010,[23])	e-nose
Omatu,2013,[82]	réseaux de neurones	6000 odeurs de thé, de café, et de cacao	e-nose
Omatu et al.,2014,[83]	réseaux de neurones	3331 échantillons, vin rouge et blanc	e-nose
Jha et al.,2016,[53]	réseaux de neurones, analyse de composantes principales	22 sujets	classification des odeurs du corps humain
Kissi et al.,2008,[56]	ensembles flous, arbres de décision	158 composants	reconnaissance de l'odeur du bois de sandal
Marsala et al.,1998,[75] Kissi et al.,2004,[55]	classification floue	99 molécules camphrées	reconnaissance de l'odeur du camphre
Ramdani et al.,2004,[91]	règles de décision floues, arbre de décision, algorithmes génétiques	71 molécules camphrées	reconnaissance de l'odeur du camphre
Sharma and Kumar,2016,[100]	classification basée sur l'entropie floue, réseaux de neurones	80 échantillons, 4 odeurs	e-nose
Bosc et al.,2015,[16] Bosc et al.,2016,[17]	découverte des sous-groupes locaux, règles de décision	**1600 molécules odorantes** (Arctander,1969,[8])	reconnaissance des qualités olfactives

Tableau 4.1: Tendances dans le domaine de la reconnaissance de l'odeur

4.2 Comparaison des approches de classification MLG sur les données des molécules odorantes

4.2.1 Préparation des données

La liste initiale des molécules odorantes que nous avons utilisé contient 1677 molécules odorantes (Arctander,1969,[8]). Chaque molécule peut émettre jusqu'à 7 odeurs ordonnées selon l'intensité. L'ensemble d'odeurs (qualité olfactives) présentes dans le jeu de données contient 80 odeurs différentes (étape (1) dans le Tableau 4.2). Les odeurs de moins de 20 occurrences ont été éliminées pour assurer l'obtention de résultats fiables par une validation croisée. Les molécules qui ne sont plus associées à une qualité olfactive après l'élimination des odeurs rares sont aussi éliminées (étape (2) dans le Tableau 4.2). Le logiciel Dragon (Mauri et al.,2006,[76]) est utilisé pour calculer les valeurs de 3839 propriétés physico-chimiques pour les molécules odorantes (étape (3) dans le Tableau 4.2). les propriétés physico-chimiques qui ont la même valeur pour plus de 0.99 des molécules sont éliminées (étape (4) dans le Tableau 4.2). Pour chaque molécule, un degré d'association dans l'ensemble $[\![0,7]\!]$ est affecté à chaque odeur tel que:

- les odeurs non émises par la molécule reçoivent le degré 0 ;
- l'odeur la plus forte reçoit le degré 7 et la plus faible reçoit le degré 1.

L'ensemble des données finalement obtenu est décrit dans le Tableau 4.3. La cardinalité moyenne d'odeurs associées à une molécule est 2.71. Le nombre de molécules associées à chaque cardinalité d'odeurs est présenté dans la Figure 4.1. Le nombre de molécules associées à chaque odeur est présenté dans la Figure 4.2.

Etape	Molécules	propriétés (attributs)	Odeurs (labels)
(1) initiale	1677	3	80
(2) éliminer les odeurs de cardinalité inférieur à 20	1656	3	52
(3) calcul des propriétés des molécules par le logiciel Dragon	1656	3842	52
(4) élimination des descripteurs pour lesquels plus de 0.99 des molécules ont la même valeur	1656	1928	52

Tableau 4.2: Préparation des données

Instances	Attributs	Labels	Grades	$\mathbb{AG}$	$\mathbb{DG}(0)$	$\mathbb{DG}(1)$	$\mathbb{DG}(2)$	$\mathbb{DG}(3)$	$\mathbb{DG}(4)$	$\mathbb{DG}(5)$	$\mathbb{DG}(6)$	$\mathbb{DG}(7)$
1623	1838	30	8	0.48	0.918	0.0001	0.0007	0.0022	0.0069	0.0152	0.0259	0.0311

Tableau 4.3: Description du jeu de données des molécules odorantes

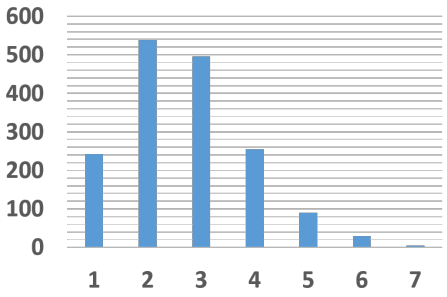


Figure 4.1: Nombre de molécules pour chaque cardinalité de labels associés

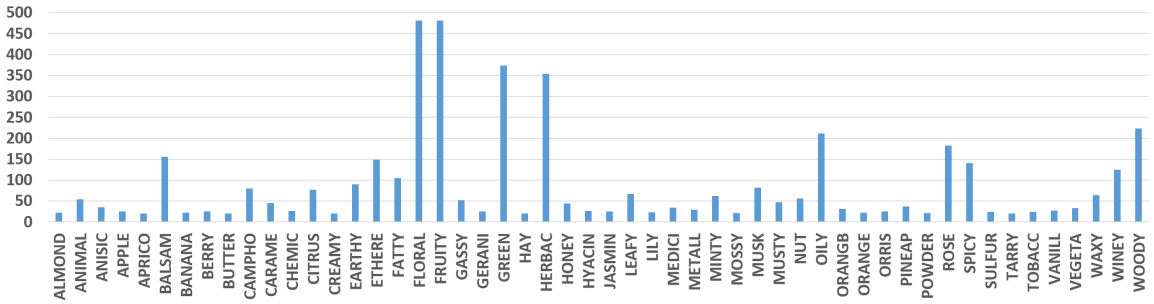


Figure 4.2: Nombre de molécules pour chaque odeur

4.2.2 Expérimentations préliminaires

L'expérimentation consiste à comparer les approches de classification multi-labels graduée de l'état de l'art : Vertical_BR, Complete_BR, Full_CLR, Horizontal_CLR, et Joined_CLR, avec les approches combinées introduites dans ce travail: Vertical_PSI, Complete_PSI, Full_CLR_PSI, Full_Stacked_RPC_PSI, Horizontal_CLR_PSI, Horizontal_Stacked_RPC_PSI, Joined_CLR_PSI, et Joined_Stacked_RPC_PSI (Section 2.5.3).

Les prédictions de chaque approche sont évaluées en utilisant les mesures d'évaluation de la classification multi-labels pour chaque sous-ensemble de labels ayant un degré d'association supérieur ou égal à un degré fixé (Tableau 4.4). La prédiction pour les dégrés d'association ≥ 1, ≥ 2, et ≥ 3 n'est pas présentée parce qu'elle est quasiment similaire au cas des degrés ≥ 4. Ceci est du au fait que les molécules ayant plus de 4 odeurs sont très rares (Figure 4.1). La moyenne des résultats par rapport à tous les degrés d'association est présentée dans le Tableau 4.5.

Les résultats obtenus montrent que la décomposition horizontale est celle qui offre les meilleurs résultats. Cependant, les résultats, en général, sont médiocres et confirment l'hypothèse que les variables descriptives des molécules ne permettent pas de reconnaître plus de 30% de l'identité des odeurs (Bosc et al.,2016,[17]).

Degré	Approche	CRHS	FMEASURE	GMEAN	EM	HL	PRECISION	RECALL	ACC−
	Vertical_BR	0.23	0.30	0.37	0.05	0.09	0.39	0.28	0.97
	Vertical_PSI	0.23	0.30	0.37	0.06	0.09	0.39	0.27	0.97
	Complete_BR	0.35	0.46	0.61	0.05	0.09	0.49	0.51	0.95
	Complete_PSI	0.22	0.29	0.34	0.04	0.08	0.39	0.26	0.98
	Full_CLR	0.25	0.31	0.36	0.09	0.07	0.39	0.29	0.98
	Full_CLR_PSI	0.25	0.31	0.36	0.10	0.07	0.40	0.29	0.98
≥ 4	Full_Stacked_CLR_PSI	0.27	0.35	0.43	0.08	0.08	0.43	0.34	0.97
	Horizontal_CLR	0.35	0.45	0.56	0.09	0.07	0.52	0.46	0.97
	Horizontal_CLR_PSI	0.33	0.44	0.57	0.09	0.08	0.49	0.47	0.96
	Horizontal_Stacked_CLR_PSI	0.34	0.45	0.58	0.07	0.08	0.49	0.48	0.96
	Joined_CLR	0.24	0.29	0.32	0.11	0.07	0.38	0.26	0.99
	Joined_CLR_PSI	0.25	0.32	0.38	0.09	0.08	0.40	0.31	0.98
	Joined_Stacked_CLR_PSI	0.25	0.32	0.41	0.06	0.08	0.39	0.32	0.97
	Vertical_BR	0.23	0.31	0.37	0.06	0.08	0.38	0.29	0.97
	Vertical_PSI	0.24	0.31	0.37	0.07	0.08	0.38	0.28	0.97
	Complete_BR	0.32	0.41	0.53	0.06	0.08	0.44	0.44	0.96
	Complete_PSI	0.23	0.30	0.35	0.05	0.08	0.39	0.26	0.98
	Full_CLR	0.21	0.26	0.28	0.10	0.06	0.33	0.23	0.99
	Full_CLR_PSI	0.24	0.29	0.33	0.10	0.07	0.36	0.27	0.99
≥ 5	Full_Stacked_CLR_PSI	0.27	0.35	0.43	0.09	0.08	0.42	0.35	0.97
	Horizontal_CLR	0.34	0.42	0.50	0.10	0.07	0.50	0.42	0.98
	Horizontal_CLR_PSI	0.35	0.43	0.52	0.15	0.07	0.50	0.43	0.97
	Horizontal_Stacked_CLR_PSI	0.33	0.43	0.54	0.09	0.07	0.49	0.45	0.96
	Joined_CLR	0.21	0.25	0.27	0.12	0.06	0.32	0.22	0.99
	Joined_CLR_PSI	0.25	0.31	0.35	0.11	0.07	0.39	0.29	0.98
	Joined_Stacked_CLR_PSI	0.25	0.33	0.41	0.07	0.08	0.38	0.34	0.97
	Vertical_BR	0.21	0.26	0.31	0.07	0.07	0.30	0.26	0.98
	Vertical_PSI	0.23	0.28	0.33	0.09	0.07	0.33	0.27	0.97
	Complete_BR	0.27	0.34	0.40	0.10	0.06	0.38	0.33	0.98
	Complete_PSI	0.22	0.27	0.31	0.08	0.06	0.33	0.25	0.98
	Full_CLR	0.17	0.20	0.21	0.10	0.05	0.24	0.18	0.99
	Full_CLR_PSI	0.20	0.23	0.26	0.10	0.06	0.28	0.22	0.99
≥ 6	Full_Stacked_CLR_PSI	0.25	0.31	0.37	0.10	0.07	0.36	0.31	0.97
	Horizontal_CLR	0.30	0.35	0.40	0.15	0.05	0.40	0.35	0.98
	Horizontal_Stacked_CLR_PSI	0.29	0.35	0.40	0.14	0.06	0.40	0.35	0.98
	Joined_CLR_PSI	0.22	0.26	0.29	0.10	0.06	0.31	0.25	0.99
	Horizontal_CLR_PSI	0.30	0.35	0.40	0.15	0.06	0.40	0.35	0.98
	Joined_Stacked_CLR_PSI	0.26	0.31	0.37	0.10	0.06	0.35	0.32	0.97
	Joined_CLR	0.17	0.19	0.19	0.10	0.05	0.22	0.17	1.00
	Vertical_BR	0.14	0.15	0.17	0.14	0.04	0.14	0.17	0.98
	Vertical_PSI	0.19	0.20	0.23	0.20	0.04	0.19	0.23	0.98
	Complete_BR	0.12	0.12	0.12	0.15	0.04	0.12	0.12	0.99
	Complete_PSI	0.19	0.20	0.20	0.20	0.04	0.19	0.20	0.98
	Full_CLR	0.09	0.09	0.09	0.12	0.03	0.09	0.09	1.00
	Full_CLR_PSI	0.15	0.15	0.15	0.19	0.03	0.15	0.15	0.99
≥ 7	Full_Stacked_CLR_PSI	0.22	0.23	0.23	0.23	0.04	0.22	0.23	0.98
	Horizontal_CLR	0.17	0.17	0.17	0.20	0.03	0.17	0.17	1.00
	Horizontal_CLR_PSI	0.16	0.17	0.17	0.18	0.03	0.16	0.17	0.99
	Horizontal_Stacked_CLR_PSI	0.15	0.15	0.15	0.18	0.04	0.15	0.15	0.99
	Joined_CLR	0.12	0.12	0.12	0.14	0.03	0.12	0.12	1.00
	Joined_CLR_PSI	0.19	0.20	0.21	0.20	0.03	0.19	0.21	0.99
	Joined_Stacked_CLR_PSI	0.20	0.21	0.22	0.23	0.04	0.20	0.22	0.98

Tableau 4.4: Evaluation de la prédiction des odeurs pour chaque sous-problème de classification multi-labels

Approche	CRHS	FMEASURE	GMEAN	EM	HL	PRECISION	RECALL	ACC−
Vertical_BR	0.21	0.26	0.31	0.07	0.07	0.30	0.26	0.98
Vertical_PSI	0.23	0.28	0.33	0.09	0.07	0.33	0.27	0.97
Complete_BR	0.27	0.34	0.40	0.10	0.06	0.38	0.33	0.98
Complete_PSI	0.22	0.27	0.31	0.08	0.06	0.33	0.25	0.98
Full_CLR	0.17	0.20	0.21	0.10	0.05	0.24	0.18	0.99
Full_CLR_PSI	0.20	0.23	0.26	0.10	0.06	0.28	0.22	0.99
Full_Stacked_CLR_PSI	0.25	0.31	0.37	0.10	0.07	0.36	0.31	0.97
Horizontal_CLR	0.30	0.35	0.40	0.15	0.05	0.40	0.35	0.98
Horizontal_CLR_PSI	0.30	0.35	0.40	0.15	0.06	0.40	0.35	0.98
Horizontal_Stacked_CLR_PSI	0.29	0.35	0.40	0.14	0.06	0.40	0.35	0.98
Joined_CLR	0.17	0.19	0.19	0.10	0.05	0.22	0.17	1.00
Joined_CLR_PSI	0.22	0.26	0.29	0.10	0.06	0.31	0.25	0.99
Joined_Stacked_CLR_PSI	0.26	0.31	0.37	0.10	0.06	0.35	0.32	0.97

Tableau 4.5: Evaluation de la prédiction en moyennant les résultats obtenus dans les sous-problèmes de classification multi-labels

4.3 Adaptation au jeu de données des molécules odorantes

Puisqu'il est difficile de reconnaître une odeur à partir des variables descriptives, l'idée dans ce travail est d'essayer de reconnaître un sous-groupe d'odeurs plus facile à discriminer par les variables descriptives.

Soit $\lambda : X \rightarrow \mathscr{P}(C)$ la fonction qui donne pour chaque instance l'ensemble de labels auxquels elle est associée. Soit $Supp(c_l) = \{x_i \in X, c_l \in \lambda(x_i)\}$ l'ensemble d'instances associées au label c_l appelé support de c_l. Soit $Co(c_l) = \{c_{l'} \in C, \exists x_i \in X : \{c_l, c_{l'}\} \subseteq \lambda(x_i)\}$ l'ensemble de labels ayant une relation de co-occurrence avec le label c_l.

L'idée de l'approche que nous proposons est de trouver pour chaque label c_l un ensemble minimal de labels $Co^*(c_l) \subseteq Co(c_l)$ dont le support inclut le support du label c_l: $Supp(Co^*(c_l)) \supseteq Supp(c_l)$. Un nouveau label $\mathscr{C}_l$ correspondant à $Co^*(c_l)$ est introduit. Un classifieur qui prédit la présence ou l'absence du label $\mathscr{C}_l$ est équivalent à un classifieur qui prédit la présence d'au moins un des labels dans $Co^*(c_l)$ ou l'absence de tous les labels dans $Co^*(c_l)$. L'avantage est que le nombre des instances associées à au moins un label dans $Co^*(c_l)$ est plus grand que le nombre des instances associées uniquement au label c_l. Le problème des classes déséquilibrées est donc réduit pour les nouveaux labels introduits.

Soit q le nombre maximal autorisé pour la cardinalité de l'ensemble $Co^*(c_l)$. Soit $\mathbb{C}$ un ensemble de labels initialement vide: $\mathbb{C} = \emptyset$. Soit $c_{l'}$ un label ayant une relation de co-occurrence avec le label c_l: $c_{l'} \in Co(c_l)$, et associé au maximum d'instances associées au label c_l: $\forall\, c_{l''} \in Co(c_l)$: $|Supp(c_{l'}) \cap Supp(c_l)| \geq |Supp(c_{l''}) \cap Supp(c_l)|$.

Le label $c_{l'}$ est ajouté à l'ensemble $\mathbb{C}$: $\mathbb{C} \leftarrow \mathbb{C} \cup \{c_{l'}\}$.

Le prochain label $c_{l''} \in Co(c_l)$ à ajouter dans $\mathbb{C}$ est sélectionné tel que: $\forall\, c_f \in Co(c_l)$: $|Supp(c_{l''}) \cap (Supp(c_l) - Supp(\mathbb{C}))| \geq |Supp(c_f) \cap (Supp(c_l) - Supp(\mathbb{C}))|$.

Les labels sont ajoutés itérativement à l'ensemble $\mathbb{C}$ jusqu'à ce que l'une des deux conditions suiv-

antes est vérifiée:

- $Supp(\mathbb{C}) \supseteq Supp(c_l)$;
- $|\mathbb{C}| = q$.

Dans le cas où l'ensemble $\mathbb{C}$ obtenu ne vérifie pas $(Supp(\mathbb{C}) \supseteq Supp(c_l))$, le dernier label ajouté à $\mathbb{C}$ est remplacé par le label c_l lui même.

En se basant sur l'heuristique que $Co^*(c_l)$ contient les labels qui ont le plus grand nombre de relations de co-occurrence avec le label c_l, l'ensemble construit $\mathbb{C}$ correspond à un ensemble minimal de labels associés à toutes les instances associées à c_l: $\mathbb{C} = Co^*(c_l)$.

L'approche que nous proposons est constituée de deux étapes:

- la première étape consiste à apprendre un ensemble de classifieurs $\{H_{\mathscr{C}_l}\}_{1 \leq l \leq k}$. Chaque classifieur $H_{\mathscr{C}_l}$ permet de prédire l'absence ou la présence du label $\mathscr{C}_l$ en se basant uniquement sur les attributs descriptifs.
- la deuxième étape consiste à apprendre un ensemble de classifieurs $\{H_{c_l}\}_{1 \leq l \leq k}$. Chaque classifieur H_{c_l} permet de prédire l'absence ou la présence du label c_l en se basant sur les prédictions des classifieurs $\{H_{\mathscr{C}_l}\}_{1 \leq l \leq k}$ et des classifieurs $\{H_{c_{l'}}\}_{c_{l'} \neq c_l}$ en utilisant l'approche PSI (Section 2.1).

4.4 Expérimentation et résultats

La sous-tâche de classification multi-labels graduée correspondant à la prédiction des labels qui ont un degré d'association au moins supérieur ou égal à 1 est équivalente à la classification multi-labels en ignorant les degrés d'association. Dans ce travail nous essayons d'abord d'améliorer la prédiction des labels associés avant d'améliorer la prédiction des degrés d'association des labels associés aux instances.

La qualité des prédictions est évaluée en utilisant la F-measure (Section 1.2.2) et en ignorant l'information sur les degrés d'association. Une validation croisée de 10 plis est appliquée sur le jeu de données des molécules odorantes. La F-measure est évaluée pour chaque pli, puis la valeur moyenne par rapport aux 10 plis est reportée. La prédiction des odeurs par l'approche directe BR (Section 1.2.4.5) est comparée à la prédiction des odeurs par l'approche que nous proposons notée PSI^*.

Le nombre maximal choisi pour la taille des sous-ensembles de labels à générer est $q = 7$. La cardinalité de chaque sous-ensemble de labels généré est illustrée dans la Figure 4.3.

La Figure 4.4 illustre les résultats obtenus pour chaque qualité olfactive pour les deux approches BR et PSI^*. L'approche PSI^* permet de prédire 7 qualités olfactives que l'approche BR ne

permet pas du tout de prédire $\{BANANA, BERY, BUTTER, CHEMIC, LILY, MOSSY, MUSTY\}$. Cependant, l'approche PSI^* n'a pas maintenu la capacité de prédire 4 qualités olfactives $\{ANESIC, HONEY, MEDICI, METALL\}$. La performance en prédiction pour les qualités olfactives $\{MUSK, VANILL\}$ est clairement réduite pour l'approche PSI^*. Cependant, l'approche PSI^* donne de meilleures résultats par rapport à l'approche BR pour presque toutes les autres qualités olfactives.

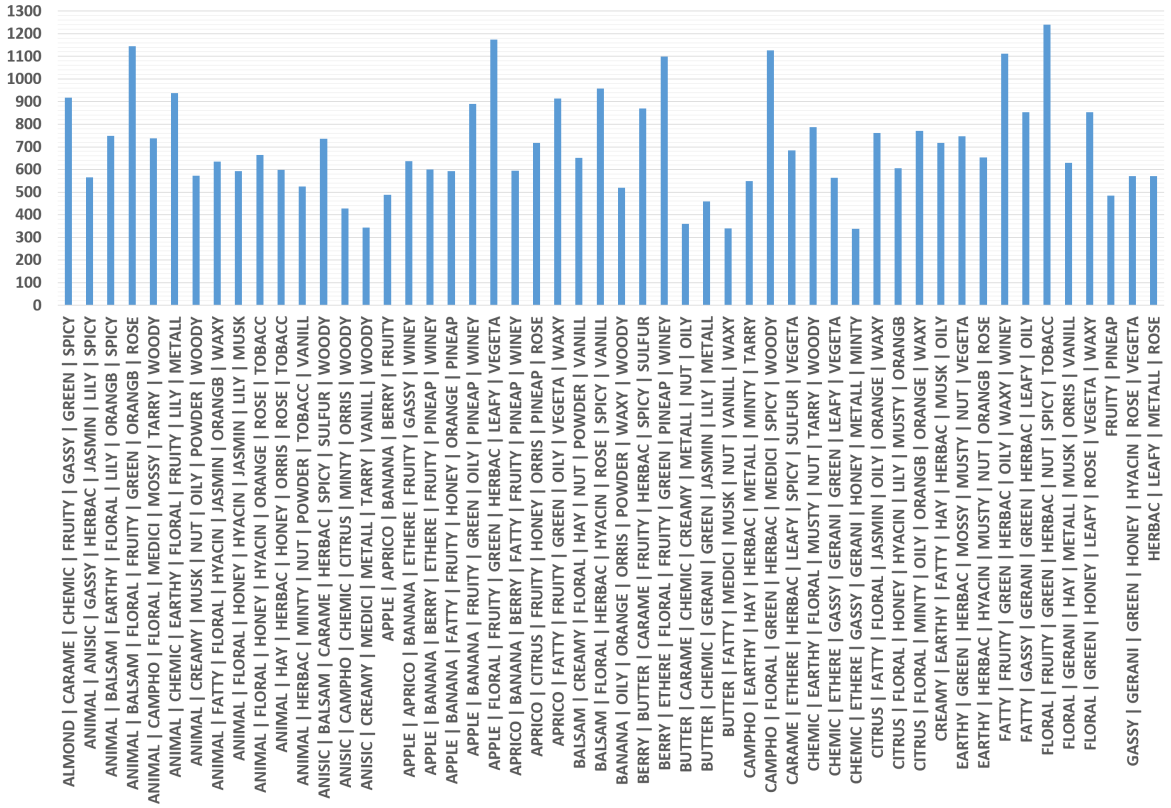


Figure 4.3: Nombre de molécules pour chaque ensemble d'odeurs

La Figure 4.5 illustre l'arbre de décision obtenu à la 10 ème itération de la validation croisée pour la qualité olfactive 'VANILL'. La Figure 4.6 illustre l'arbre de décision de la qualité olfactive 'VANILL' obtenu par l'approche PSI^* à la 10 ème itération de la validation croisée. La signification des attributs descriptifs illustrés dans les deux arbres de décision peut être consultée à partir de la page offcielle du logiciel 'Dragon' [1]. La dégradation de la performance en prédiction pour la qualité olfactive 'VANILL' peut être expliquée par le fait que l'arbre de décision 'VANILL' de l'approche PSI^* est basé sur les prédictions d'un autre arbre de décision pour l'ensemble d'odeurs $\{BUTTER, FATTY, MEDICI, MUSK, NUT, VANILL, WAXY\}$. L'arbre de décision correspondant à cet ensemble d'odeurs est caractérisé par une très faible performance en prédiction selon les résultats illustrés dans la Figure 4.7.

L'avantage de l'approche PSI^* est qu'elle permet de fournir une information supplémentaire sur les

[1] $http://www.talete.mi.it/products/dragon_molecular_descriptor_list.pdf$

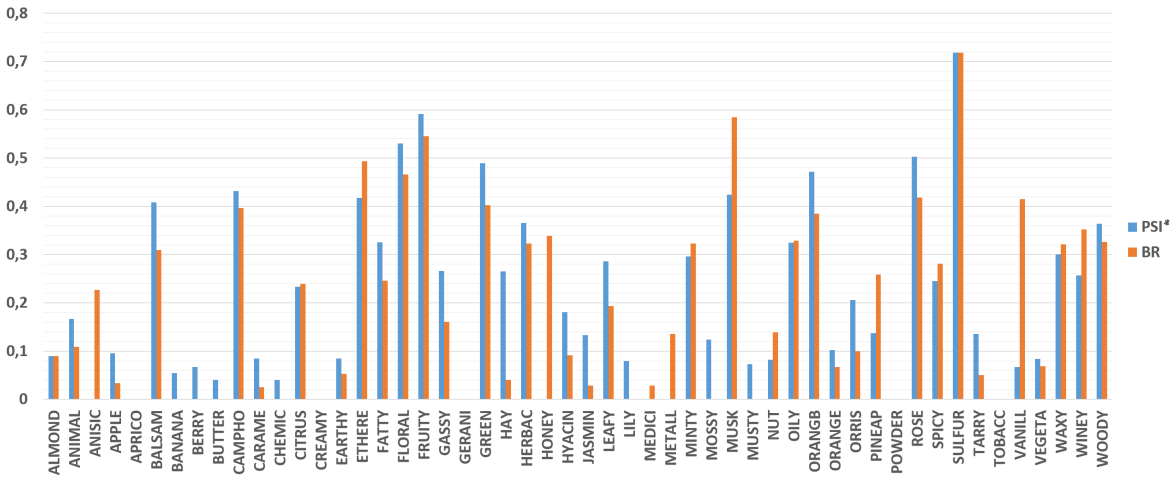


Figure 4.4: Performance en prédiction pour les qualités olfactives

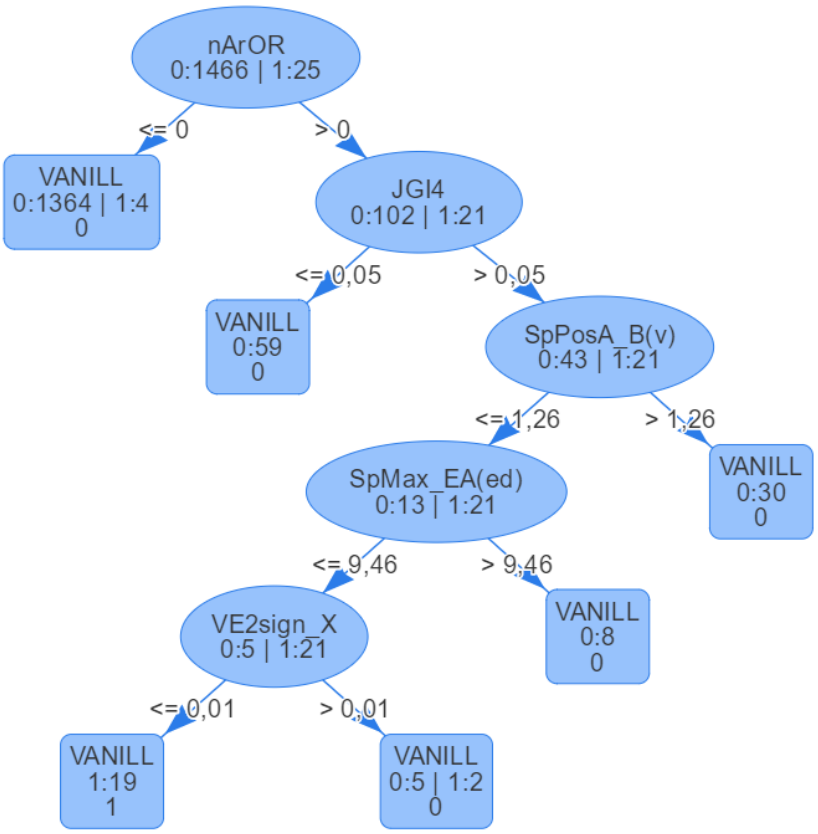


Figure 4.5: L'arbre de décision de la 'VANILL' obtenu par l'approche BR

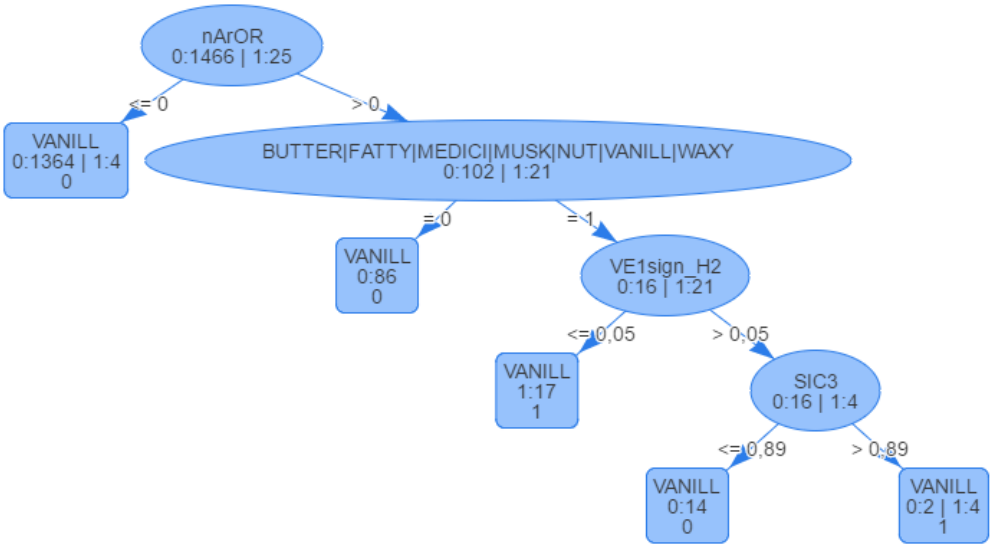


Figure 4.6: L'arbre de décision de la 'VANILL' obtenu par l'approche *PSI**

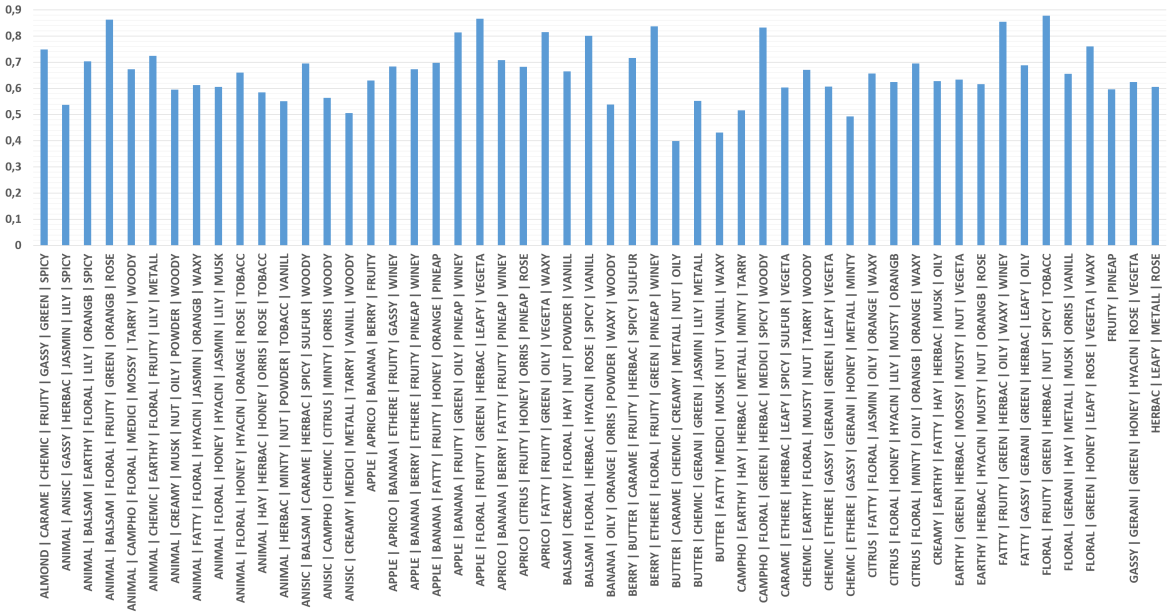


Figure 4.7: Performance en prédiction pour les ensembles de qualités olfactives générées

dépendances entre les qualités olfactives et les sous-ensembles de qualités olfactives. La Figure 4.8 illustre le fait que la qualité olfactive 'ROSE' dépend de la qualité olfactive 'HERBAC'. Ceci traduit le fait qu'au niveau du noeud 'HERBAC' dans l'arbre de décision de la qualité olfactive 'ROSE', la qualité olfactive 'HERBAC' permet de mieux identifier la qualité olfactive 'ROSE' que toutes propriétés physico-chimiques des molécules disponibles. Le graphe de dépendances illustré dans la Figure 4.9 illustre les autres dépendances entre les qualités olfactives extraites à la 10ème étape de la validation croisée.

La performance en prédiction pour les arbres de décision basés sur les propriétés physico-chimiques seulement (approche BR) est très limité pour la plupart des qualités olfactives. L'exploitation des relations entre les qualités olfactives (approche PSI^*) peut permettre d'améliorer les résultats de prédiction. Cependant, la performance en prédiction peut être aussi dégradée en raison de la propagation d'erreur. Le point clé de l'amélioration des résultats dans l'approche PSI^* est de trouver les bons sous-ensembles de labels pour réduire le risque de la propagation d'erreur. Les valeurs de la F-measure pour les sous-ensembles de labels sont comprises entre 0.5 et 0.9. La prédiction d'un sous-ensemble de labels contenant moins de 7 odeurs peut être plus intéressante que la prédiction de la présense ou l'absence d'un seul label avec une performance en prédiction très limitée.

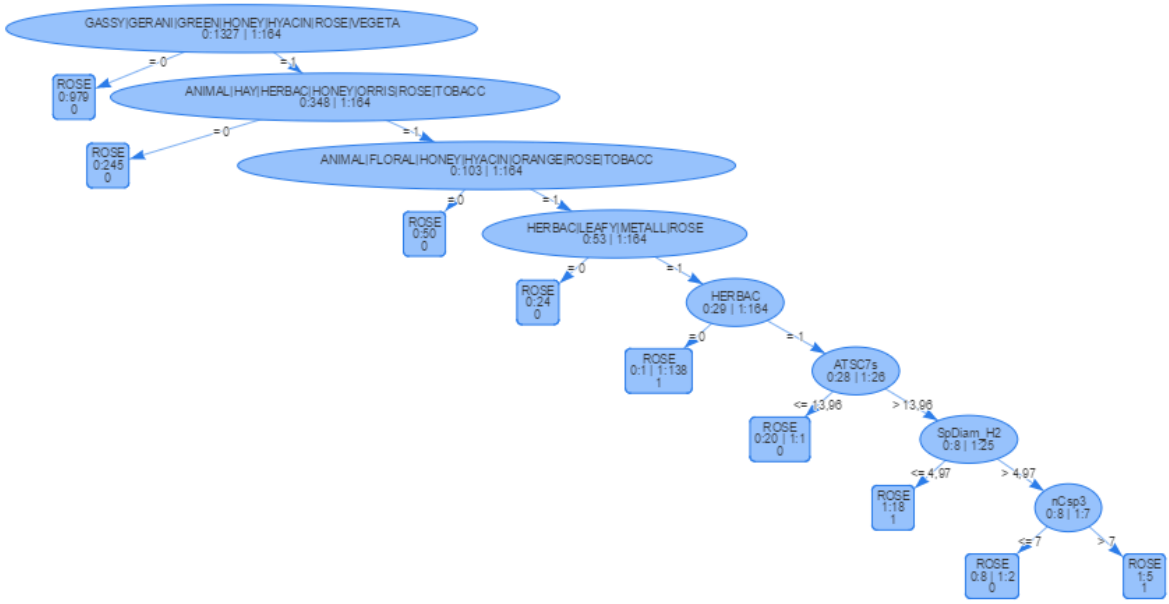


Figure 4.8: L'arbre de décision de la 'ROSE' obtenu par l'approche PSI à la 10ème étape de la validation croisée

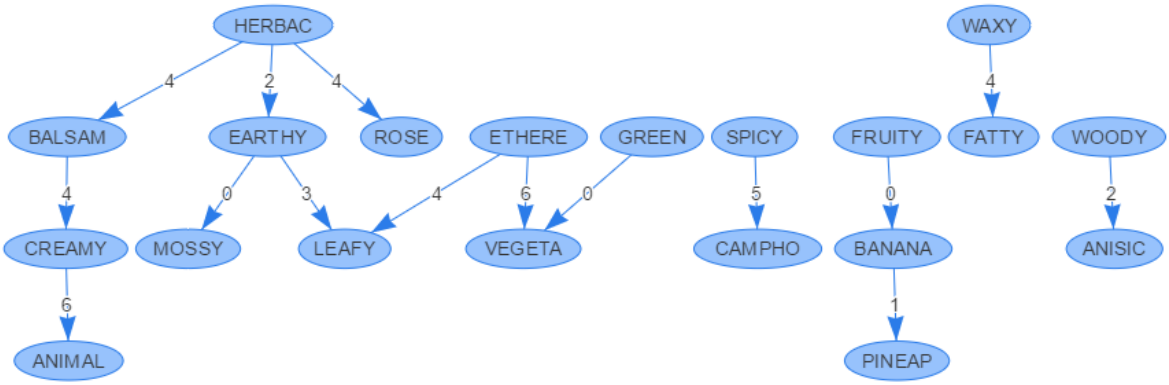


Figure 4.9: Graphe de dépendances obtenu par l'approche PSI à la 10ème étape de la validation croisée

4.5 Conclusion

La tâche de prédire les qualités olfactives des molécules à partir des propriétés physico-chimiques est difficile. Certaines qualités olfactives peuvent être très rares, et les propriétés disponibles ne sont pas suffisantes pour discriminer les qualités olfactives.

L'idée de l'approche que nous proposons est de considérer des sous ensembles de qualités olfactives les plus corrélées entre eux. Ensuite, un classifieur est construit pour chaque sous ensemble de qualités olfactives pour prédire soit la présence d'au moins une qualité olfactive, soit l'absence de toutes les qualités olfactives du sous ensemble considéré. L'avantage de cette approche est que le nombre d'instances associées à au moins une des qualités olfactives du sous ensemble considéré est naturellement plus grand que le nombre des instances associées à exactement une qualité olfactive. Ceci permet de résoudre le problème des qualités olfactives rares.

L'approche que nous proposons construit une deuxième couche de classifieurs pour prédire l'absence ou la présence de chaque qualité olfactive. Chaque classifieur de la deuxième couche fournit des prédictions en se basant à la fois sur les attributs descriptifs et sur les prédictions des classifieurs de la première couche. Ceci permet de résoudre le problème des attributs descriptifs insuffisants à la discrimination des qualités olfactives.

Les résultats d'expérimentation de notre approche montrent que l'exploitation des relations entre les qualités olfactives peut aider à l'amélioration des prédictions. Cependant, la précision en prédiction risque de se dégrader à cause de la propagation des erreurs de prédictions. Le point clé permettant d'obtenir de meilleurs résultats est de trouver les bons sous-ensembles de qualités olfactives à considérer. Le fait d'éliminer les classifieurs de la première couche ne fournissant pas de bonnes prédictions peut aussi améliorer les résultats de notre approche.

Conclusion et perspectives

Dans cette thèse nous nous sommes intéressés au défi d'apprentissage des relations entre les labels à partir des données multi-labels graduées. Nous avons mis en avant les différentes limites des approches existantes dans l'introduction. Notre contribution dans cette thèse est la proposition de nouvelles approches de classification multi-labels graduée permettant de remédier aux limites des approches existantes.

L'approche PSI que nous proposons permet d'apprendre les dépendances entre les labels sans limiter au préalable les dépendances qui peuvent être apprises. Elle permet de fournir des prédictions sans nécessiter une étape de prédiction préliminaire. L'approche PSI garantit des prédictions cohérentes par rapport aux dépendances apprises. Elle permet d'apprendre des dépendances différentes entre les labels en fonction des trois mesures caractérisant l'approche PSI. Les mesures PSI permettent donc de gérer le compromis entre l'apprentissage des relations entre les labels et la réduction du risque de la propagation des erreurs de prédiction. Les résultats de nos expérimentations sur des données multi-labels et multi-labels graduées montrent que l'approche PSI est très compétitive avec les approches de l'état de l'art auxquelles elle a été comparée. L'approche PSI possède deux inconvénients majeurs: elle nécessite de reconstruire certains classifieurs de base dans le processus d'élimination des dépendances cycliques, et elle permet l'apprentissage des dépendances faibles qui augmentent le risque de la propagation des erreurs de prédiction.

L'approche PSI2 que nous proposons permet de remédier aux limites de l'approche PSI. L'approche PSI2 préconise l'utilisation des arbres de décision comme classifieurs de base. Ceci permet de reconstruire uniquement des sous arbres à l'étape de l'élimination des dépendances cycliques. L'approche PSI2 introduit deux paramètres permettant de contrôler les dépendances qui peuvent être apprises par rapport à la profondeur et aux branches de l'arbre de décision. Nous avons évalué les résultats de l'approche PSI2 en utilisant plusieurs configurations de paramètres. Les expérimentations que nous avons mené permettent de ressortir deux configurations particulières offrant des résultats de prédiction similaires. La première configuration consiste à ne pas dépasser deux dépendances dans la même branche de l'arbre de décision, et à autoriser l'apprentissage des dépendances uniquement dans des niveaux de profondeur proches des racines des arbres de décision. Cette configuration permet de

minimiser le risque de la propagation des erreurs de prédiction, et d'éviter l'apprentissage des dépendances faibles entre les labels (dépendances apprises au niveau des nœuds proches des feuilles). La deuxième configuration consiste à autoriser l'apprentissage du maximum de dépendances entre les labels ce qui revient à utiliser l'approche PSI. Ceci s'explique par le fait que les prédictions basées sur une prédiction erronée peuvent être correctes. La prédiction erronée dans ce cas corrige les prédictions qui en dépendent. L'apprentissage des relations entre les labels peut donc augmenter à la fois le nombre des erreurs de prédiction propagées et corrigées. L'avantage principale de l'approche PSI2 se limite donc à l'optimisation du temps d'apprentissage en reconstruisant uniquement les sous arbres de décision impliqués dans une dépendance cyclique. L'inconvénient des approches PSI et PSI2 est le fait qu'elles ne permettent pas d'apprendre les relations de préférence entre les labels.

Les approches CLR_PSI et Stacked_RPC_PSI que nous proposons permettent de combiner l'apprentissage des relations de dépendance et des relations de préférence entre les labels. L'approche CLR_PSI combine notre approche PSI avec l'approche Calibrated Label Ranking (CLR) permettant d'apprendre les relations de préférence entre les labels. La combinaison consiste à construire certains classifieurs de l'approche CLR selon l'approche PSI pour apprendre des dépendances entre les labels. Les prédictions finales sont fournies selon l'approche CLR. L'approche Stacked_RPC_PSI consiste à combiner l'approche Ranking by Pairwise Comparisons (RPC) avec notre approche PSI. La combinaison consiste à apprendre les classifieurs de l'approche RPC, puis les utiliser pour prédire les préférences entre les labels pour les données d'apprentissage. L'approche PSI construit ensuite un ensemble de classifieurs fournissant des prédictions finales en se basant sur les prédictions de préférences entre les labels fournies par l'approche RPC. Les expérimentations que nous avons mené montrent que les approches CLR_PSI et Stacked_RPC_PSI sont très compétitives avec les approches de l'état de l'art par rapport aux résultats de prédiction. Ceci confirme l'hypothèse que nous avons fait à l'introduction sur le fait que la combinaison de plusieurs types de relations entre les labels permet d'améliorer les résultats de prédiction.

Dans cette thèse nous nous somme intéressés aux systèmes de recommandation traitant des données multi-labels graduées arrivant en flux. Nous avons proposé une approche de recommandation d'items basée sur les variables descriptives des utilisateurs, les variables descriptives des items, et les degrés d'appréciation des items auparavant fournis par les utilisateurs. L'approche que nous proposons est basée sur deux modèles de classification multi-labels graduée. Chaque modèle de classification est basé sur un ensemble de classifieurs binaires. Chaque classifieur binaire permet d'apprendre des relations de dépendance entre les labels. Il est construit d'une façon incrémentale pour tenir en compte les données arrivant en flux. Les résultats de nos expérimentations sur des jeux de données différents montrent que notre approche est compétitive avec les approches de l'état de l'art que nous avons comparé. L'inconvénient de notre approche est que le degré d'appréciation minimal et le degré d'appréciation maximal n'ont pas la même probabilité de prédiction par rapport aux autres degrés d'appréciation. Les résultats de prédiction de notre approche peuvent être améliorés davantage en utilisant une approche d'agrégation des prédictions ne pénalisant pas les degrés d'appréciation aux extrémités.

Nous nous somme intéressés dans cette thèse au cas du déséquilibre de la distribution de labels et au manque de variables discriminantes pour la tâche de classification. L'approche que nous avons proposé pour lever ces deux verrous est basée sur l'exploitation des relations entre les labels. Notre approche extrait pour chaque label les données qui lui sont associées. Elle extrait ensuite un sous ensemble de labels dont l'union est associée à ces données. Un classifieur préliminaire est construit pour prédire si une donnée est associée à au moins un label dans le sous ensemble extrait de labels. L'avantage de notre approche est qu'elle permet de gérer le problème des labels rares (associées à un nombre limité de données). En effet, l'ensemble de données associées à l'union de labels que notre approche identifie contient les données associées au label de départ. Le fait que le classifieur préliminaire prédit qu'une donnée est associée à cette union de labels augmente les chances que cette donnée soit aussi associée au label de départ. Le classifieur final permettant de prédire l'absence ou la présence du label de départ peut donc se baser sur la prédiction du classifieur préliminaire en cas de manque de variables discriminantes. Notre approche permet à chaque classifieur final d'exploiter les prédictions de tous les classifieurs préliminaires pour fournir la prédiction finale. Nous avons utilisé dans notre expérimentation un jeu de données de molécules odorantes présentant les deux problèmes du déséquilibre de labels et du manque de variables discriminantes. Les résultats obtenus montrent que notre approche permet d'améliorer la prédiction pour plusieurs labels mais elle est moins performante pour certains labels. L'analyse des liens de dépendances entre les classifieurs préliminaires et finaux nous permet de justifier les résultats obtenus. En effet, les classifieurs permettant de prédire les labels pour lesquels notre approche est moins performante dépendent des classifieurs préliminaires les moins précis en prédiction. La chute de performance est donc expliquée par la propagation de l'erreur de prédiction ayant plus d'impact en absence de variables descriptives discriminantes.

Notre approche adaptée aux contraintes des données de molécules odorantes peut être améliorée en éliminant les classifieurs préliminaires n'ayant pas une bonne performance en prédiction. Chaque classifieur préliminaire éliminé peut être remplacé en changeant l'union de labels qui lui correspond. En effet, il est possible de trouver pour chaque label de départ plusieurs sous ensembles de labels associés aux mêmes données. Une autre voie d'amélioration de notre approche est de modifier la stratégie de sélection des sous ensembles de labels pour les classifieurs préliminaires. En effet notre approche extrait les labels associés au plus de données en commun avec le label de départ. L'ensemble de labels obtenu n'est pas nécessairement plus facile à prédire même si le problème des labels rares est limité. Une stratégie de sélection de sous ensembles de labels qui tient compte de la distribution des données par rapport aux variables descriptives peut améliorer les résultats de notre approche.

Dans cette thèse nous avons proposé des approches de classification multi-labels (PSI, PSI2, CLR_PSI, et Stacked_RPC_PSI) que nous avons combiné avec des approches de décomposition de la classification multi-labels graduée (décomposition verticale, horizontale, et complète). La limite de notre stratégie est que la même approche est appliquée sur l'ensemble des labels. En effet, il se peut que la prédiction pour un sous ensemble de labels soit plus performante en combinant l'approche PSI avec la décomposition verticale, et que la prédiction pour un autre sous ensemble de labels soit plus performante en combinant l'approche Stacked_RPC_PSI avec la décomposition horizontale. La per-

formance globale en prédiction peut donc être améliorée en identifiant la meilleure stratégie a adopter pour chaque sous ensemble de labels.

Nous nous somme concentré dans cette thèse sur les approches de décomposition de la classification multi-labels graduée. Ceci permet d'exploiter les approches existantes pour la classification multi-labels et pour la classification mono-label. Les approches de décomposition ont aussi l'avantage de simplifier la tâche d'apprentissage des relations entre les labels, de leur interprétation et de leur exploitation pour l'amélioration des résultats de prédiction. Une autre voie de recherche consiste à ramener la classification multi-labels graduée au cas plus générale de la classification floue. Cette approche a l'inconvénient que seuls les classifieurs flous peuvent être exploités, et qu'une étape d'agrégation est nécessaire pour convertir les prédictions floues en prédictions multi-labels graduées. L'avantage majeur de cette voie de recherche est que la complexité est largement réduite par rapport aux approches de décomposition nécessitant l'apprentissage de plusieurs classifieurs.

Références

[1] Andrea E. Abele-Brehm and Mahena Stief. Die prognose des berufserfolgs von hochschulabsolventinnen und -absolventen: Befunde zur ersten und zweiten erhebung der erlanger l'angsschnittstudie Bela-E[predicting career success of university graduates: Findings of the first and second wave of the erlangen longitudinal study Bela-E]. *Zeitschrift für Arbeits- und Organisationspsychologie A&O*, 48(1):4–16, 2004.

[2] Shikha Agrawal, Jitendra Agrawal, Shilpy Kaur, and Sanjeev Sharma. A comparative study of fuzzy pso and fuzzy svd-based rbf neural network for multi-label classification. *Neural Computing and Applications*, pages 1–12, 2016. ISSN 1433-3058.

[3] Jose Aguilar, Priscila Valdiviezo-Díaz, and Guido Riofrio. A general framework for intelligent recommender systems. *Applied Computing and Informatics*, 13(2):147 – 160, 2017. ISSN 2210-8327.

[4] M. A. Aizerman, E. A. Braverman, and L. Rozonoer. Theoretical foundations of the potential function method in pattern recognition learning. In *Automation and Remote Control,*, number 25 in Automation and Remote Control,, pages 821–837, 1964.

[5] Everton Alvares-Cherman, Jean Metz, and Maria Carolina Monard. Incorporating label dependency into the binary relevance framework for multi-label classification. *Expert Systems with Applications*, 39(2):1647 – 1655, 2012. ISSN 0957-4174.

[6] Clare Amanda and Ross D. King. Knowledge discovery in multi-label phenotype data. In *Proceedings of the 5th European Conf. on Principles of Data Mining and Knowledge Discovery*, PKDD '01, pages 42–53, London, UK, UK, 2001. Springer-Verlag. ISBN 3-540-42534-9.

[7] A. Q. Ansari, A. Khusro, and M. R. Ansari. Performance evaluation of classifier techniques to discriminate odors with an e-nose. In *2015 Annual IEEE India Conference (INDICON)*, pages 1–5, Dec 2015.

[8] Steffen Arctander. *Perfume and Flavor Chemicals: (aroma Chemicals)*. Perfume and Flavor Chemicals: Aroma Chemicals. Allured Publishing Corporation, 1969. ISBN 9780931710377.

[9] L. R. Bachtiar, C. P. Unsworth, R. D. Newcomb, and E. J. Crampin. Predicting odorant chemical class from odorant descriptor values with an assembly of multi-layer perceptrons. In *2011*

Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pages 2756–2759, Aug 2011.

[10] L. R. Bachtiar, C. P. Unsworth, and R. D. Newcomb. x201c;super e-noses x201d;: Multi-layer perceptron classification of volatile odorants from the firing rates of cross-species olfactory receptor arrays. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 954–957, Aug 2014.

[11] L. R. Bachtiar, R. D. Newcomb, A. V. Kralicek, and C. P. Unsworth. Improving odorant chemical class prediction with multi-layer perceptrons using temporal odorant spike responses from drosophila melanogaster olfactory receptor neurons. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 6393–6396, Aug 2016.

[12] A. Betancourt. A sequential classifier for hand detection in the framework of egocentric vision. In *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 600–605, June 2014.

[13] Arpit Bhardwaj, Aruna Tiwari, Harshit Bhardwaj, and Aditi Bhardwaj. A genetically optimized neural network model for multi-class classification. *Expert Systems with Applications*, 60:211 – 221, 2016. ISSN 0957-4174.

[14] Albert Bifet and Ricard Gavaldà. Learning from time-changing data with adaptive windowing. In *SDM*, pages 443–448. SIAM, 2007. ISBN 978-1-61197-277-1.

[15] J. Bobadilla, F. Ortega, A. Hernando, and A. GutiéRrez. Recommender systems survey. *Know.-Based Syst.*, 46:pp. 109–132, July 2013. ISSN 0950-7051.

[16] Guillaume Bosc, Mehdi Kaytoue, Marc Plantevit, Fabien De Marchi, Moustafa Bensafi, and Jean-François Boulicaut. Vers la découverte de modèles exceptionnels locaux : des règles descriptives liant les molécules à leurs odeurs. In *15e Journées Internationales Francophones Extraction et Gestion des Connaissances (EGC 2015)*, number E.28 in Extraction et Gestion des Connaissances. EGC'2015, Luxembourg, Luxembourg, January 2015. Actes des 15e Journées Internationales Francophones Extraction et Gestion des Connaissances (EGC 2015) qui se sont déroulées à Luxembourg du 27 au 30 janvier 2015.

[17] Guillaume Bosc, Jérôme Golebiowski, Moustafa Bensafi, Céline Robardet, Marc Plantevit, Jean-François Boulicaut, and Mehdi Kaytoue. *Local Subgroup Discovery for Eliciting and Understanding New Structure-Odor Relationships*, pages 19–34. Springer International Publishing, Cham, 2016. ISBN 978-3-319-46307-0.

[18] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational*

Learning Theory, COLT '92, pages 144–152, New York, NY, USA, 1992. ACM. ISBN 0-89791-497-X.

[19] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984.

[20] C. Brinker, E. L. Mencía, and J. Fürnkranz. Graded multilabel classification by pairwise comparisons. In *2014 IEEE International Conference on Data Mining*, pages 731–736, Dec 2014.

[21] Christian Brinker, Eneldo Loza Mencía, and Johannes Fürnkranz. Graded multilabel classification by pairwise comparisons. In Ravi Kumar, Hannu Toivonen, Jian Pei, Joshua Zhexue Huang, and Xindong Wu, editors, *ICDM*, pages 731–736. IEEE Computer Society, 2014. ISBN 978-1-4799-4302-9.

[22] Linda Buck and Richard Axel. A novel multigene family may encode odorant receptors: A molecular basis for odor recognition. *Cell*, 65(1):175 – 187, 1991. ISSN 0092-8674.

[23] Allison Carey, Guirong Wang, Chih-Ying Su, Laurence J. Zwiebel, and John R. Carlson. Odourant reception in the malaria mosquito anopheles gambiae. *Nature*, 464(7285):66–71, Mar 2010. ISSN 0028-0836. 20130575[pmid].

[24] Quansheng Chen, Aiping Liu, Jiewen Zhao, and Qin Ouyang. Classification of tea category using a portable electronic nose based on an odor imaging sensor array. *Journal of Pharmaceutical and Biomedical Analysis*, 84:77 – 83, 2013. ISSN 0731-7085.

[25] Weiwei Cheng, Krzysztof Dembczynski, and Eyke Hüllermeier. Graded multilabel classification: The ordinal case. In Martin Atzmüller, Dominik Benz, Andreas Hotho, and Gerd Stumme, editors, *Proceedings of LWA2010 - Workshop-Woche: Lernen, Wissen & Adaptivitaet*, Kassel, Germany, 2010.

[26] Edith Cohen and Martin J. Strauss. Maintaining time-decaying stream aggregates. *J. Algorithms*, 59(1):pp. 19–36, April 2006. ISSN 0196-6774.

[27] S. M. Daud, M. S. Najib, and N. Zahed. Classification of lubricant oil odor-profile using case-based reasoning. In *2016 IEEE Conference on Systems, Process and Control (ICSPC)*, pages 207–212, Dec 2016.

[28] Claire A. de March, SangEun Ryu, Gilles Sicard, Cheil Moon, and Jérôme Golebiowski. Structure–odour relationships reviewed in the postgenomic era. *Flavour and Fragrance Journal*, 30 (5):342–361, 2015. ISSN 1099-1026. FFJ-14-0196.R2.

[29] Sebastien Destercke. *Multilabel Prediction with Probability Sets: The Hamming Loss Case*, pages 496 – 505. Springer International Publishing, Cham, 2014. ISBN 978-3-319-08855-6.

[30] Shifei Ding, Xiekai Zhang, Yuexuan An, and Yu Xue. Weighted linear loss multiple birth support vector machine based on information granulation for multi-class classification. *Pattern Recognition*, 67:32 – 46, 2017. ISSN 0031-3203.

[31] Pedro Domingos and Geoff Hulten. Mining high-speed data streams. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '00, pages 71–80, New York, NY, USA, 2000. ACM. ISBN 1-58113-233-6.

[32] Michael D. Ekstrand, John T. Riedl, and Joseph A. Konstan. Collaborative filtering recommender systems. *Found. Trends Hum.-Comput. Interact.*, 4(2):81–173, February 2011. ISSN 1551-3955.

[33] André Elisseeff and Jason Weston. A kernel method for multi-labelled classification. In *In Advances in Neural Information Processing Systems 14*, pages 681–687. MIT Press, 2001.

[34] H. R. Estakhroueiyeh and E. Rashedi. Detecting moldy bread using an e-nose and the knn classifier. In *2015 5th International Conference on Computer and Knowledge Engineering (ICCKE)*, pages 251–255, Oct 2015.

[35] Eibe Frank and Mark Hall. A simple approach to ordinal classification. In *Proceedings of the 12th European Conference on Machine Learning*, EMCL '01, pages 145–156, London, UK, UK, 2001. Springer-Verlag. ISBN 3-540-42536-5.

[36] J. Friedman. Another approach to polychotomous classification. *Dept. Statistics, Stanford Univ., Tech. Rep*, 1996.

[37] Johannes Fürnkranz, Eyke Hüllermeier, Eneldo Loza Mencía, and Klaus Brinker. Multilabel classification via calibrated label ranking. *Machine Learning*, 73(2):133–153, 2008. ISSN 1573-0565.

[38] Imen Gaied, Farah Jemili, and Ouajdi Korbaa. *A Genetic-Fuzzy Classification Approach to Improve High-Dimensional Intrusion Detection System*, pages 319–329. Springer International Publishing, Cham, 2017. ISBN 978-3-319-53480-0.

[39] JoÃo Gama, Raquel Sebastião, and Pedro Pereira Rodrigues. Issues in evaluation of stream learning algorithms. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 329–338, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-495-9.

[40] João Gama, Raquel Sebastião, and Pedro Pereira Rodrigues. On evaluating stream learning algorithms. *Mach. Learn.*, 90(3):pp. 317–346, March 2013. ISSN 0885-6125.

[41] A Gepperth and B Hammer. Incremental learning algorithms and applications. pages 357–368, 2016.

[42] Shantanu Godbole and Sunita Sarawagi. *Advances in Knowledge Discovery and Data Mining: 8th Pacific-Asia Conference, PAKDD 2004, Sydney, Australia, May 26-28, 2004. Proceedings*, chapter Discriminative Methods for Multi-labeled Classification, pages 22–30. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004. ISBN 978-3-540-24775-3.

[43] Ken Goldberg, Theresa Roeder, Dhruv Gupta, and Chris Perkins. Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval*, 4(2):pp. 133–151, 2001. ISSN 1573-7659.

[44] Elissa A. Hallem and John R. Carlson. Coding of odors by a receptor repertoire. *Cell*, 125(1): 143 – 160, 2006. ISSN 0092-8674.

[45] A. Haque, L. Khan, M. Baron, B. Thuraisingham, and C. Aggarwal. Efficient handling of concept drift and concept evolution over stream data. In *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*, pages 481–492, May 2016.

[46] F. Maxwell Harper and Joseph A. Konstan. The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 5(4):pp. 1–19, December 2015. ISSN 2160-6455.

[47] Trevor Hastie and Robert Tibshirani. Classification by pairwise coupling. In *Proceedings of the 1997 Conference on Advances in Neural Information Processing Systems 10*, NIPS '97, pages 507–513, Cambridge, MA, USA, 1998. MIT Press. ISBN 0-262-10076-2.

[48] Francisco Herrera, Francisco Charte, Antonio J. Rivera, and María J. del Jesus. *Multilabel Classification Problem Analysis, Metrics and Techniques*, chapter Multilabel Classification, pages 17–31. 2016. ISBN 978-3-319-41111-8.

[49] W Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, pages 13–30, 1963.

[50] Eyke Hüllermeier, Johannes Fürnkranz, Weiwei Cheng, and Klaus Brinker. Label ranking by learning pairwise preferences. *Artificial Intelligence*, 172(16–17):1897 – 1916, 2008. ISSN 0004-3702.

[51] Geoff Hulten, Laurie Spencer, and Pedro Domingos. Mining time-changing data streams. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '01, pages 97–106, New York, NY, USA, 2001. ACM. ISBN 1-58113-391-X.

[52] Sunil Kr. Jha and Kenshi Hayashi. Body odor classification by selecting optimal peaks of chemical compounds in gc–ms spectra using filtering approaches. *International Journal of Mass Spectrometry*, 415:92 – 102, 2017. ISSN 1387-3806.

[53] Sunil Kr. Jha, Filip Josheski, Ninoslav Marina, and Kenshi Hayashi. Gc–ms characterization of body odour for identification using artificial neural network classifiers fusion. *International Journal of Mass Spectrometry*, 406:35 – 47, 2016. ISSN 1387-3806.

[54] R. Jindal and S. Taneja. Ranking in multi label classification of text documents using quantifiers. In *2015 IEEE International Conference on Control System, Computing and Engineering (ICCSCE)*, pages 162–166, Nov 2015.

[55] Mohamed Kissi, Mohammed Ramdani, Mustapha Tollabi, and Driss Zakarya. Determination of fuzzy logic membership functions using genetic algorithms: application to structure–odor modeling. *Journal of molecular modeling*, 10(5-6):335–341, 2004.

[56] Mohamed Kissi, Mohammed Ramdani, Bernadette Bouchon-Meunier, and Driss Zakarya. Pattern recognition system based on empirical knowledge: Sandalwood and camphoraceous odours application. *Mathematics and computers in simulation*, 77(5):453–463, 2008.

[57] E. Kroupi, J. M. Vesin, and T. Ebrahimi. Subject-independent odor pleasantness classification using brain and peripheral signals. *IEEE Transactions on Affective Computing*, 7(4):422–434, Oct 2016. ISSN 1949-3045.

[58] Miroslav Kubat, Robert Holte, and Stan Matwin. *Learning when negative examples abound*, pages 146–153. Springer Berlin Heidelberg, Berlin, Heidelberg, 1997. ISBN 978-3-540-68708-5.

[59] Roshan Kumari and Saurabh Kr. Srivastava. Machine learning: A review on binary classification. *International Journal of Computer Applications*, 160(7):11–15, Feb 2017. ISSN 0975-8887.

[60] K. Laghmari, C. Marsala, and M. Ramdani. Graded multi-label classification: Compromise between handling label relations and limiting error propagation. In *11th Inter. Conf. on Intelligent Systems: Theories and Applications (SITA)*, pages 1–6, Oct 2016.

[61] Khalil Laghmari, Christophe Marsala, and Mohammed Ramdani. *A Distributed Recommender System Based on Graded Multi-label Classification*, pages 101–108. Springer International Publishing, Cham, 2017. ISBN 978-3-319-59647-1.

[62] Khalil Laghmari, Christophe Marsala, and Mohammed Ramdani. An adapted incremental graded multi-label classification model for recommendation systems. *Progress in Artificial Intelligence*, Aug 2017. ISSN 2192-6360. doi: 10.1007/s13748-017-0133-5.

[63] Khalil Laghmari, Christophe Marsala, and Mohammed Ramdani. Classification multi-labels graduée: apprendre les relations entre les labels ou limiter la propagation d'erreur ? Extraction et Gestion des Connaissances, RNTI-E-33:381–386, 2017.

[64] Khalil Laghmari, Christophe Marsala, and Mohammed Ramdani. *Learning Label Dependency and Label Preference Relations in Graded Multi-Label Classification*, chapter Multilabel Classification. Springer-Verlag, 2018.

[65] Khalil Laghmari, Christophe Marsala, and Mohammed Ramdani. Dynamic label relation learning for multi-label classification. *Data Science and Analytics*, 2018.

[66] Khalil Laghmari, Christophe Marsala, and Mohammed Ramdani. Apprendre les relations de préférences et de co-occurrence entre les labels en classification multi-labels. Extraction et Gestion des Connaissances, RNTI, 2018.

[67] A. Lanata, A. Guidi, A. Greco, G. Valenza, F. Di Francesco, and E. P. Scilingo. Automatic recognition of pleasant content of odours through electroencephalographic activity analysis. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 4519–4522, Aug 2016.

[68] Gerardo Lastra, Oscar Luaces, and Antonio Bahamonde. Interval prediction for graded multilabel classification. *Pattern Recognition Letters*, 49:171 – 176, 2014. ISSN 0167-8655.

[69] Chunming Liu and Longbing Cao. *A Coupled k-Nearest Neighbor Algorithm for Multi-label Classification*, pages 176–187. Springer International Publishing, Cham, 2015. ISBN 978-3-319-18038-0.

[70] Pierre-Xavier Loeffel, Christophe Marsala, and Marcin Detyniecki. Memory management for data streams subject to concept drift. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pages 387–392, 2016.

[71] Pasquale Lops, Marco de Gemmis, and Giovanni Semeraro. *Content-based Recommender Systems: State of the Art and Trends*, pages 73–105. Springer US, Boston, MA, 2011. ISBN 978-0-387-85820-3.

[72] Eneldo Loza Mencía and Johannes Fürnkranz. Efficient pairwise multilabel classification for large-scale problems in the legal domain. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases - Part II*, ECML PKDD '08, pages 50–65, Berlin, Heidelberg, 2008. Springer-Verlag. ISBN 978-3-540-87480-5.

[73] Oscar Luaces, Jorge Díez, José Barranquero, Juan José del Coz, and Antonio Bahamonde. Binary relevance efficacy for multilabel classification. *Progress in Artificial Intelligence*, 1(4): 303–313, Dec 2012. ISSN 2192-6360.

[74] Ingo T. Mahlke, Peter H. Thiesen, and Bernd Niemeyer. Chemical indices and methods of multivariate statistics as a tool for odor classification. *Environmental Science & Technology*, 41(7):2414–2421, 2007. PMID: 17438794.

[75] Christophe Marsala, Mohammed Ramdani, Mustapha Tollabi, and Driss Zakarya. Recognition of Odors: a Fuzzy Decision Tree Approach. In *Seventh International Conference IPMU*, pages 532–539, 1998.

[76] Andrea Mauri, Viviana Consonni, Manuela Pavan, and Roberto Todeschini. Dragon software: An easy approach to molecular descriptor calculations. *MATCH / Communications In Mathematical & In Computer Chemistry*, 56:237–248, 2006.

[77] Neha Mehra and Surendra Gupta. Survey on multiclass classification methods. *International Journal of Computer Science and Information Technologies*, 4(4):572–576, 2013. ISSN 0975-9646.

[78] Deiner Mena, Elena Montañés, José Ramón Quevedo, and Juan José del Coz. An overview of inference methods in probabilistic classifier chains for multilabel classification. *Wiley Interdisc. Rew.: Data Mining and Knowledge Discovery*, 6(6):215–230, 2016.

[79] Elena Montañés, José Ramón Quevedo, and Juan José del Coz. Aggregating independent and dependent models to learn multi-label classifiers. In Dimitrios Gunopulos, Thomas Hofmann, Donato Malerba, and Michalis Vazirgiannis, editors, *ECML/PKDD (2)*, volume 6912 of *Lecture Notes in Computer Science*, pages 484–500. Springer, 2011. ISBN 978-3-642-23782-9.

[80] Elena Montañés, Robin Senge, Jose Barranquero, José Ramón Quevedo, Juan José del Coz, and Eyke Hüllermeier. Dependent binary relevance models for multi-label classification. *Pattern Recogn.*, 47(3):1494–1508, March 2014. ISSN 0031-3203.

[81] Gonzalo Nápoles, Rafael Falcon, Elpiniki Papageorgiou, Rafael Bello, and Koen Vanhoof. Partitive granular cognitive maps to graded multilabel classification. In *2016 IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2016, Vancouver, BC, Canada, July 24-29, 2016*, pages 1363–1370. IEEE, 2016. ISBN 978-1-5090-0626-7.

[82] S. Omatu. Odor classification by neural networks. In *2013 IEEE 7th International Conference on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS)*, volume 01, pages 309–314, Sept 2013.

[83] S. Omatu, D. Hayashi, and M. Yano. Odor classification of wines by using neural networks. In *17th International Conference on Information Fusion (FUSION)*, pages 1–6, July 2014.

[84] P. Patil and V. Kulkarni. Odour detection and classification. In *2015 International Conference on Green Computing and Internet of Things (ICGCIoT)*, pages 426–428, Oct 2015.

[85] Michael J. Pazzani. A framework for collaborative, content-based and demographic filtering. *Artificial Intelligence Review*, 13(5):pp. 393–408, 1999. ISSN 1573-7462.

[86] Ignazio Pillai, Giorgio Fumera, and Fabio Roli. Designing multi-label classifiers that maximize f measures: State of the art. *Pattern Recognition*, 61:394 – 404, 2017. ISSN 0031-3203.

[87] R. C. Prati. Fuzzy rule classifiers for multi-label classification. In *2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–8, Aug 2015.

[88] J. R. Quinlan. Induction of decision trees. *MACH. LEARN*, 1:81–106, 1986.

[89] J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993. ISBN 1558602402.

[90] M. M. Rahman, C. Charoenlarpnopparut, and P. Suksompong. Signal processing for multi-sensor e-nose system: Acquisition and classification. In *2015 10th International Conference on Information, Communications and Signal Processing (ICICS)*, pages 1–5, Dec 2015.

[91] Mohammed Ramdani, Mohamed Kissi, and Bernadette Bouchon-Meunier. Man-machine interaction to extract features of odorous molecules. In *Intelligent Sensory Evaluation*, pages 255–268. Springer Berlin Heidelberg, 2004.

[92] J. Read. A Pruned Problem Transformation Method for Multi-label classification. In *Proc. 2008 New Zealand Computer Science Research Student Conference (NZCSRS 2008)*, pages 143–150, 2008.

[93] J. Read, A. Puurula, and A. Bifet. Multi-label classification with meta-labels. In *2014 IEEE International Conference on Data Mining*, pages 941–946, Dec 2014.

[94] Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. Classifier chains for multi-label classification. *Mach. Learn.*, 85(3):333–359, December 2011. ISSN 0885-6125.

[95] Jesse Read, Luca Martino, Pablo M. Olmos, and David Luengo. Scalable multi-output label prediction: From classifier chains to classifier trellises. *Pattern Recognition*, 48(6):2096 – 2109, 2015. ISSN 0031-3203.

[96] C. Salperwyck and V. Lemaire. Incremental decision tree based on order statistics. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, Aug 2013.

[97] J. C. Schlimmer and D. Fisher. A case study of incremental concept induction. In *Proceedings of the Fifth National Conference on Artificial Intelligence*, pages 496–501, 1986.

[98] T. Seesaard, C. Sriphrapradang, T. Kitiyakara, and T. Kerdcharoen. Self-screening for diabetes by sniffing urine samples based on a hand-held electronic nose. In *2016 9th Biomedical Engineering International Conference (BMEiCON)*, pages 1–4, Dec 2016.

[99] Robin Senge, Juan José del Coz, and Eyke Hüllermeier. *On the Problem of Error Propagation in Classifier Chains for Multi-label Classification*, pages 163–170. Springer International Publishing, Cham, 2014. ISBN 978-3-319-01595-8.

[100] S. Sharma and R. Kumar. Fuzzy relative entropy based classification scheme for discrimination of odors/gases using a poorly selective sensor array. In *2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1195–1200, July 2016.

[101] B. Soonsiripanichkul and T. Murata. Domination dependency analysis of sales marketing based on multi-label classification using label ordering and cycle chain classification. In *2016 5th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, pages 1048–1053, July 2016.

[102] Zhongwei Sun, Zhongwen Guo, Mingxing Jiang, Xi Wang, and Chao Liu. *Research and Application of Fast Multi-label SVM Classification Algorithm Using Approximate Extreme Points*, pages 39–52. Springer International Publishing, Cham, 2016. ISBN 978-3-319-42553-5.

[103] Farbound Tai and Hsuan-Tien Lin. Multilabel classification with principal label space transformation. *Neural Comput.*, 24(9):2508–2542, September 2012. ISSN 0899-7667.

[104] Aik Choon Tan, Daniel Q. Naiman, Lei Xu, Raimond L. Winslow, and Donald Geman. Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics*, 21 (20):3896–3904, 10 2005. ISSN 1367-4803.

[105] X. Tang and A. Xu. Multi-class classification using kernel density estimation on k-nearest neighbours. *Electronics Letters*, 52(8):600–602, 2016. ISSN 0013-5194.

[106] Fengchun Tian, Zhifang Liang, Lei Zhang, Yan Liu, and Zhenzhen Zhao. A novel pattern mismatch based interference elimination technique in e-nose. *Sensors and Actuators B: Chemical*, 234:703 – 712, 2016. ISSN 0925-4005.

[107] Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *Int J Data Warehousing and Mining*, 2007:1–13, 2007.

[108] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. Mining multi-label data. In *Data Mining and Knowledge Discovery Handbook*, pages 667–685, 2010.

[109] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. *Data Mining and Knowledge Discovery Handbook*, chapter Mining Multi-label Data, pages 667–685. Springer US, Boston, MA, 2010. ISBN 978-0-387-09823-4.

[110] Ayşegül Uçar and Recep Özalp. Efficient android electronic nose design for recognition and perception of fruit odors using kernel extreme learning machines. *Chemometrics and Intelligent Laboratory Systems*, 166:69 – 80, 2017. ISSN 0169-7439.

[111] Ondrej Šuch and Santiago Barreda. Bayes covariant multi-class classification. *Pattern Recognition Letters*, 84:99 – 106, 2016. ISSN 0167-8655.

[112] Paul E. Utgoff. Id5: An incremental id3. In John E. Laird, editor, *ML*, pages 107–120. Morgan Kaufmann, 1988. ISBN 0-934613-64-8.

[113] Paul E. Utgoff. Incremental induction of decision trees. *Mach. Learn.*, 4(2):pp. 161–186, November 1989. ISSN 0885-6125.

[114] Paul E. Utgoff. An improved algorithm for incremental induction of decision trees. In *In Proceedings of the Eleventh International Conference on Machine Learning*, pages 318–325. Morgan Kaufmann, 1994.

[115] V. Vapnik and A. Lerner. Pattern Recognition using Generalized Portrait Method. *Automation and Remote Control*, 24, 1963.

[116] Shangfei Wang, Jun Wang, Zhaoyu Wang, and Qiang Ji. Enhancing multi-label classification by modeling dependencies among labels. *Pattern Recognition*, 47(10):3405 – 3413, 2014. ISSN 0031-3203.

[117] Xiaoxue Wang, Shuang An, Hong Shi, and Qinghua Hu. *Fuzzy Rough Decision Trees for Multi-label Classification*, pages 207–217. Springer International Publishing, Cham, 2015. ISBN 978-3-319-25783-9.

[118] Y. Wang, S. C. F. Chan, and G. Ngai. Applicability of demographic recommender system to tourist attractions: A case study on trip advisor. In *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, volume 3, pages 97–101, Dec 2012.

[119] W. Xie, Y. Ouyang, J. Ouyang, W. Rong, and Z. Xiong. User occupation aware conditional restricted boltzmann machine based recommendation. In *2016 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, pages 454–461, Dec 2016.

[120] Eleftherios Spyromitros Xioufis, Myra Spiliopoulou, Grigorios Tsoumakas, and Ioannis Vlahavas. Dealing with concept drift and class imbalance in multi-label stream classification. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Two*, IJCAI'11, pages 1583–1588. AAAI Press, 2011. ISBN 978-1-57735-514-4.

[121] Xuesong Yan, Wei Li, Qinghua Wu, and Victor S. Sheng. *A Double Weighted Naive Bayes for Multi-label Classification*, pages 382–389. Springer Singapore, Singapore, 2016. ISBN 978-981-10-0356-1.

[122] Song Yan-yan and Lu Ying. Decision tree methods: applications for classification and prediction. *Shanghai Arch Psychiatry*, 27(20):130–135, 4 2015. ISSN 1002-0829.

[123] Xiwang Yang, Yang Guo, Yong Liu, and Harald Steck. A survey of collaborative filtering based social recommender systems. *Comput. Commun.*, 41:1–10, March 2014. ISSN 0140-3664.

[124] Zoulficar Younes, Fahed Abdallah, Thierry Denoeux, and Hichem Snoussi. A dependent multilabel classification method derived from the k-nearest neighbor rule. *EURASIP J. Adv. Sig. Proc.*, 2011, 2011.

[125] L.A. Zadeh. Fuzzy sets. *Information and Control*, 8(3):338 – 353, 1965. ISSN 0019-9958.

[126] Z. Zaier, R. Godin, and L. Faucher. Recommendation quality evolution based on neighbors discrimination. In *2008 International MCETECH Conference on e-Technologies (mcetech 2008)*, pages 148–153, Jan 2008.

[127] L. Zhang and P. Deng. Abnormal odor detection in electronic nose via self-expression inspired extreme learning machine. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, PP(99):1–11, 2017. ISSN 2168-2216.

[128] L. Zhang and D. Zhang. Efficient solutions for discreteness, drift, and disturbance (3d) in electronic olfaction. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, PP(99): 1–13, 2017. ISSN 2168-2216.

[129] Lei Zhang, Fengchun Tian, Lijun Dang, Guorui Li, Xiongwei Peng, Xin Yin, and Shouqiong Liu. A novel background interferences elimination method in electronic nose using pattern recognition. *Sensors and Actuators A: Physical*, 201:254 – 263, 2013. ISSN 0924-4247.

[130] M. L. Zhang and Z. H. Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837, Aug 2014. ISSN 1041-4347.

[131] Min-Ling Zhang and Zhi-Hua Zhou. A k-nearest neighbor based algorithm for multi-label classification. In *Granular Computing, 2005 IEEE International Conference on*, volume 2, pages 718–721 Vol. 2, July 2005.

[132] Min-Ling Zhang and Zhi-Hua Zhou. Ml-knn: A lazy learning approach to multi-label learning. *Pattern Recogn.*, 40(7):2038–2048, July 2007. ISSN 0031-3203.