Apprendre les relations de préférence et de co-occurrence entre les labels en classification multi-labels

Khalil Laghmari[*,**]
Christophe Marsala[**]
Mohammed Ramdani[*]

[*]Laboratoire Informatique de Mohammedia,
FSTM, Hassan II University of Casablanca,
BP 146 Mohammedia 20650 Maroc.
laghmari.khalil@gmail.com
ramdani@fstm.ac.ma
[**]Sorbonne Universités,
UPMC Univ Paris 06,
CNRS, LIP6 UMR 7606,
4 place Jussieu 75005 Paris, France.
christophe.marsala@lip6.fr

Résumé. En classification multi-labels, chaque instance est associée à un ou plusieurs labels. Par exemple, un morceau de musique peut être associé aux labels 'heureux' et 'relaxant'. Des relations de co-occurrence peuvent exister entre les labels : par exemple, les labels 'heureux' et 'triste' ne peuvent pas être associés au même morceau de musique. Les labels peuvent aussi avoir des relations de préférence : par exemple, pour un morceau de musique contenant plusieurs piques, le label 'heureux' est préféré par rapport au label 'relaxant'. Les relations entre les labels peuvent aider à mieux prédire les labels associés aux instances. Les approches existantes peuvent apprendre soit les relations de co-occurrence, soit les relations de préférence. Ce travail introduit une approche permettant de combiner l'apprentissage des deux types de relations. Les expérimentations menées montrent que la nouvelle approche introduite offre les meilleurs résultats de prédiction par rapports à cinq approches de l'état de l'art.

1 Introduction

La reconnaissance de formes (pattern recognition) constitue une compétence d'intelligence fondamentale. Par exemple, en considérant un ensemble de labels disponibles : {'heureux', 'relaxant', 'triste', 'nerveux'}, l'intelligence humaine est capable d'associer ces labels à des morceaux de musique selon les émotions qu'ils expriment.

Le défi de transférer la compétence de reconnaissance de formes aux ordinateurs fait partie du domaine de l'apprentissage artificiel. L'apprentissage est dit supervisé

lorsqu'un ensemble d'instances dont les labels associés sont connus est disponible. L'apprentissage consiste à établir des liens entres les attributs descriptifs des instances et les labels associés. L'objectif de l'apprentissage est de pouvoir prédire les labels associés aux instances en se basant sur les attributs descriptifs. L'ensemble d'instances est dit multi-labels lorsque chaque instance est associée à un ou plusieurs labels parmi un ensemble de labels disponibles.

La classification multi-labels (Herrera et al. (2016)) est une tâche d'apprentissage supervisé sur des instances multi-labels. La classification floue (Bouchon-Meunier et al. (1997)) où l'association entre chaque instance et les labels disponibles possède un degré dans l'intervalle $[0, 1]$ est une généralisation de la classification multi-labels où le degré d'association est binaire (0 ou 1). Dans cet article on s'intéresse à la classification multi-labels, et plus particulièrement au défi d'apprendre des relations entre les labels et les exploiter pour améliorer les prédictions (Loza Mencía et Janssen (2014) ;Loza Mencía et Janssen (2016)). Il existe deux types principaux de relations entre les labels :

— les relations de co-occurrence : par exemple, les émotions 'heureux' et 'triste' sont rarement associées au même morceau de musique, et les émotions 'heureux' et 'relaxant' peuvent être associées au même morceau de musique.
— les relations de préférence : par exemple, pour un morceau de musique contenant plusieurs piques, le label à préférer entre 'relaxant' et 'heureux' pour l'associer à ce morceau de musique est le label 'heureux'.

Les approches de classification multi-labels existantes peuvent apprendre soit uniquement des relations de co-occurrence, soit uniquement des relations de préférence (Gibaja et Ventura (2015)). Dans cet article, nous faisons l'hypothèse qu'une approche qui permet d'apprendre les deux types de relations entre les labels peut avoir une meilleure performance en prédiction que les approches existantes.

La suite de cet article est organisée comme suit : l'état de l'art des approches de classification multi-labels est discuté dans la Section 2 ; la nouvelle approche de classification multi-labels permettant l'apprentissage des relations de co-occurrence et de préférence entre les labels est présentée dans la Section 3 ; l'étude expérimentale comparant la nouvelle approche avec les approches existantes est présentée dans la Section 4.

2 Classification multi-labels

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 $(x_{i,a_1}, \ldots, x_{i,a_p}) = (x_{i,a_j})_{1 \leq j \leq p}$. Soit $C = \{c_l\}_{1 \leq l \leq k}$ un ensemble de labels. Chaque instance x_i est associée à un sous-ensemble de labels $y_i \subseteq C$.
L'ensemble de sous-ensembles de labels est noté $\mathcal{P}(C)$. Soit $\lambda : X \rightarrow \mathcal{P}(C)$ la fonction qui associe chaque instance $x_i \in X$ au sous-ensemble de labels correspondant $\lambda(x_i) = y_i$. La fonction λ est dite *fonction de supervision* de l'ensemble d'apprentissage X.
Soit $E : \mathcal{P}(C) \times \mathcal{P}(C) \rightarrow [0, 1]$ une fonction objectif à optimiser. La classification multi-labels consiste à apprendre à partir de l'ensemble d'apprentissage supervisé

(X, λ) un classifieur $H : a_1 \times \ldots \times a_p \rightarrow \mathcal{P}(C)$ qui prédit pour chaque instance $x \in a_1 \times \ldots \times a_p$ l'ensemble de labels correspondants $H(x)$. L'objectif du classifieur multi-labels est d'optimiser la fonction objectif E évaluant la prédiction $H(x)$ par rapport au véritable sous-ensemble de labels $y \subseteq C$ associé à x.

2.2 Approches de classification multi-labels

La classification multi-labels est une généralisation de la classification mono-label où chaque instance $x_i \in X$ ne peut être associée qu'à un seul label à la fois : $|y_i| = 1$. La classification multi-labels peut être effectuée en adaptant les classifieurs mono-label au cas des instances multi-labels (Sun et al. (2016) ;Agrawal et al. (2016) ;Wang et al. (2015)). L'inconvénient de cette catégorie d'approches est qu'il faut modifier l'algorithme de classification pour modifier la stratégie d'apprentissage de relations entre les labels.
Une autre stratégie pour construire un classifieur multi-labels consiste à appliquer des transformations sur les instances multi-labels pour se ramener au cas mono-label (Tsoumakas et Katakis (2007)). Cette catégorie d'approches possède deux principaux avantages :

— les classifieurs mono-label existants peuvent être utilisés pour gérer les sous problèmes de classification mono-label générés par la transformation des instances.
— il est possible de modifier la stratégie d'apprentissage de relations entre les labels en modifiant juste la méthode de transformation sans modifier l'algorithme de classification de base.

Les trois principales catégories d'approches de transformation du cas multi-labels au cas mono-label sont discutées dans la suite.

2.2.1 Approches basées sur la prédiction directe de sous-ensembles de labels

La classification multi-labels peut être transformée en classification mono-label en considérant chaque sous-ensemble de labels en tant qu'un nouveau label (Read (2008)). Soit C' l'ensemble de nouveaux labels, et soit $L : X \rightarrow C'$ la fonction qui associe à chaque instance de l'ensemble d'apprentissage $x_i \in X$ un nouveau label $L(x_i) \in C'$. Le nouveau label $L(x_i)$ correspond au sous-ensemble de labels y_i initialement associé à x_i. L'ensemble de nouveaux labels C' peut contenir au plus n labels différents $C' = \{c'_l\}_{1 \leq l \leq n}$ dans le cas où les sous-ensembles de labels $\{y_i\}_{1 \leq i \leq n}$ dans l'ensemble d'apprentissage X sont tous différents.
Soit $L^{-1} : C' \rightarrow X$ la fonction qui fournit pour un nouveau label $c'_l \in C'$ une instance $x_i \in X$ telle que $L^{-1}(c'_l) = x_i$ et $L(x_i) = c'_l$.
Un classifieur mono-label $h : a_1 \times \ldots \times a_p \rightarrow C'$ peut être construit à partir de X muni de la fonction de supervision L. Le classifieur multi-labels H fournit une prédiction pour une instance donnée $x \in a_1 \times \ldots \times a_p$ en faisant la conversion du nouveau label prédit $h(x)$ en un sous-ensemble de labels dans C. Le classifieur H est donné par : $H(x) = \lambda(L^{-1}(h(x)))$.
L'inconvénient de cette catégorie d'approches de transformation est que deux sous-ensembles de labels ayant des labels en commun sont considérés comme totalement

différents dans le problème transformé. Par conséquent, les relations entre les labels appartenant à deux sous-ensembles de labels différents ne peuvent pas être apprises.

2.2.2 Approches basées sur la prédiction intermédiaire de la présence de chaque label

Soit $\lambda_{c_l} : X \rightarrow \{0,1\}$ la fonction qui associe à chaque instance $x_i \in X$ la valeur 1 si le label c_l est associée à l'instance x_i ($c_l \in \lambda(x_i)$), et la valeur 0 sinon.

L'approche 'Binary Relevance' (BR) consiste à construire un classifieur multi-labels H à partir d'un ensemble de k classifieurs mono-labels $\{H_{c_l}\}_{1\leq l\leq k}$ (Tsoumakas et Katakis (2007)). Chaque classifieur $H_{c_l} : a_1 \times \ldots \times a_p \rightarrow \{0,1\}$ apprend de l'ensemble X muni de la fonction de supervision λ_{c_l} à prédire pour une instance donnée x si elle est associée au label c_l ($H_{c_l}(x) = 1$) ou non ($H_{c_l}(x) = 0$). Le classifieur multi-labels H est donné par : $H(x) = \{c_l, H_{c_l}(x) = 1\}_{1\leq l\leq k}$. L'inconvénient de l'approche BR est que les classifieurs $\{H_{c_l}\}_{1\leq l\leq k}$ sont tous indépendants. L'approche BR ne permet donc pas d'apprendre des relations de co-occurrence entre les labels.

L'approche 'Classifier Chains' (CC) est une extension de l'approche 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 est étendue telle que $x_i^e = (x_{i,a_1}, \ldots, x_{i,a_p}, \lambda_{c_1}(x_i), \ldots, \lambda_{c_k}(x_i))$. L'ensemble d'apprentissage X_l de chaque classifieur H_{c_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_{c_l}. Le classifieur $H_{c_2} : a_1 \times \ldots \times a_p \times b_1 \rightarrow \{0,1\}$ 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 $H_{c_1} : x = (x_{a_1}, \ldots, x_{a_p}, H_{c_1}(x_i))$ avant d'être reçue par le classifieur H_{c_2}. Chaque classifieur H_{c_l} a donc la possibilité de fournir des prédictions en se basant sur les prédictions des autres classifieurs qui le précèdent : $\{H_{c_{l'}}\}_{1\leq l'<l\leq k}$. L'inconvénient de l'approche CC est que les relations de co-occurrence 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-occurrence avec les labels $c_{l'}, l' > l$.

L'approche 'Aggregating Independent and Dependent classifiers' (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)). Le premier ensemble de classifieurs $\{h_{c_l}\}_{1\leq l\leq k}$ est construit par l'approche BR (chaque classifieur h_l est indépendant des autres classifieurs). Le deuxième ensemble de classifieurs $\{H_{c_l}\}_{1\leq l\leq k}$ est construit de façon similaire à l'approche CC. La différence est que l'ensemble d'apprentissage X_{c_l} pour le classifieur H_{c_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_{c_l} d'établir sa prédiction en se basant sur la présence ou l'absence des 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 $x^e = (x_{a_1}, \ldots, x_{a_p}, h_{c_1}(x), \ldots, h_{c_{l-1}}(x), h_{c_{l+1}}(x), \ldots, h_{c_k}(x))$ avant d'être passée en entrée au classifieur H_{c_l}. Chaque classifieur dépendant H_{c_l} fournit sa prédiction en

se basant sur les prédictions initiales $\{h_{c_{l'}}(x)\}_{l' \neq l}$ et en ignorant les prédictions finales des autres classifieurs $\{H_{c_{l'}}(x)\}_{l' \neq l}$. La prédiction du classifieur multi-labels H est donnée par : $H(x) = \{c_l, H_{c_l}(x) = 1\}_{1 \leq l \leq k}$.
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 $\{h_{c_{l'}}(x)\}_{l' \neq l}$ qui sont remplacées par les prédictions finales $\{H_{c_{l'}}\}_{1 \leq l' < l \leq k}$.

L'approche 'Pre-selection, Selection, and Interest of chaining based classifier' (PSI) permet de combiner les avantages des approches CC et AID (Laghmari et al. (2016)). Un ensemble de classifieurs initiaux $\{H_{c_l}\}_{1 \leq l \leq k}$ est construit de la même façon que les classifieurs dépendants dans l'approche AID. Ceci permet d'apprendre initialement les relations entre les labels sans restriction. Un ordre de prédiction est établi ensuite afin de fournir des prédictions cohérentes avec les relations apprises comme dans l'approche CC. Le fait qu'un classifieur H_{c_l} soit appris en considérant l'ensemble étendu d'attributs $A \cup C - \{c_l\}$ n'implique pas nécessairement que la prédiction de H_{c_l} dépend de tous les attributs $A \cup C - \{c_l\}$. L'ordre des classifieurs $\{H_{c_l}\}$ est établi tel que chaque classifieur H_{c_l} soit précédé par les classifieurs dont il dépend.
Le défi relevé par l'approche PSI est le cas d'une dépendance cyclique : par exemple, le cas où un classifieur H_{c_l} dépend de l'attribut $b_{l'}$, et le classifieur $H_{c_{l'}}$ dépend de l'attribut b_l. L'ordre de prédiction entre H_{c_l} et $H_{c_{l'}}$ ne peut pas être donné dans ce cas. L'approche PSI est basée sur trois mesures appelées pré-selection, sélection, et intérêt de chaînage qui permettent d'éliminer les dépendances cycliques. La mesure de pré-selection fournit l'ensemble de classifieurs impliqués dans une dépendance cyclique. La mesure de sélection sélectionne un classifieur à remplacer par un nouveau classifieur. La mesure d'intérêt de chaînage fournit l'ensemble d'attributs à considérer par le nouveau classifieur de façon à apprendre les relations entre les labels sans retomber dans une dépendance cyclique. Lorsqu'un classifieur est remplacé certaines dépendances cycliques disparaissent mais pas nécessairement toutes. Les trois mesures PSI sont appliquées itérativement jusqu'à l'élimination de toutes les dépendances cycliques.
L'approche PSI possède deux principaux avantages :

— même si certains classifieurs peuvent être appris deux fois pour éliminer une dépendance cyclique, seulement les k classifieurs finaux sont gardés par l'approche PSI.
— l'ordre de prédiction est établi après l'apprentissage des classifieurs. Ceci permet de ne pas empêcher l'apprentissage de certaines relations au préalable comme dans l'approche CC.

L'inconvénient de toutes les approches basées sur la prédiction intermédiaire de la présence de chaque label avant de fournir une prédiction multi-labels (BR, CC, AID, et PSI) est que les relations exprimant une préférence entre les labels ne sont pas apprises.

2.2.3 Approches basées sur la prédiction intermédiaire de la préférence entre chaque paire de labels

L'approche 'Ranking by Pairwise Comparisons' (RPC) est basée sur $\frac{1}{2}k(k-1)$ classifieurs $\{H_{c_l,c_{l'}}\}_{1\leq l<l'\leq k}$ permettant la prédiction d'une préférence entre chaque couple labels $(c_l, c_{l'})$ (Hüllermeier et al. (2008)).
L'ensemble d'apprentissage $X_{c_l,c_{l'}}$ du classifieur $H_{c_l,c_{l'}}$ est le sous-ensemble de X contenant uniquement les instances associées exclusivement à l'un des deux labels c_l ou $c_{l'}$: $X_{c_l,c_{l'}} = \{x_i \in X, (c_l \in y_i \ et \ c_{l'} \notin y_i) \ ou \ (c_l \notin y_i \ et \ c_{l'} \in y_i)\}$.
Soit $\lambda_{c_l,c_{l'}} : X \to \{c_l, c_{l'}, \emptyset\}$ la fonction donnée par :

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

La fonction $\lambda_{c_l,c_{l'}}$ fournit le label préféré entre c_l et $c_{l'}$ pour les instances de l'ensemble d'apprentissage. Le symbole $\emptyset$ indique que l'instance x_i n'appartient pas à l'ensemble d'apprentissage du classifieur $H_{c_l,c_{l'}}$. Chaque classifieur $H_{c_l,c_{l'}}$ apprend à partir de l'ensemble $X_{c_l,c_{l'}}$ muni de la fonction de supervision $\lambda_{c_l,c_{l'}}$ à prédire le label préféré entre c_l et $c_{l'}$ pour une instance donnée x.
Soit $V_{c_l} : a_1 \times \ldots \times a_p \to [\![0, k-1]\!]$ la fonction qui fournit pour chaque label $c_l \in C$ le nombre de fois où il a été préféré pour une instance x (le nombre de votes pour le label c_l) : $V_{c_l}(x) = |\{(c_{l'}, c_{l''}), H_{c_{l'},c_{l''}}(x) = c_l\}_{1\leq l'<l''\leq k}|$.
L'approche RPC ne fixe pas une méthode de prédiction et permet juste d'ordonner les labels selon le nombre de fois qu'ils ont été préférés par les classifieurs $\{H_{c_l,c_{l'}}\}_{1\leq l<l'\leq k}$. 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 :
$H(x) = \{c_l \in C, V_{c_l}(x) \geq v\}$.

L'approche 'Calibrated Label Ranking' (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 (Fürnkranz et al. (2008)). L'approche CLR introduit un label virtuel c_0 et apprend k classifieurs de plus $\{H_{c_l,c_0}\}_{1\leq l\leq k}$ par rapport à l'approche RPC. La fonction de supervision correspondant au classifieur H_{c_l,c_0} est donnée par :
$\lambda_{c_l,c_0}(x_i) = c_l$ si $c_l \in y_i$, et $\lambda_{c_l,c_0}(x_i) = c_0$ 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'avantage des approches basées sur l'apprentissage de préférences (RPC, et CLR) est que l'ensemble d'apprentissage est réduit pour chaque classifieur permettant de prédire une préférence. Ceci est aussi un inconvénient car cela empêche d'apprendre les relations de co-occurrence puisqu'il n'y a pas nécessairement assez d'instances en commun entres les ensembles d'apprentissage de deux classifieurs.

3 Nouvelle approche de classification multi-labels

Soit $X = \{x_1, x_2, x_3, x_4, x_5, x_6\}$ un ensemble d'apprentissage, et $C = \{c_1, c_2, c_3, c_4\}$ un ensemble de labels ($|C| = k = 4$). La Table 1 représente les fonctions de supervision liant les instances aux labels correspondants. Par exemple, le classifieur H_{c_1,c_3}

permettant de prédire la préférence entre c_1 et c_3 dans l'approche RPC (Section 2.2.3) est construit à partir du sous-ensemble d'apprentissage $X_{c_1,c_3} = \{x3, x6\}$ muni de la fonction de supervision λ_{c_1,c_3}. Les instances associées au symbole $\emptyset$ dans la Table 1 sont ignorées à l'étape d'apprentissage du classifieur H_{c_1,c_3}.

	λ_{c_1,c_2}	λ_{c_1,c_3}	λ_{c_1,c_4}	λ_{c_2,c_3}	λ_{c_2,c_4}	λ_{c_3,c_4}	λ_{c_1}	λ_{c_2}	λ_{c_3}	λ_{c_4}
x_1	c_1	$\emptyset$	$\emptyset$	c_3	c_4	$\emptyset$	1	0	1	1
x_2	c_1	$\emptyset$	$\emptyset$	c_3	c_4	$\emptyset$	1	0	1	1
x_3	c_1	c_1	c_1	$\emptyset$	$\emptyset$	$\emptyset$	1	0	0	0
x_4	c_2	$\emptyset$	$\emptyset$	c_2	c_2	$\emptyset$	0	1	0	0
x_5	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	1	1	1	1
x_6	$\emptyset$	c_3	$\emptyset$	c_3	$\emptyset$	c_3	0	0	1	0

TAB. 1: Exemple de données d'apprentissage

Un exemple d'une règle de décision basée sur une combinaison des relations de dépendance et de préférence est donné par : 'si pour une instance x, c_1 est préféré au c_2, et c_3 est associé à x, alors prédire que c_4 est associé à x'.
Afin d'apprendre cette règle de décision, le classifieur H_{c_4} permettant de prédire l'absence ou la présence du label c_4 doit considérer λ_{c_1,c_2} et λ_{c_3} dans la Table 1 en tant qu'attributs descriptifs supplémentaires. Cependant, certaines instances n'ont pas une valeur définie pour les attributs de préférences supplémentaires (symbole $\emptyset$).
Pour remédier à ce problème des valeurs manquantes, l'idée de notre approche est de les prédire en utilisant les classifieurs de préférences $\{H_{c_l,c_{l'}}\}_{1\leq l<l'\leq k}$ décrits dans l'approche RPC (Section 2.2.3). Ensuite, pour apprendre à la fois les relations de préférence et de dépendance, chaque classifieur $H_{c_l}, l \in [\![1, k]\!]$ est construit en considérant $\{\lambda_{c_{l'},c_{l''}}\}_{1\leq l'<l''\leq k}$ et $\{\lambda_{c_{l'}}\}_{l'\neq l}$ en tant qu'attributs descriptifs supplémentaires. Ceci risque de produire des dépendances cycliques qui peuvent être éliminées en utilisant l'approche PSI (Section 2.2.2). Notre approche appelée Stacked_RPC_PSI, est donc une combinaison des approches RPC et PSI permettant de tirer avantage à la fois des relations de préférence et de dépendance entre les labels, et en utilisant le minimum nombre de classifieurs binaires.

4 Expérimentation

4.1 Mesures de description des instances multi-labels

La performance des classifieurs multi-labels peut dépendre de la répartition des labels par rapport aux instances. Trois mesures sont souvent utilisées pour décrire la distribution des labels dans un jeu de données multi-labels (Tsoumakas et Katakis (2007)) :

— la cardinalité (label cardinality) qui évalue la moyenne du nombre de label associés à une instance : $\mathbb{LC} = \frac{1}{n}\sum_{i=1}^{n}|y_i|$.

Apprendre les relations de préférence et de co-occurrence entre les labels

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

TAB. 2: Données multi-labels.

— la densité (label density) qui évalue la moyenne du nombre de labels associés à une instance par rapport au nombre total de labels : $\mathbb{LD} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i|}{k} = \frac{\mathbb{LC}}{k}$.

— le nombre de combinaisons distinctes de labels (distinct label combinations) qui évalue le nombre de sous-ensembles de labels différents qui sont associés aux instances : $\mathbb{DLC} = |\{y_i\}_{1\leq i\leq n}|$.

4.2 Données et procédure d'expérimentation

Trois jeux de données multi-labels provenant de domaines différents sont utilisés pour comparer la performance de prédiction des classifieurs multi-labels (Table. 2). Le jeu de données des émotions (Trohidis et al. (2008)) contient un ensemble de 594 instances. Chaque instance représente un morceau de musique décrit par 72 attributs et associé à une ou plusieurs émotions parmi l'ensemble {amazed-suprised, happy-pleased, relaxing-calm, quiet-still, sad-lonely, angry-aggresive}. Le jeu de données des scènes (Boutell et al. (2004)) contient 2407 instances. Chaque instance représente une image décrite par 294 attributs et associée à un sous-ensemble de labels dans {Beach, Sunset, FallFoliage, Field, Mountain, Urban}. Le jeu de données des protéines contient 2417 instances décrites par 103 attributs (Elisseeff et Weston (2001)). Chaque protéine est associée à une localisation dite composant cellulaire. L'objectif est de prédire les localisations des protéines dans les cellules de levure.

La nouvelle approche introduite Stacked_RPC_PSI est comparée avec cinq autres approches existantes (AID, BR, CC, CLR, et PSI). Les arbres de décision (Quinlan (1993)) sont utilisés pour construire les classifieurs mono-labels de base. La méthode de validation croisée en 10 groupes est appliquée sur chaque jeu de données. La moyenne par rapport aux 10 plis pour les mesures d'évaluation de la prédiction est calculée pour chacun des trois jeux de données 'emotions', 'scenes', et 'yeast'.

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
CLR	0.45	0.54	0.59	0.17	0.24	0.55	**0.60**	0.83
PSI	**0.48**	0.56	0.61	**0.25**	0.24	0.59	0.58	0.84
Stacked_RPC_PSI	**0.48**	**0.57**	**0.62**	0.24	**0.23**	**0.60**	0.59	**0.86**

TAB. 3: Évaluation de la prédiction sur les données 'emotions'.

Approche	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
CLR	0.50	0.53	0.58	0.40	0.13	0.51	0.59	0.94
PSI	0.57	0.59	0.60	0.54	0.13	0.59	0.59	0.93
Stacked_RPC_PSI	**0.60**	**0.62**	**0.63**	**0.56**	**0.12**	**0.63**	0.62	0.94

TAB. 4: Évaluation de la prédiction sur les données 'scenes'.

Approche	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
CLR	**0.47**	**0.59**	**0.66**	0.10	**0.21**	**0.67**	**0.59**	**0.88**
PSI	0.43	0.53	0.60	0.15	0.26	0.58	0.53	0.84
Stacked_RPC_PSI	0.44	0.54	0.61	0.13	0.23	0.62	0.53	**0.88**

TAB. 5: Évaluation de la prédiction sur les données 'yeast'

4.3 Mesures d'évaluation de la qualité de prédiction

La qualité de prédiction peut être évaluée en se basant sur plusieurs mesures (Herrera et al. (2016)) :

La mesure de l'erreur de hamming (Hamming-loss) (Destercke (2014)) donnée par $\mathbb{HL} = \frac{|y_i \triangle H(x_i)|}{k}$, avec $y_i \triangle H(x_i)$ étant la différence symétrique entre l'ensemble correct de labels associés et l'ensemble de labels prédit donnée par :
$y_i \triangle\ H(x_i) = \{c_l \in y_i - H(x_i)\}_{1 \leq l \leq k} \bigcup \{c_l \in H(x_i) - y_i\}_{1 \leq l \leq k}$.
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'erreur de Hamming est une mesure très optimiste pour les données ayant une cardinalité et une densité faibles. Le classifieur multi-labels dans ce cas apprend à prédire des sous-ensembles de labels avec une cardinalité faible. Ainsi, même si aucun des labels prédits est effectivement associé à l'instance ($y_i \cap H(x_i) = \emptyset$) le nombre d'erreurs ($|y_i \triangle H(x_i)|$) reste petit par rapport au nombre maximal d'erreurs k parce que la cardinalité de l'ensemble y_i et de l'ensemble $H(x_i)$ est faible.

Le score de Hamming (closely related Hamming score) (Godbole et Sarawagi (2004)) n'est pas sensible à la cardinalité et à la densité des labels. Il mesure le nombre de labels prédits correctement par rapport au nombre de l'union des labels prédits et des labels effectivement associés à l'instance :
$\mathbb{CRHS} = \frac{|y_i \cap H(x_i)|}{|y_i \cup H(x_i)|}$.

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

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

La mesure F_β (van Rijsbergen (1974)) 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)) est une mesure d'évaluation de la qualité de prédiction adaptée aux données avec un déséquilibre de labels (présence de labels associés à presque toutes les instances, et des labels associés à très peu d'instances). En effet, toutes les mesures précédentes favorisent un classifieur qui prédit le label majoritaire en cas de déséquilibre de labels. La moyenne géométrique combine la précision positive donnée par $acc^+ = \frac{|\{c_l, c_l \in y_i \ et \ c_l \in H(x_i)\}_{1 \leq l \leq k}|}{|y_i|} = \mathbb{RECALL}$, et la précision négative donnée par $acc^- = \frac{|\{c_l, c_l \notin y_i \ et \ c_l \notin H(x_i)\}_{1 \leq l \leq k}|}{|C - y_i|}$ en une seule mesure donnée par : $\mathbb{GMEAN} = \sqrt{acc^+ \times acc^-}$.

La correspondance exacte (exact match) 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 associés à l'instance : $\mathbb{EM} = 1$ if $y_i = H(x_i)$, 0 sinon.

4.4 Résultats et discussion

Toutes les approches exploitant les relations entre les labels fournissent généralement des résultats meilleurs que l'approche BR qui ne permet pas l'apprentissage des relations entre les labels (Tables 3 à 5).

La nouvelle approche Stacked_RPC_PSI fournit les meilleurs résultats pour les données 'emotions' et 'scenes', mais pas pour les données 'yeast'. En effet les données 'yeast' présentent un déséquilibre dans la distribution de labels : certains labels sont très rares ou sont prédominants. L'approche CLR fournit les meilleurs résultats pour les données 'yeast' parce qu'elle est basée sur l'apprentissage de préférence utilisant juste un sous-ensemble réduit de l'ensemble d'apprentissage. L'effet du déséquilibre de la distribution de labels est donc réduit pour l'approche CLR. Les approches CC, AID, et PSI sont basées sur les dépendances entre les labels qui peuvent propager les erreurs de prédiction. Dans le cas où la prédiction d'un label dépend de la prédiction d'un label rare qui ne sera presque jamais prédit, l'erreur de prédiction pour le label rare peut être propagée pour les labels dépendants. La nouvelle approche Stacked_RPC_PSI fournit des résultats meilleurs que les approches CC, AID, et PSI parce qu'elle est basée aussi sur l'apprentissage de préférences. Ceci confirme l'hypothèse que la combinaison des relations de co-occurrence et des relations de préférence peut améliorer les prédictions.

5 Conclusion

Apprendre les relations entre les labels et les exploiter pour améliorer les prédictions est un défi intéressant dans la classification multi-labels. L'approche RPC permet l'apprentissage des relations de préférences entre les labels, et l'approche PSI permet l'apprentissage des relations de co-occurrence entre les labels sans restriction au

préalable. Ce travail introduit l'approche Stacked_RPC_PSI qui combine les deux approches RPC et PSI afin de bénéficier à la fois des relations de préférence et de co-occurrence pour améliorer la prédiction. L'expérimentation sur trois jeux de données montrent que l'approche Stacked_RPC_PSI est très compétitive avec les approches de l'état de l'art. L'inconvénient de l'approche Stacked_RPC_PSI est qu'elle est sensible au problème du déséquilibre de la distribution de labels (présence de labels rares ou de labels prédominants). Une idée qui pourrait réduire l'impact du déséquilibre de la distribution de labels dans l'approche Stacked_RPC_PSI consiste à éviter d'apprendre des dépendances par rapport aux classifieurs de préférence triviaux qui prédisent le label majoritaire.

Références

Agrawal, S., J. Agrawal, S. Kaur, et S. Sharma (2016). A comparative study of fuzzy pso and fuzzy svd-based rbf neural network for multi-label classification. *Neural Computing and Applications*, 1–12.

Bouchon-Meunier, B., C. Marsala, et M. Ramdani (1997). *Learning from Imperfect Data*. John Wiley & Sons.

Boutell, M. R., J. Luo, X. Shen, et C. M. Brown (2004). Learning multi-label scene classification. *Pattern Recognition 37*(9), 1757 – 1771.

Destercke, S. (2014). *Multilabel Prediction with Probability Sets : The Hamming Loss Case*, pp. 496 – 505. Cham : Springer International Publishing.

Elisseeff, A. et J. Weston (2001). A kernel method for multi-labelled classification. In *In Advances in Neural Information Processing Systems 14*, pp. 681–687. MIT Press.

Fürnkranz, J., E. Hüllermeier, E. Loza Mencía, et K. Brinker (2008). Multilabel classification via calibrated label ranking. *Machine Learning 73*(2), 133–153.

Gibaja, E. et S. Ventura (2015). A tutorial on multilabel learning. *ACM Comput. Surv. 47*(3), 52 :1–52 :38.

Godbole, S. et S. Sarawagi (2004). *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, pp. 22–30. Berlin, Heidelberg : Springer Berlin Heidelberg.

Herrera, F., F. Charte, A. J. Rivera, et M. J. del Jesus (2016). *Multilabel Classification Problem Analysis, Metrics and Techniques*, Chapter Multilabel Classification, pp. 17–31.

Hüllermeier, E., J. Fürnkranz, W. Cheng, et K. Brinker (2008). Label ranking by learning pairwise preferences. *Artificial Intelligence 172*(16–17), 1897 – 1916.

Kubat, M., R. Holte, et S. Matwin (1997). *Learning when negative examples abound*, pp. 146–153. Berlin, Heidelberg : Springer Berlin Heidelberg.

Laghmari, K., C. Marsala, et M. Ramdani (2016). Graded multi-label classification : Compromise between handling label relations and limiting error propagation. In *11th Inter. Conf. on Intelligent Systems : Theories and Applications (SITA)*, pp. 1–6.

Loza Mencía, E. et F. Janssen (2014). Stacking label features for learning multilabel rules. In S. Džeroski, P. Panov, D. Kocev, et L. Todorovski (Eds.), *Discovery Science - 17th Inter. Conf. DS 2014, Bled, Slovenia, October 8-10, 2014, Proceedings*, Volume 8777 of *Lecture Notes in Computer Science*, pp. 192–203. Springer.

Loza Mencía, E. et F. Janssen (2016). Learning rules for multi-label classification : a stacking and a separate-and-conquer approach. *Machine Learning 105*(1), 77–126.

Montañés, E., J. R. Quevedo, et J. J. del Coz (2011). Aggregating independent and dependent models to learn multi-label classifiers. In D. Gunopulos, T. Hofmann, D. Malerba, et M. Vazirgiannis (Eds.), *ECML/PKDD (2)*, Volume 6912 of *Lecture Notes in Computer Science*, pp. 484–500. Springer.

Quinlan, J. R. (1993). *C4.5 : Programs for Machine Learning*. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

Read, J. (2008). A Pruned Problem Transformation Method for Multi-label classification. In *Proc. 2008 New Zealand Computer Science Research Student Conference (NZCSRS 2008)*, pp. 143–150.

Sun, Z., Z. Guo, M. Jiang, X. Wang, et C. Liu (2016). *Research and Application of Fast Multi-label SVM Classification Algorithm Using Approximate Extreme Points*, pp. 39–52. Cham : Springer International Publishing.

Trohidis, K., G. Tsoumakas, G. Kalliris, et I. P. Vlahavas (2008). Multi-label classification of music into emotions. In J. P. Bello, E. Chew, et D. Turnbull (Eds.), *ISMIR*, pp. 325–330.

Tsoumakas, G. et I. Katakis (2007). Multi-label classification : An overview. *Int J Data Warehousing and Mining 2007*, 1–13.

van Rijsbergen, C. J. (1974). Foundations of evaluation. *Journal of Documentation 30*, 365–373.

Wang, X., S. An, H. Shi, et Q. Hu (2015). *Fuzzy Rough Decision Trees for Multi-label Classification*, pp. 207–217. Cham : Springer International Publishing.

Summary

In multi-label classification each instance can be associated to more than one label. For example, a music record can be associated to both labels 'happy' and 'relaxing'. Labels can be related with co-occurrence dependencies: for example, labels 'happy' and 'sad' can not be associated to the same music record. Labels can also be related with preference relations: for example, the label 'happy' is preferred over the label 'relaxing' to be associated to a music record containing several pikes. Label relations can help to better predict labels associated to instances. Existing approaches can learn either co-occurrence relations or preference relations. This work introduces an approach allowing to learn the two types of relations in order to improve the predictive performance. Experiments carried out show that the new introduced approach gives the best prediction results compared to five approaches from the state of the art.