Sélection et transformation de variables pour la classification Multi-Label par une approche MDL

Sènami C. Fréjus Ahomagnon *, Nicolas Voisine**
Marc Boullé**

*Polytech Nantes
jusberlin@gmail.com
**Orange Labs Lannion
prénom.nom@orange.com

Résumé. La classification multi-label est une extension de la classification supervisée au cas de plusieurs labels. Elle a connu un regain d'intérêt récent dans la communauté du machine learning de par son utilité dans plusieurs domaines. Comme pour tout problème de machine learning, le besoin de prétraiter les données multi-label est apparu comme une nécessité afin d'améliorer les performances des classifieurs. Dans cet article, nous introduisons une nouvelle méthode permettant de prétraiter des variables descriptives par discrétisation ou groupement de valeur, dans le cas de plusieurs labels à prédire. Le choix du meilleur prétraitement est posé comme un problème de sélection de modèle, et est résolu au moyen d'une approche bayésienne. Une étude comparative est réalisée avec d'autres méthodes de l'état de l'art afin de positionner la nouvelle méthode et de montrer l'intérêt de la sélection de variables pour la classification.

1 Introduction

Cet article se place dans le cadre de la classification multi-label où l'ont veut prédire un ensemble de labels pour une instance. La classification de sons, de vidéos et de textes fait partie de ces problèmes où un élément peut être associé à plusieurs labels. En prenant par exemple un article sur le Dalai Lama : en classification de texte, il peut être associé à la politique et à la religion. Suite à l'accroissement des gisements de données multi-label, le problème de la grande dimensionnalité dans les données s'est posé. Comme dans tout problème de machine learning, la mise en place de méthodes de prétraitement des données est donc apparue comme une nécessité. Des études variées [Dendamrongvit et al. (2011), Trohidis et al. (2008), Spolaôr et al. (2012),Zhao et al. (2010)] ont montré que la réduction de l'espace des variables explicatives pouvait être faite sans détériorer les performances des classifieurs.
Différentes études ayant abondé dans ce sens ont abouti à des méthodes de sélection de variables pour les classifieurs multi-label. Par analogie à la typologie des méthodes de classification multi-label (Madjarov et al., 2012), trois grandes familles de méthodes de sélection peuvent être dégagées de ces études. La première famille est celle des méthodes de sélection de variables par transformation du problème, où en transformant le problème de sélection multi-label en plusieurs problèmes de sélection mono-label (BR) ou multi-classe (LP), les résultats

obtenus en multi-classe/mono-label sont agrégés pour obtenir les résultats en multi-label. Cette première approche est la plus couramment utilisée dans la majorité des méthodes proposées. La deuxième famille est celle des méthodes par réduction de dimension où l'objectif est de déterminer un espace réduit des variables et/ou des labels qui facilite la tâche d'apprentissage. La troisième est celle des méthodes par adaptation du problème où les méthodes se basant sur des critères de sélection pour les données multi-classe sont adaptées pour le cas du multi-label.

Cet article présente les résultats d'un travail de recherche effectué sur les méthodes de sélection de variables pour les données multi-label. Nous proposons une nouvelle méthode ML-MODL de sélection et de transformation de variable pour la classification multi-label. Cette nouvelle méthode appartient à la troisième famille de méthodes de sélection de variables et est inspirée de la méthode supervisée de discrétisation de variables pour les données multi-classe MODL (Boullé, 2006). L'objectif est de positionner cette nouvelle méthode par rapport aux autres méthodes de l'état de l'art mais également de déterminer son impact sur les performances des classifieurs multi-label. L'article est organisé comme suit : dans un premier temps, nous introduisons la nouvelle méthode de sélection de variables ML-MODL, puis nous présentons l'analyse des expérimentations de cette méthode pour la classification multi-label ; enfin nous terminons par la conclusion et les perspectives de notre étude.

2 Approche MODL Multi Label

Pour des raisons de place, cette section décrit uniquement le cas de la discrétisation pour la classification multi-label. Il s'agit d'une généralisation de l'approche de discrétisation MODL multi-classe (Boullé, 2006).

Dans l'article nous utilisons les notations suivantes :

— N : nombre d'instances
— J : nombre de labels $l_1, l_2, \ldots, l_J$
— I : nombre d'intervalles
— N_i : nombre d'instances dans l'intervalle i
— $N_{.j}$: nombre d'instances de label j
— N_{ij} : nombre d'instances dans l'intervalle i du label j

Le critère d'évaluation C_D que nous proposons est le log négatif de la probabilité a posteriori du modèle connaissant les données.

Pour $I > 1$ C_D est égal à :

$$C_D(\text{Modèle ML}) = \log(2) + L(I-1) + \log\binom{N+I-1}{I-1} \quad (1)$$

$$+ J \times \sum_{1 \leq i \leq I} \log(N_i + 1) \quad (2)$$

$$+ \sum_{1 \leq i \leq I} \sum_{1 \leq j \leq J} \log \frac{N_i!}{N_{ij}!(N_i - N_{ij})!} \quad (3)$$

Les termes de la ligne (1) sont liés pour le premier à la présence ou non du modèle nul, pour le second au choix du nombre d'intervalles et pour le troisième aux choix de I intervalles parmi N instances. L est le recodage universel de Rissanen ((Rissanen, 1983)). Le terme (2)

décrit la distribution de chaque label pour chaque intervalle i, $1 \leq i \leq I$. Le terme (3) décrit la probabilité d'observer les données connaissant le modèle. Considérant les distributions des labels équiprobables et indépendantes, cela revient à calculer pour chaque label le nombre de distributions binomiales de N_i individus. Le coût du modèle nul $M_\emptyset$ défini pour $I = 1$ est égal à :

$$\begin{aligned} C_D(M_\emptyset) = & \log(2) + \\ & + J \times \log(N_i + 1) \\ & + \sum_{1 \leq j \leq J} \log \frac{N!}{N_{.j}!(N - N_{.j})!} \end{aligned}$$

Pour notre algorithme de recherche, nous utilisons un algorithme glouton qui partitionne récursivement la variable en deux parties en minimisant le coût C_D. Il est particulièrement adapté dans le cas d'un critère d'évaluation de bipartition, local à deux intervalles. La complexité algorithmique de MODL-ML est en $\mathcal{O}(n^2)$ dans le pire des cas. En pratique, si nous bornons le nombre d'intervalles I, la complexité est bornée par $\mathcal{O}(n * I)$.
Le critère de discrétisation permettent d'effectuer une sélection de variables de type filtre (en classant celles-ci par valeur de critère décroissante). Quand le meilleur critère est obtenu pour le modèle nul $I = 1$ alors il n'y a pas de discrétisation. Par conséquent nous ne sélectionnons pas la variable pour l'apprentissage du modèle. Ce qui fait de ML-MODL une méthode de sélection et de transformation sans paramètre.

3 Expérimentation

Cette section présente des résultats d'expérimentation permettant d'évaluer et comparer notre méthode de sélection de variables pour la classification multi-label.

3.1 Protocole

Pour évaluer notre méthode ML-MODL nous allons la comparer à plusieurs autres méthodes sur la base de différents critères. Pour notre étude, nous considérerons :

— l'ensemble DS de k jeux de données notés ds_1, ds_2,..., ds_k ;
— m méthodes de prétraitement notées M_1, M_2,..., M_m ;
— MLKNN (Zhang et Zhou, 2007) comme classifieur de référence ;
— Et l'ensemble E des critères d'évaluation : Ranking Loss, Accuracy et Root Means Square Error (RMSE)

Nous avons étudié les performances sur cinq classifieurs issus de Mulan (Tsoumakas et al., 2011) : Binary Relevance(BR, avec J48 comme classifieur de base), Classifier Chain (CC, avec J48 comme classifieur de base), PairWise(avec un Naive Bayes comme classifieur de base), Multi-label kNN (ML-KNN) et RAndom k labEL sets (RAKEL, avec du LabelPowerset sur du J48). Cependant au vu des résultats nettement supérieurs (avec ou sans sélection) de ML-KNN et du manque de place, nous ne présentons que les résultats de ML-KNN.

Le protocole consiste à exécuter en 10-cross validation l'enchaînement des tâches suivantes :

1. appliquer aux k jeux d'apprentissage du fold courant, les m méthodes de sélection et/ou de transformation de variables. Nous avons en sortie $m * k$ jeux d'apprentissage prétraités,
2. construire sur les $m * k$ jeux d'apprentissage prétraités et les jeux d'apprentissage non traités, le modèle de prédiction multi-label MLKNN en utilisant Mulan . Comme résultat, nous avons $m * k$ modèles de prédiction,
3. évaluer l'ensemble des modèles obtenus précédemment sur les données de test respectifs en se basant sur les 3 critères d'évaluation.

Le tableau 1 récapitule les quatres jeux de données de la littérature choisis. Quatre jeux de données bruitées ont été créé à partir de ceux d'origine. Le bruitage consiste à construire de nouvelles variables en permutant aléatoirement les lignes. Ainsi ces nouvelles variables ne sont plus corrélées aux labels. Nous joignons les variables d'origines avec celles bruitées. A part le nombre de variables qui double, les autres statistiques restent identiques.

	domain	**instances**	**nominal**	**numeric**	**labels**	**cardinalité**	**densité**	**distinct**
Birds	audio	645	2	258	19	1.014	0.053	133
Emotions	music	593	0	72	6	1.869	0.311	27
Scene	image	2407	0	294	6	1.074	0.179	15
Yeast	bology	2417	0	103	14	4.237	0.303	198

TAB. 1 – *jeux de données utilisés avec leurs caractéristiques*

Spolaôr et al. (2013) ont introduit quatre nouvelles méthodes de sélection de variables basées sur l'approche par transformation. Dans cet article, les auteurs utilisent les critères ReliefF (RF) et Information Gain (IG) afin d'évaluer les variables. Ces deux critères appliqués après les transformations LP et BR produisent les quatre méthodes RF-BR, IG-BR, RF-LP et IG-LP. Nous utilisons ces quatre méthodes de sélection pour les comparer à $ML - MODL$:

— $ML - MODL$: la nouvelle méthode ;
— $RF - BR$: approche BR se basant sur le critère ReliefF ;
— $IG - LP$: approche LP se basant sur le critère Information Gain ;
— $IG - BR$: approche BR se basant sur le critère Information Gain ;
— $RF - LP$: approche LP se basant sur le critère ReliefF.

3.2 Résultats

Le tableau 2 affiche les moyennes des critères sélectionnés sur les 8 bases de tests par méthode de sélection de variables. Il est à noter que le classifieur sans prétraitement donne déjà d'assez bon résultats. Nous constatons qu'en utilisant du MLKNN, $ML - MODL$ est la méthode qui obtient les meilleurs résultats moyens sur les 3 critères. Derrière MODL-ML, aucune autre méthode ne sort véritablement du lot.

De façon plus précise le tableau 3 affiche les moyennes des critères RMSE et Accuracy sur les 4 bases d'origine. Nous constatons que $ML - MODL$ domine les autres méthodes sur le critère RMSE et il se place juste derrière les meilleurs pour le critère Accuracy. Quand nous ajoutons du bruit, nous constatons sur le tableau 3 que le modèle MLKNN sans sélection baisse fortement ses performances. $ML - MODL$ est la seule méthode qui garde à l'identique

ses performances entre bases avec ou sans bruit. Les autres méthodes, à part $IG - BR$, ont des résultats qui décroissent significativement. Il est à noter (cf. tableau 2) que $IG - LP$ est la méthode qui sélectionne le moins de variable, au prix de moins bonne performance prédictive.

	no select	**RF-BR**	**RF-LP**	**IG-BR**	**IG-LP**	**ML_MODL**
Ranking Loss	0.137	0.129	0.127	0.126	0.133	**0.119**
Accuracy	0.525	0.553	0.564	0.555	0.544	**0,568**
Root Mean Squared Error	0.970	0.955	0.948	0.952	0.964	**0.937**
% Variables Sélectionnées	100	52.8	56.4	39.1	**29.3**	48.5

TAB. 2 – *Moyenne des critères sur les 8 jeux de données tests*

	no select		**RF-BR**		**RF-LP**		**IG-BR**		**IG-LP**		**ML_MODL**	
dataset	**Acc.**	**RMSE**	**Acc.**	**RMSE**	**Acc.**	**RMSE**	**Acc.**	**RMSE**	**Acc.**	**RMSE**	**Acc.**	**RMSE**
Emotions	0.533	0.909	0.533	0.913	0.542	0.909	0.543	0.911	0.530	0.925	**0.551**	**0.903**
Birds	**0.549**	0.846	0.532	0.849	0.547	0.845	0.536	0.845	0.507	0.888	0.542	**0.839**
Scene	0.667	0.612	0.663	0.620	**0.668**	0.623	0.665	0.619	0.661	0.616	0.667	**0.612**
Yeast	**0.516**	1.391	0.502	1.403	0.515	**1.391**	0.467	1.438	0.516	**1.391**	0.512	1.392
EmotionsNoise	0.453	0.959	0.539	0.908	**0.555**	**0.902**	0.543	0.911	0.530	0.925	0.551	0.903
BirdsNoise	0.501	0.898	0.505	0.895	0.499	0.898	**0.551**	0.840	0.478	0.911	0.542	**0.839**
SceneNoise	0.507	0.710	0.664	0.627	**0.670**	0.622	0.665	0.619	0.661	0.616	0.667	**0.612**
YeastNoise	0.473	1.438	0.487	1.426	**0.516**	**1.391**	0.467	1.438	0.473	1.438	0.512	1.392

TAB. 3 – *Performances par base non bruitées des méthodes de sélection de variables pour Accuracy et le RMSE en utilisant MLKNN comme classifieur*

4 Conclusion

Cet article a introduit ML-MODL qui est une nouvelle méthode sans paramètre de sélection et de transformation de variables pour la classification multi-label. Cette nouvelle méthode est une adaptation de la méthode MODL qui elle, est dédiée aux données multi-classe. Les résultats obtenus des différentes expérimentations montrent que :

— les performances d'un classifieur utilisant des données prétraitées par ML-MODL sont au moins aussi bonnes que celles d'un classifieur n'ayant pas bénéficié de la sélection de variables ;
— ML-MODL est la meilleure et la seule méthode de sélection de variables parmi celles évaluées, qui améliore ou au moins ne détériore pas les performances de MLKNN sur l'ensemble des critères d'évaluation retenues.
— la nouvelle méthode résiste au bruit car elle maintient ses performances en présence de bruit.

Les résultats montrent l'apport de la méthode $ML - MODL$ pour l'amélioration de la performance en classification multi-label. Dans des travaux futurs, nous étendrons cette méthode

avec des modèles de discrétisation multi-label plus expressifs, avec notamment la possibilité de choisir un sous-ensemble de labels informatifs pour chaque variable descriptive.

Références

Boullé, M. (2006). A Bayes optimal discretization method for continuous attributes. *Machine Learning 65*, 131–165.

Dendamrongvit, S., P. Vateekul, et M. Kubat (2011). Irrelevant attributes and imbalanced classes in multi-label text-categorization domains. *Intelligent Data Analysis 15*(6), 843–859.

Madjarov, G., D. Kocev, D. Gjorgjevikj, et S. Džeroski (2012). An extensive experimental comparison of methods for multi-label learning. *Pattern Recognition 45*(9), 3084–3104.

Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *Annals of Statistics 11*(2), 416–431.

Spolaôr, N., E. A. Cherman, M. C. Monard, et H. D. Lee (2013). A comparison of multi-label feature selection methods using the problem transformation approach. *Electronic Notes in Theoretical Computer Science 292*, 135–151.

Spolaôr, N., M. Monard, et H. Lee (2012). A systematic review to identify feature selection publications in multi-labeled data. *Relatório Técnico do ICMC No 374*(31), 3.

Trohidis, K., G. Tsoumakas, G. Kalliris, et I. P. Vlahavas (2008). Multi-label classification of music into emotions. In *ISMIR*, Volume 8, pp. 325–330.

Tsoumakas, G., E. Spyromitros-Xioufis, J. Vilcek, et I. Vlahavas (2011). Mulan: A java library for multi-label learning. *Journal of Machine Learning Research 12*(Jul), 2411–2414.

Zhang, M.-L. et Z.-H. Zhou (2007). Ml-knn: A lazy learning approach to multi-label learning. *Pattern Recogn. 40*(7), 2038–2048.

Zhao, Z., F. Morstatter, S. Sharma, S. Alelyani, A. Anand, et H. Liu (2010). Advancing feature selection research. *ASU feature selection repository*, 1–28.

Summary

The multi-label classification got recent interest in the machine learning community by its usefulness in many areas. As with any machine learning problem, the need to preprocess multi-label data has emerged as a need to improve the performance of learners. In this paper, we introduce a new method selection and variable processing for multi-label classification. This method is an adaptation of MDL criterion and is based on a Bayesian approach. A comparative study is made with other methods of the state of the art to position the new method but also to show interest of the features selection for the multi-label classification.