Un nouveau cadre de travail pour la classification associative dans les données aux classes disproportionnées

Dominique Gay*, Loïc Cerf**, Nazha Selmaoui-Folcher* et Jean-François Boulicaut**

*ERIM EA3791, PPME EA3325, Université de la Nouvelle-Calédonie
**Université de Lyon, CNRS
INSA-Lyon, LIRIS, UMR5205, F-69621, France

Résumé. Lorsque les classes sont disproportionnées les performances des classifieurs associatifs sont détériorées : *(i)*, les approches existantes sont biaisées vers la classe majoritaire et *(ii)*, la précision sur la classe minoritaire en pâtit. Dans cet article, nous proposons un nouveau cadre de travail plus approprié aux données dont les classes sont disproportionnées. Un nouveau classifieur associatif, `fitcare`, s'inscrit dans ce cadre de travail. Les expérimentations montrent que `fitcare` évite les deux problèmes majeurs liés aux classes disproportionnées. Notre nouvelle approche est aussi appliquée avec succès à des données géologiques réelles sur l'érosion des sols de Nouvelle-Calédonie.

1 Introduction

Les méthodes de classification supervisée à base de motifs (e.g., la classification associative) sont connues pour être efficaces, performantes et facilement interprétables (voir e.g. Bringmann et al. (2009) pour une vue d'ensemble). Le principe général est le suivant : un ensemble de motifs intéressants (selon des contraintes sur une mesure d'intérêt) est extrait des données d'apprentissage. Une sélection de cet ensemble est ensuite utilisée pour prédire la classe de nouveaux objets entrants. Toutefois, lorsque les classes sont disproportionnées, les classifieurs présentent un biais vers la classe majoritaire au détriment de la précision dans les classes minoritaires. Nous pensons que ces faiblesses sont dues au type même des approches existantes : elles sont `OVA` (One-Versus-All), i.e., étant donné un problème à p classes, les règles extraites sont caractéristiques d'une classe c_i par rapport au reste de la base (i.e., l'union des autres classes c_j $(j \neq i)$). Ainsi, la répartition des erreurs des règles dans les différentes classes c_j n'est pas prise en compte. Dans cet article, nous proposons un nouveau cadre de travail `OVE` (One-Versus-Each) dans lequel les motifs intéressants seront caractéristiques d'une classe c_i par rapport à chacune des autres classes c_j séparément.

2 Le cadre de travail `One-Versus-Each`

Afin de tenir compte de la distribution des classes et de la répartition des erreurs des motifs dans les différentes classes du problème, nous utilisons, pour chaque classe c_i, un seuil de fréquence minimum dans cette classe et un seuil de fréquence maximum pour chaque autre

classe c_j. Ainsi pour un problème à p classes, p^2 paramètres seuils sont nécessaires. Nous représentons les p^2 paramètres de manière concise dans une matrice Γ de la manière suivante :

$$\Gamma = \begin{pmatrix} \gamma_{1,1} & \gamma_{1,2} & \cdots & \gamma_{1,p} \\ \gamma_{2,1} & \gamma_{2,2} & \cdots & \gamma_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{p,1} & \gamma_{p,2} & \cdots & \gamma_{p,p} \end{pmatrix}$$

Sur une ligne i de Γ, nous retrouvons les seuils de fréquence et d'infréquence que doit respecter tout motif X caractérisant la classe c_i. $\gamma_{i,i}$ est la fréquence relative minimale de X dans c_i ; les $\gamma_{i,j}$ ($j \neq i$) sont les fréquences relatives maximales de X dans c_j. Étant donné Γ, les motifs à extraire sont définis de la manière suivante.

Définition 2.1 (Règle de caractérisation OVE) *Une règle d'association $X \rightarrow c_i$ est une règle de caractérisation OVE selon Γ pour la classe c_i, notée OVE-CR, ssi les trois conditions suivantes sont vérifiées :*

1. *X est fréquent dans r_{c_i}, i.e., $freq_r(X, r_{c_i}) \geq \gamma_{i,i}$*
2. *X est infréquent dans chaque autre classe, i.e., $\forall j \neq i, freq_r(X, r_{c_j}) < \gamma_{i,j}$*
3. *X est un corps minimal, i.e., $\forall Y \subset X, \exists j \neq i \mid freq_r(Y, r_{c_j}) \geq \gamma_{i,j}$*

Les contraintes suivantes sur Γ sont nécessaires à un cadre cohérent :

- Contrainte ligne : $\mathbb{C}_{ligne} \equiv \forall i \in \{1, \ldots, n\}, \forall j \neq i, \gamma_{i,j} < \gamma_{i,i}$
- Contrainte colonne : $\mathbb{C}_{colonne} \equiv \forall i \in \{1, \ldots, n\}, \forall j \neq i, \gamma_{j,i} < \gamma_{i,i}$

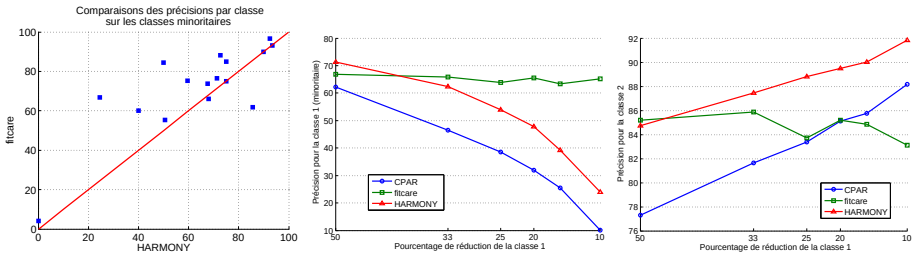
Intuitivement, une OVE-CR $X \rightarrow c_i$ sera caractéristique de c_i et non pertinente pour chaque autre classe c_j si $\gamma_{i,i}$ est plus grand que tous les $\gamma_{i,j}$ sur la ligne i de Γ ($\mathbb{C}_{ligne}$). Considérons d'autre part une OVE-CR $Y \rightarrow c_j$. On comprend que $\gamma_{j,i}$ doit être plus petit que $\gamma_{i,i}$, sans quoi Y pourrait être suffisamment fréquent dans r_{c_i} pour être pertinent pour c_i ($\mathbb{C}_{colonne}$). $\mathbb{C}_{colonne}$ garantit l'extraction d'un ensemble de règles non-conflictuelles.

Proposition 2.1 *Soient Γ une matrice satisfaisant $\mathbb{C}_{colonne}$ et S_Γ l'ensemble des OVE-CRs dont les corps respectent les seuils de fréquence et d'infréquence de Γ. Alors S_Γ est sans conflit de corps de règles, i.e., il n'existe pas, dans S_Γ, deux règles $X \rightarrow c_i$ et $Y \rightarrow c_j$ telles que $Y \subseteq X$ et $j \neq i$.*

3 fitcare

Dans le cadre OVE, nous avons développé une méthode de classification associative : fitcare. Seuls les points clés de la méthode détaillée dans Cerf et al. (2008) et Gay (2009) sont rappelés ici. fitcare est composé de trois algorithmes :

1. Il n'est pas concevable de positionner manuellement p^2 paramètres. fitcare ajuste Γ automatiquement en utilisant un algorithme d'optimisation basé sur l'approche hill-climbing. Il atteint ainsi un optimum local dans l'espace de recherche des paramètres.
2. Étant donné Γ, fitcare extrait, par niveaux, l'ensemble S_Γ des OVE-CRs.
3. Pour prédire la classe d'un nouvel objet entrant t, fitcare additionne les fréquences relatives, dans chaque classe, des règles de S_Γ qui s'appliquent à t. La plus grande somme indique la classe à prédire pour t.



(a) (b) (c)

FIG. 1 – *Comparaisons des performances sur données* `UCI` *et artificielles*

4 Expérimentations et applications

Afin de valider le nouveau cadre `OVE` et la méthode `fitcare`, nous la comparons avec des approches `OVA` récentes : `HARMONY` (Wang et Karypis, 2006) et `CPAR` (Yin et Han, 2003). De plus nous appliquons notre approche à des données réelles traitant de l'érosion des sols.

La figure 1(a) rapporte les résultats de précisions de `fitcare` et d'`HARMONY` sur 19 classes minoritaires issues de données `UCI`. Face à `HARMONY`, `fitcare` est gagnant 12 fois et égal 4 fois. Les points éloignés de la droite indiquent un grand écart de précision. Les figures b et c rapportent les précisions par classe obtenues sur la base de données `UCI-waveform`. Les trois classes, originellement équilibrées, sont ici artificiellement déséquilibrées. La figure 1(b) montre que les précisions de `CPAR` et `HARMONY` sur c_1 empirent lorsque cette classe se réduit. En parallèle, la figure 1(c) montre que les précisions de `CPAR` et `HARMONY` sur c_2 augmentent lorsque l'on réduit la classe c_1. Au contraire, les résultats de `fitcare` sont, quelque soit la classe, plutôt stables lorsque le déséquilibre augmente. Ceci confirme que dans le cadre `OVE`, `fitcare` évite le biais des approches `OVA` vers la classe majoritaire et est plus performant que ces concurrents sur les classes minoritaires.

Application aux données géologiques. Les données "érosion" sont constituées d'objets (les pixels représentants les zones du sol étudié) qui sont décrits par des attributs sélectionnés par les experts. On trouve notamment la pente, la pluviométrie, la nature du sol, sa couverture, etc. Chaque pixel est aussi étiqueté *sol érodé/sol non-érodé* – ce qui constitue la classe. Les sols érodés représentent 3% des données. On a donc bien un problème de classification dans des classes disproportionnées. Les `OVE`-CRs extraites de ces données confirment ce que pensent les experts sur les combinaisons de facteurs propices à l'érosion. De plus, il est possible de calculer des valeurs de mesures d'intérêt basées sur les fréquences des motifs extraits et, ainsi, de qualifier certains phénomènes d'érosion. Enfin, `fitcare` nous permet aussi de classer de nouvelles zones. La figure 2 rapporte graphiquement ces prédictions. Elles sont normalisées (entre 0 et 1) pour nous offrir une estimation de l'aléa érosion, i.e., de la probabilité d'apparition de l'érosion dans une zone en fonction de ses caractéristiques.

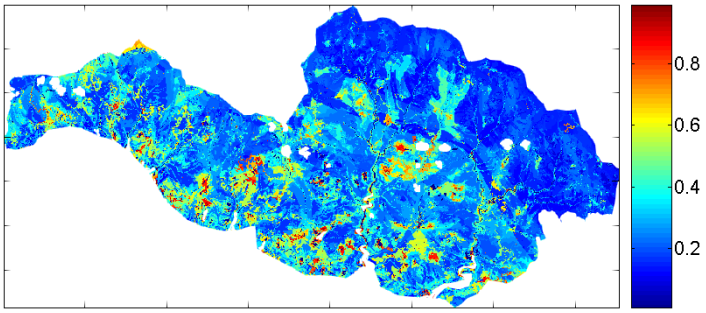


FIG. 2 – *Carte résultat de l'aléa érosion dans le bassin versant de la Dumbéa*

5 Conclusion

Dans cet article, un nouveau cadre de travail pour la classification associative dans les données multi-classes disproportionnées a été développé. Dans ce nouveau cadre OVE, fitcare montre de bonnes performances sur des données UCI artificielles et réelles. L'application aux données d'érosion des sols de Nouvelle-Calédonie confirme le bien-fondé et l'utilité du cadre et de la méthode.

Références

Bringmann, B., S. Nijssen, et A. Zimmermann (2009). Pattern based classification : a unifying perspective. In *LeGo'09 worskhop colocated with ECML/PKDD'09*.

Cerf, L., D. Gay, N. Selmaoui, et J.-F. Boulicaut (2008). A parameter free associative classifier. In *Proceedings DaWaK'08*, Volume 5182 of *LNCS*, pp. 238–247. Springer.

Gay, D. (2009). *Calcul de motifs sous contraintes pour la classification supervisée*. Ph. D. thesis, Université de la Nouvelle-Calédonie / INSA-Lyon.

Wang, J. et G. Karypis (2006). On mining instance-centric classification rules. *IEEE Transactions on Knowledge and Data Engineering 18*(11), 1497–1511.

Yin, X. et J. Han (2003). CPAR : Classification based on predictive association rules. In *Proceedings SIAM SDM'03*, pp. 369–376.

Summary

In this paper, a new framework for multi-class imbalanced classification problems is proposed. In the so-called One-Versus-Each framework, we also develop a new associative classifier. The application to real data sets shows the added-value of our approach.