Sous-échantillonnage topographique par apprentissage semi-supervisé

Mustapha Lebbah, Younès Bennani

LIPN UMR CNRS 7030, Université Paris 13,
99, avenue Jean-Baptiste Clément, 93430 Villetaneuse
prénom.nom@lipn.univ-paris13.fr

Résumé. Plusieurs aspects pourraient influencer les systèmes d'apprentissage existants. Un de ces aspects est lié au déséquilibre des classes dans lequel le nombre d'observations appartenant à une classe, dépasse fortement celui des observations dans les autres classes. Dans ce type de cas assez fréquent, le système d'apprentissage a des difficultés au cours de la phase d'entraînement liées au déséquilibre inter-classe. Nous proposons une méthode de sous-échantillonnage adaptatif pour traiter ce type de bases déséquilibrées. Le processus procède par le sous-échantillonnage des données majoritaires, guidé par les données minoritaires tout au long de la phase d'un apprentissage semi-supervisée. Nous utilisons comme modèle d'apprentissage les cartes auto-organisatrices. L'approche proposée a été validée sur plusieurs bases de données en utilisant les arbres de décision comme classificateur avec une validation croisée. Les résultats expérimentaux ont montré des performances très prometteuses.

1 Introduction

La plupart des algorithmes d'apprentissage sont basés sur deux hypothèses : La première est le critère à minimiser qui est le nombre d'erreurs. La deuxième est que les données d'apprentissage doivent êtres un échantillon représentatif de la population sur laquelle le modèle sera appliqué. Ces deux hypothèses ne sont pas respectées pour certains modèles quand ils sont construits à partir de données déséquilibrées. Nous pouvons l'illustrer par un exemple simple pris souvent en littérature : si 99% des données appartiennent à une seule classe, il sera difficile de faire mieux que le 1% d'erreur obtenue en classant tous les individus dans cette classe. Il convient donc de trouver d'autres solutions et hypothèses adaptées au problème de déséquilibre sans remettre en cause les fondements des algorithmes. Weiss (2003) propose de distinguer six catégories de problèmes liés aux données déséquilibrées, et à l'apprentissage des classes rares. Ces catégories sont (Marcellin et al. (2008)) : (**a**) Métriques inappropriées : dans ce cas, les mesures utilisées au cours du processus d'apprentissage ne sont pas adaptées aux classes déséquilibrées. (**b**) Manque "absolu" de données : ce problème est observé lorsque les données disponibles ne sont pas assez suffisantes pour définir clairement les frontières de la classe. (**c**) Manque "relatif" de données : c'est un problème similaire au manque absolu, sauf que dans ce cas ce manque est relatif à la taille de la base de données majoritaires. (**d**) Fragmentation des données : ce problème est lié aux algorithmes ayant une approche descendante, qui partent de l'espace de tous les individus et le partitionnent récursivement en sous-espaces

de plus en plus petits. (**e**) Marge d'induction inappropriée : Il s'agit de la marge appliquée à la règle apprise sur les données d'apprentissage pour pouvoir généraliser. (**f**) Données bruitées : Le bruit a plus d'impact sur les classes rares que sur les classes fréquentes. La distribution inégale des classes n'est pas le seul problème responsable de l'échec des algorithmes d'apprentissage. Plusieurs méthodes ont été proposées pour traiter les problèmes de déséquilibre, que la plupart regroupent en deux catégories principales. Au niveau algorithmique nous trouvons des méthodes qui tiennent intrinsèquement compte du déséquilibre en compensant les données sans altérer la distribution des classes, Raskutti et Kowalczyk (2004); Kubat et al. (1997); Barandela et al. (2003). Au niveau des données, les stratégies d'échantillonnage permettent d'équilibrer les données, ou de constituer des échantillons de manière à encourager les algorithmes d'apprentissage à converger vers un type de modèle spécifique. Deux catégories sont considérées : le sous-échantillonnage de la classe majoritaire et le sur-échantillonnage de la classe minoritaire. Le sur-échantillonnage à pour objectif de rééquilibrer les données en augmentant le nombre d'individus appartenant à la classe minoritaire, Chawla et al. (2002).

A l'inverse du sur-échantillonnage, un moyen pour rééquilibrer les données est la suppression d'un certain nombre d'individus appartenant à la classe majoritaire. Dans ce papier nous nous intéressons à cette catégorie d'algorithmes. La méthode la plus évidente et la plus simple est celle qui consiste à supprimer aléatoirement des individus de la classe majoritaire. Pour éviter les inconvénients du sous-échantillonnage aléatoire d'autres techniques proposent de guider l'échantillonnage de la classe majoritaire pour le rendre moins aveugle. Nous retrouvons principalement l'algorithme basé sur les liens de Tomek (Tomek (1976)). Considérons deux individus $\mathbf{x}_1$ et $\mathbf{x}_2$ appartenant respectivement à la classe i et à la classe j, et $d(\mathbf{x}_1, \mathbf{x}_2)$ la distance entre ces deux individus. La paire $(\mathbf{x}_1, \mathbf{x}_2)$ est un lien de Tomek s'il n'existe aucun individu $\mathbf{x}_3$ tel que $d(\mathbf{x}_3, \mathbf{x}_1) < d(\mathbf{x}_1, \mathbf{x}_2)$ ou $d(\mathbf{x}_3, \mathbf{x}_2) < d(\mathbf{x}_1, \mathbf{x}_2)$. Si ces deux individus forment un lien de Tomek, c'est que l'un des deux est du bruit, ou que les deux sont des points frontières. Nous trouvons aussi une autre technique qui se base sur la règle des plus proches voisins condensé (CNN :Condensed Nearest Neighbor), Hart (1968). D'autres méthodes consistent à combiner différents algorithmes de sous-échantillonnage, Batista et al. (2004); Marcellin et al. (2008).

2 Sous-Echantillonnage Topographique Adaptatif : Topographic Neighborhood Cleaning Rule (TNCR)

2.1 Règle de nettoyage avec voisinage

Cette méthode utilise la règle des plus proches voisins de Wilson pour supprimer des individus de la classe majoritaire ("Neighborhood Cleaning Rule" (NCR), Laurikkala (2001)). Après la sélection des trois voisins les plus proches pour chaque exemple $\mathbf{x}_i$ l'une des règles suivantes est appliquée : - si $\mathbf{x}_i$ appartient à la classe majoritaire ("**-**"), et les trois voisins les plus proches sont classés dans la classe minoritaire ("+"), alors l'observation $\mathbf{x}_i$ est supprimée. - si $\mathbf{x}_i$ appartient à la classe minoritaire, et les trois voisins les plus proches sont classés dans la classe majoritaire, alors les trois voisins les plus proches sont supprimés.

2.2 Quantification topographique et nettoyage adaptatif

Afin de mieux guider le sous-échantillonnage et de le rendre moins aveugle, l'utilisation des cartes auto-organisatrices SOM[1] (Kohonen (2001)), nous paraît une solution efficace pour choisir d'une façon intelligente les données à supprimer de la classe majoritaire en tenant compte de leurs topologies. La méthode proposée consiste à modifier l'algorithme d'apprentissage en proposant de supprimer à chaque itération les observations qui "gênent" les données minoritaires. L'approche consiste à intégrer les règles de nettoyage de l'algorithme NCR (§2.1) comme une troisième étape dans l'algorithme SOM. Cette règle sera appliquée localement au niveau de chaque cellule de la carte. Par conséquent, l'algorithme SOM sera utilisé dans le cas semi-supervisé, puisque les étiquettes associées à la classe positive ("+", minoritaire) seront utilisées. Ces étiquettes ne sont pas utilisées comme variable de la base, mais uniquement dans la phase de nettoyage. L'élimination au cours de l'apprentissage d'observations implique la modification de la base d'apprentissage $\mathcal{A}$ qui diminue au fur et à mesure des itérations ($\mathcal{A} = \{\mathbf{x}_i \in \mathcal{R}^n, i = 1..N\}$ où l'individu $\mathbf{x}_i = (x_{i1}, x_{i2}..., x_{ij}, ..., x_{in})$). Chaque observation $\mathbf{x}$ dispose d'une étiquette $label(\mathbf{x})$ positive ("+", données minoritaires) ou négative ("-", données majoritaires). L'étiquette négative "-" est utilisée uniquement dans la règle de nettoyage. On notera par la suite $\mathcal{A}^+$ l'ensemble des données positives et par $\mathcal{A}^-$ l'ensemble des données négatives. L'approche que nous proposons est donc une approche hybride : une action sur les données avec la phase de nettoyage par voisinage et une modification algorithmique de SOM. A l'opposé de NCR qui est dans l'obligation de supprimer des données de la classe majoritaire, notre approche TNCR ne l'est pas puisque le nettoyage par voisinage s'applique d'une manière locale au niveau de chaque cellule. Le modèle classique des cartes auto-organisatrices utilisé se présente sous forme d'une carte possédant un ordre topologique de C cellules. Les cellules sont réparties sur les nœuds d'un maillage. La prise en compte dans la carte de taille C de la notion de proximité impose de définir une relation de voisinage topologique. L'influence mutuelle entre deux cellules c et r est donc définie par la fonction $\mathcal{K}(\delta(c, r))$ où $\delta(c, r)$ est la distance de graphe entre les deux cellules c et r. A chaque cellule c de la grille $\mathcal{C}$ est associée un vecteur référent $\mathbf{w}_c = (w_{c1}, w_{c2}..., w_{ck}, ..., w_{cn})$ de dimension n. On note par la suite par $\mathcal{W} = \{\mathbf{w}_c, \mathbf{w}_c \in \mathcal{R}^n\}_{c=1}^{|W|}$ l'ensemble des référents associés à la carte. A chaque référent est associé un sous ensemble de données affectées à la cellule c qui sera noté P_c. L'ensemble des sous ensembles forment la partition de l'ensemble des données $\mathcal{A}$, $\mathcal{P} = \{P_1, ...P_c, ..., P_{|W|}\}$. Nous proposons de minimiser la fonction de coût suivante :

$$\mathcal{R}(\mathcal{A}, \chi, \mathcal{W}) = \sum_{\mathbf{x}_i \in \mathcal{A}} \sum_{r \in C} \mathcal{K}(\delta(\chi(\mathbf{x}_i), r)) ||\mathbf{x}_i - \mathbf{w}_r||^2 \quad (1)$$

Où χ affecte chaque observation $\mathbf{x}$ à une cellule unique de la carte $\mathcal{C}$.
Les phases principales de l'algorithme d'apprentissage TNCR sont :

- **Entrées** : Base d'apprentissage, $\mathcal{A}_0 = \mathcal{A}_0^- \cup \mathcal{A}_0^+$; le paramètre k : taille du voisinage
- **Sortie** :
  1. Référents de la carte $\mathcal{C}$.
  2. Base d'apprentissage sous échantillonnée $\mathcal{A}_{final}$ ($|\mathcal{A}_{final}| \leq |\mathcal{A}_0|$), telle que $|\mathcal{A}_{final}^+| = |\mathcal{A}_0^+|$.

[1] Self-Organizing Map

- **Phase d'affectation** : chaque observation $\mathbf{x}_i$ est affectée au referent $\mathbf{w}_c$, dont elle est la plus proche au sens de la distance euclidienne : $\chi(\mathbf{x}_i) = \arg\min_c \left(\|\mathbf{x}_i - \mathbf{w}_c\|^2\right)$
- **Phase d'adaptation** : les vecteurs référents sont mis-à-jour avec l'expression suivante :

$$\mathbf{w}_c = \frac{\sum_{r \in C} \mathcal{K}^T(\delta(c,r)) \sum_{\mathbf{x}_i \in \mathcal{A}_t, \chi(\mathbf{x}_i)=r} \mathbf{x}_i}{\sum_{r \in C} \mathcal{K}^T(\delta(c,r)) n_r} \quad (2)$$

- **Phase de nettoyage** : pour chaque cellule $c \in \mathcal{C}$ à l'itération t
  - si $\mathbf{x}_i \in \mathcal{A}_t^- \wedge k\text{-}ppv(\mathbf{x}_i) \subset P_{\chi(\mathbf{x}_i)} \subset \mathcal{A}_t^+$ alors $\mathcal{A}_{t+1} \leftarrow \mathcal{A}_t - \{\mathbf{x}_i\}$;
  - si $\mathbf{x}_i \in \mathcal{A}_t^+ \wedge k\text{-}ppv(\mathbf{x}_i) \subset P_{\chi(\mathbf{x}_i)} \subset \mathcal{A}_t^-$ alors $\mathcal{A}_{t+1} \leftarrow \mathcal{A}_t - k\text{-}ppv(\mathbf{x}_i)$;

Les trois phases permettent de minimiser la fonction de coût (eq. 1). Les deux premières phases sont similaires à l'algorithme de type nuées dynamiques classique. La troisième phase permet de minimiser la fonction de coût par rapport à $\mathcal{A}$. A la fin de l'apprentissage, la carte auto-organisatrice détermine une partition des données en $|\mathcal{W}|$ groupes associés à chaque référent $\mathbf{w}_c \in \mathcal{R}^n$ de la carte. Il est important de noter qu'en plus de la carte topologique, nous obtenons aussi une nouvelle base d'apprentissage de taille inférieure ou égale à la base initiale $\mathcal{A}$ ($|\mathcal{A}_{final}| \leq |\mathcal{A}|$).

3 Validation

Nous avons utilisé différents types de bases de données provenant du répertoire UCI, Asuncion et Newman (2007), qui sont utilisés de telle manière a avoir des degrés de déséquilibre variables pour évaluer notre approche (tableau 1). Plusieurs indices d'évaluation existent en littérature, mais pour nos expérimentations nous avons choisi de calculer deux indices synthétiques. Le premier est l'indice classique AUC "Area under curve" et un nouvel indice appelé IBA "Index of Balanced Accuracy", présenté par García et al. (2009). $IBA = (1 + (TPrate - TNrate)) \times (TPrate \times TNrate)$ où $TPrate$ et $TNrate$ indiquent respectivement le taux de vrai positives et le taux de vrai négatives. Tous les résultats présentés ci-dessous sont obtenus avec le paramètre de voisinage local $k = 3$.

Bases	Taille	Dim. (quanti, quali)	Taille min/maj	Classe % (min, maj)
Post-operative	90	8(1,7)	24,66	26.67, 73.33
Thyroid	215	5 (5,0)	30,185	13.95, 86.04
Ecoli	336	7 (7,0)	35,301	10.42, 89.58
Satimage	6435	36 (36,0)	626, 5809	9.72, 90.27
Glass	214	9 (9,0)	17, 197	7.94, 92.06
Flag	194	28(10,18)	17,177	8.76, 91.24

TAB. 1 – *Bases d'apprentissage. Les colonnes présentent : le nombre d'exemples, le nombre d'attributs (quantitatifs et qualitatifs), la classe minoritaire et majoritaire et leurs distribution*

Le tableau 2 présente les mesures AUC et IBA qui sont calculées dans le cas des arbres de décision. Les nombres entre parenthèses indiquent l'écart type calculé sur 100 expériences correspondant à une validation croisée, en divisant la base en 10 sous ensembles et répétant ce

Bases	AUC (%)		IBA (%)		TS (%)	
	NCR	TNCR	NCR	TNCR	NCR	TNCR
Post-operative	45.62 (16.49)	**47.89** (18.46)	9.89 (10.92)	**12.08** (18.56)	75.76	46.97
Thyroid	**95.84** (6.51)	95.12 (8.58)	80.64 (27.95)	**81.85** (27.71)	4.87	7.03
Ecoli	83.70 (12.24)	**84.46** (12.33)	25.42 (19.37)	**26.69** (18.78)	14.29	19.61
Satimage	82.30 (2.42)	**83.42** (2.21)	24.05 (5.21)	**24.92** (4.87)	10.61	13.74
Flag	**61.61** (19.32)	61.18 (18.66)	12.10 (23.10)	**14.20** (25.84)	25.43	38.42
Glass	97.37 (7.58)	**97.62** (7.61)	85.92 (28.44)	**91.26** (24.84)	7.11	12.7

TAB. 2 – *AUC (%), IBA (%) et TS (%) calculés dans le cas des arbres de décision. Les nombres entre parenthèses correspondent à l'écart type calculé après une validation croisée. AUC : Area under curve. IBA : Index of Balanced Accuracy. TS (%) :Taux des données supprimées après application de l'algorithme NCR et TNCR.* 100% *correspond à la taille de la classe majoritaire*

processus 10 fois. L'analyse des résultats permet en premier lieu de confirmer que le sous-échantillonnage topographique et adaptatif permet d'obtenir de meilleurs performances en terme de l'indice AUC ou IBA, sur la plupart des bases de données. Nous avons constaté une légère baisse pour la base Flag et la base thyroid en observant uniquement l'indice AUC. Cette baisse est due uniquement à la faible valeur du taux $TNrate$ sur les deux bases. Par exemple pour la base Flag, $TNrate$ passe de 93.59% avec NCR à 93.33% avec TNCR, par contre concernant la classe positive, $TPrate$ passe de 20.82% (NCR) à 22.82% (TNCR). Ceci se traduit par une augmentation de l'indice IBA qui donne un avantage à la classe positive. Pour mieux comprendre le comportement des deux méthodes TNCR et NCR, nous avons calculé le taux de suppression (TS) obtenu à la fin de l'apprentissage semi-supervisé (table 2). Nous observons clairement que la méthode TNCR fournit majoritairement un taux de suppression élevé en le comparant à celui de NCR. Nous avons constaté que nous atteignons des performances meilleures ou similaires avec moins de données que la méthode NCR.

4 Conclusion et perspectives

Nous nous sommes intéressés dans ce travail au problème de déséquilibre des classes et aux différentes méthodes et solutions existantes. Ensuite nous avons présenté une approche de sous-échantillonnage qui se base sur les cartes topologiques. Cette solution guide le choix des données à supprimer dans un voisinage local, en prenant en considération la distribution et la topologie des données. Une série d'expériences ont été réalisées pour valider la méthode proposée. Les résultats obtenus ont été comparés avec une méthode de sous échantillonnage connue qui nous a permis de mieux évaluer notre approche qui s'est avérée prometteuse comme solution au problème de déséquilibre des classes. Les perceptives issues de ce travail touchent un grand nombre d'étapes. Dans un premier temps, nous comptons comparer notre approche avec d'autres méthodes de sous et sur échantillonnage et étudier l'influence du recouvrement des données sur les performances. Dans un deuxième temps, nous envisageons étudier les performances de la méthode TNCR en présence d'autres méthodes de classement type SVM ou simplement le k-ppv. D'autres indices d'évaluation seront aussi étudiés, Hand (2009).

Références

Asuncion, A. et D. Newman (2007). UCI machine learning repository. http ://www.ics.uci.edu/∼mlearn/MLRepository.html.

Barandela, R., J. S. Sánchez, V. García, et E. Rangel (2003). Strategies for learning in class imbalance problems. *Pattern Recognition 36*(3), 849–851.

Batista, G. E. A. P. A., R. C. Prati, et M. C. Monard (2004). A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explor. Newsl. 6*(1), 20–29.

Chawla, N. V., K. W. Bowyer, et P. W. Kegelmeyer (2002). Smote : Synthetic minority oversampling technique. *Journal of Artificial Intelligence Research 16*, 321–357.

García, V., R. A. Mollineda, et J. S. Sánchez (2009). Index of balanced accuracy : A performance measure for skewed class distributions. In H. Araújo, A. M. Mendonça, A. J. Pinho, et M. I. Torres (Eds.), *IbPRIA*, Volume 5524 of *Lecture Notes in Computer Science*, pp. 441–448. Springer.

Hand, D. (2009). Measuring classifier performance : a coherent alternative to the area under the roc curve. *Machine Learning 77*(1), 103–123.

Hart, P. (1968). The condensed nearest neighbor rule. *IEEE Trans. on Inform. Th. 11*, 515–516.

Kohonen, T. (2001). *Self-organizing Maps*. Springer Berlin.

Kubat, M., R. Holte, et S. Matwin (1997). Learning when negative examples abound. In *ECML '97 : Proceedings of the 9th European Conference on Machine Learning*, London, UK, pp. 146–153. Springer-Verlag.

Laurikkala, J. (2001). Improving identification of difficult small classes by balancing class distribution. In *AIME '01 : Proceedings of the 8th Conference on AI in Medicine in Europe*, London, UK, pp. 63–66. Springer-Verlag.

Marcellin, S., D. A. Zighed, et G. Ritschard (2008). Evaluating decision trees grown with asymmetric entropies. In A. An, S. Matwin, Z. W. Ras, et D. Slezak (Eds.), *ISMIS*, Volume 4994 of *Lecture Notes in Computer Science*, pp. 58–67. Springer.

Raskutti, B. et A. Kowalczyk (2004). Extreme re-balancing for svms : a case study. *SIGKDD Explor. Newsl. 6*(1), 60–69.

Tomek, I. (1976). An experiment with the edited nearest-neighbor rule. *IEEE Transactions on Systems, Man, and Cybernetics 6*(6), 448–452.

Weiss, G. M. (2003). *The effect of small disjuncts and class distribution on decision tree learning*. Ph. D. thesis, New Brunswick, NJ, USA. Director-Hirsh, Haym.

Summary

Several aspects could affect the existing machine learning algorithm. One of these is related imbalance classes in which the number of observations belonging to a class, greatly exceeds the observations in other classes. We propose a method of adaptive subsampling to treat this type of imbalanced databases. The process proceeds by subsampling of the majority class guided with minority data (semi-supervised learning). We use as a learning model the self-organizing maps. The proposed approach has been validated on multiple databases using decision trees as a classifier with cross validation. The experimental results showed very promising performance.