Sélection d'échantillons d'apprentissage pour ensembles de classifieurs basée sur le concept de marge

Samia Boukir[1], Li Guo[1]

[1]ENSEGID/IPB - Université de Bordeaux - Laboratoire G&E, 1 Allée F. Daguin, 33607 Pessac Cedex
samia.boukir@egid.u-bordeaux3.fr, li.guo@egid.u-bordeaux3.fr

Résumé – L'échantillon d'apprentissage est un élément essentiel en apprentissage supervisé. Traditionnellement, toutes les instances de la base d'apprentissage sont traitées de la même façon dans le processus d'apprentissage, et autant d'instances que possible sont utilisées pour créer le classifieur. Dans cet article, nous proposons une nouvelle approche pour mesurer l'importance de chaque instance dans le processus d'apprentissage basée sur la théorie de la marge des méthodes d'ensemble. Nous montrons que les instances de faible marge ont une influence majeure sur la constitution d'échantillons d'apprentissage appropriés pour la construction de classifieurs fiables. Les résultats expérimentaux montrent que notre méthode permet non seulement de réduire de façon significative la taille de l'échantillon d'apprentissage mais d'améliorer aussi la précision de classification.

Abstract – The training set is an essential element in supervised learning. Traditionally, all instances in the training set are treated equally in the learning process and as many instances as possible are used to create a classifier. In this paper, we introduce a new approach to measure the importance of each instance in the learning process, based on the margin theory of ensemble methods. We show that smaller margin instances have a major influence in forming an appropriate training set to build up a reliable classifier. Experimental results show that our method not only significantly reduces the training set size but also increases the classification accuracy.

1 Introduction

L'échantillon d'apprentissage est un élément essentiel en apprentissage supervisé. Traditionnellement, toutes les instances de la base d'apprentissage sont supposées jouer le même rôle et ainsi sont traitées de la même façon dans le processus d'apprentissage. Pour appréhender les problèmes de classification de manière efficace, chaque instance devrait être traitée de façon différente puisque chacune joue un rôle différent dans la construction du classifieur.

Certaines méthodes d'apprentissage automatique considèrent déjà leurs instances d'apprentissage différemment, telles que les Machines à Vecteurs de Support (*Support Vector Machines* ou SVM) [9] et le *boosting* [5], deux méthodes de référence en apprentissage automatique. Ces approches puissantes se focalisent effectivement sur les instances les plus *importantes*, de différentes façons. Néanmoins, elles ne mesurent pas explicitement la pertinence ou *importance* de ces instances clés vis à vis du processus d'apprentissage.

Dans cet article, nous proposons un nouveau concept : *importance d'instance*, basé sur la marge des méthodes d'ensemble, qui vise à traiter différemment les instances d'apprentissage dans le processus d'apprentissage. Ce concept est à la base de la méthode d'apprentissage d'ensemble que nous avons élaborée, qui fournit un échantillon d'apprentissage pertinent pour les méthodes d'ensemble.

2 Importance d'instance d'apprentissage basée sur la marge d'ensemble

2.1 Marge des méthodes d'ensemble

La méthode d'ensemble est un paradigme populaire d'apprentissage qui construit un modèle de classification en intégrant des composants d'apprentissage multiples [4]. Le *bagging* est l'une des méthodes d'ensemble les plus performantes [1]. L'idée de base réside dans le fait que les différentes répliques de la base d'apprentissage sont légèrement différentes de l'originale mais suffisamment diverses pour obtenir des classifieurs différents qui vont pouvoir être combinés.

La marge d'ensemble est un concept fondamental des méthodes d'ensemble [8]. Le fait que ce soit la marge d'une classification plutôt que l' erreur brute d'apprentissage qui compte est devenu un facteur clé dans l'analyse et l'utilisation de classifieurs ces dernières années. Nous utilisons notre propre définition de la marge [7] qui s'exprime par l'équation 1, où c_1 est la classe de vote majoritaire pour l'instance x et v_{c_1} le nombre de votes attribués à cette classe, c_2 est la seconde classe la plus populaire et v_{c_2} le

nombre de votes correspondant.

$$marge(x) = \frac{v_{c_1} - v_{c_2}}{\sum_{c=1}^{L}(v_c)} = \frac{\max_{c=1,\ldots,L}(v_c) - \max_{c=1,\ldots,L \cap c \neq c_1}(v_c)}{\sum_{c=1}^{L}(v_c)} \quad (1)$$

2.2 Importance d'instance d'apprentissage

Nous introduisons ici un nouveau concept inspiré de l'importance de variable de Breiman [2] qui est de plus en plus utilisée pour la sélection de variables pour la classification [6]. Au lieu de mesurer l'impact de variables sur la performance de classification, nous proposons d'évaluer ici l'influence ou *importance* d'instances d'apprentissage sur le comportement d'un classifieur. l'Importance $\mathcal{I}$ d'un sous-ensemble d'apprentissage s vis à vis d'un classifieur f peut être définie par l'équation 2, où S est l'échantillon d'apprentissage dans sa totalité, $\mathcal{E}$ est une fonction d'évaluation de classifieurs, f_S est un classifieur entrainé par toutes les instances d'apprentissage, f_{S-s} dénote le même classifieur mais entrainé sans les exemples du sous-échantillon s.

$$\mathcal{I}(s)_f = \mathcal{E}(f_S) - \mathcal{E}(f_{S-s}) \quad (2)$$

Comme un exemple seul aurait généralement une influence négligeable sur la performance d'un classifieur, nous suggérons de mesurer plutôt l'importance d'un sous-ensemble de l'échantillon d'apprentissage. Les instances de ce sous-échantillon devraient être homogènes. Pour cela, nous utilisons la marge d'ensemble comme critère d'ordonnancement des instances d'apprentissage.

2.3 Importance de sous-échantillon d'apprentissage

La marge d'une instance est une distance de la frontière de décision à cette instance. Plus la marge est faible, plus la quantité d'information fournie par l'instance correspondante est significative, et plus elle a d' importance. Toutes ensemble, les valeurs de marge peuvent révéler les propriétés de distribution de l'échantillon d'apprentissage correspondant. Nous proposons une nouvelle méthode pour mesurer l'importance d'un sous-échantillon d'apprentissage, basée sur la marge, qui passe par les étapes suivantes :

1. Calculer la marge de chaque instance d'apprentissage.
2. Trier les instances d'apprentissage en fonction de leurs valeurs de marge, par ordre croissant.
3. Composer des sous-ensembles consécutifs disjoints de taille n à partir des instances d'apprentissage ordonnées.
4. Calculer l'importance de chaque sous-échantillon d'apprentissage en utilisant l'équation 2 avec comme fonction d'évaluation de classifieur la précision globale de classification.

Pour évaluer notre méthode, nous avons utilisé le *bagging* pour créer un ensemble, et les arbres de classification et de régression (*Classification and Regression Trees (CART)*) [3] comme classifieurs de base.

2.4 Résultats experimentaux

Nous avons utilisé 100 CART élagués (*pruned*) dans nos expérimentations. La figure 2 montre les résultats d' importance de sous-échantillon d'apprentissage sur 3 jeux de données de la base UCI (voir table 1). La taille n des sous-échantillons a été fixée à 10% de celle de l'échantillon d'apprentissage. Il apparait clairement que le sous-échantillon le plus important des bases d'apprentissage des 3 jeux de données est celui de plus faible marge. Les sous-échantillons les plus importants suivants sont respectivement le 2ème et le 3ème de plus faible marge. Ces résultats démontrent l'impact des instances de faible marge sur la constitution d'échantillons d'apprentissage pertinents qui concerne ici au moins 30% de chaque jeu de données.

3 Une nouvelle méthode d'apprentissage d'ensemble

Les instances de faible marge jouent un rôle plus important que celles de forte marge dans la construction d'ensembles de classifieurs. En outre, il est plus difficile de modifier la classification d'instances de forte marge que celle de faible marge. Pour améliorer la précision de classification d'une méthode d'ensemble, il faudrait donc se focaliser sur la performance de celle-ci sur les instances de faible marge. Dans la suite, nous montrons que la taille de l'échantillon d'apprentissage d'un ensemble de classifieurs peut être réduite de façon significative, tout en améliorant la précision, en lui retirant ses instances de plus forte marge.

3.1 Elagage d'instances d'ensemble

Nous proposons une méthode itérative d'élagage (*pruning*) d'instances d'apprentissage pour ensembles de classifieurs, qui exploite notre concept d'importance de sous-échantillon d'apprentissage basé sur la marge. Elle consiste en les étapes suivantes :

1. Créer un ensemble de classifieurs E avec l'échantillon d'apprentissage S.
2. Calculer la valeur de marge de chaque instance d'apprentissage.
3. Evaluer E sur un échantillon de validation V et obtenir le taux d'erreur de classification ϵ de E.
4. Supprimer les M instances de plus forte marge pour constituer un nouvel échantillon d'apprentissage S.

Jeux	Appren.	Valida.	Test	Attributs	Classes
Letter	5000	2000	5000	16	26
Connect-4	5000	2000	5000	42	3
Waveform	2000	1000	2000	21	3
Optdigits	1400	1000	1400	64	10
Pendigit	2000	1000	2000	16	10
Statlog	2000	2000	2000	35	6
Segment	900	510	900	19	7

TABLE 1 – Jeux de données

5. Si la taille de $S > 0$, alors aller à l'étape 1, sinon choisir l'ensemble E^* ayant le taux d'erreur le plus faible comme le meilleur ensemble, S^* étant l'échantillon d'apprentissage réduit associé.

Pour valider notre méthode, nous avons utilisé, comme pour l'évaluation de notre concept d'importance de marge, le *bagging* pour créer un ensemble, et les arbres de classification et de régression (*Classification and Regression Trees (CART)*) comme classifieurs de base.

3.2 Complexité

Le complexité du *bagging* classique, qui utilise tout l'échantillon d'apprentissage, est en $O(TN\log(N))$, où N est le nombre de données d'apprentissage et T le nombre d'arbres de décision de l'ensemble. La complexité de notre méthode d'apprentissage basé sur un *bagging* sélectif, favorisant les instances de plus faible marge, est en $O(\frac{1}{M}TN\log(N))$, $0 < M < 1$, M étant le pourcentage d'instances à supprimer de l'ensemble d'apprentissage à chaque étape d'élagage.

3.3 Résultats expérimentaux

Comme dans l'analyse empirique précédente, nous avons validé notre algorithme sur 7 jeux de données de la base UCI (voir table 1), et nous avons utilisé 100 CART élagués (*pruned*). La taille M du sous-ensemble d'instances à supprimer à chaque étape d'élagage a été fixée à 5% de celle de l'échantillon entier d'apprentissage.

La figure 1 montre la courbe de précision de classification en fonction de la taille de l'échantillon d'apprentissage sélectionné par notre méthode sur l'échantillon de test du jeu de données *Letter*. La courbe de précision de classification croît de façon monotone avec l'augmentation du pourcentage d'instances supprimées de l'échantillon entier d'apprentissage jusqu'à atteindre un pic de précision, elle décroît alors de façon monotone. Nous avons utilisé moins de 40% de l'échantillon d'apprentissage tout en augmentant la précision de classification d'environ 2%.

La table 2 présente la moyenne et l'écart type (issus de 10 passes de calcul) de la précision de classification obtenue sur la base de test par l'ensemble de classifieurs sélectionné qui a engendré la précision maximale sur l'échantillon de validation, ainsi que la taille de l'échantillon d'apprentissage élagué associé (taux d'instances de plus faible marge de l'échantillon entier d'apprentissage). Les résultats du *bagging* classique, qui utilise tout l'échantillon d'apprentissage, sont montrés sur cette table, pour comparaison.

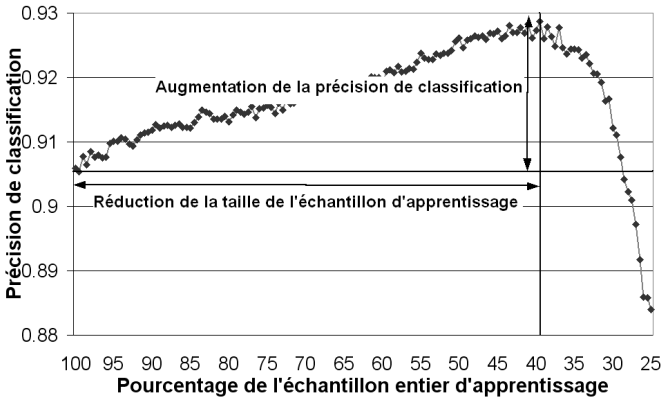


FIGURE 1 – Précision de classification en fonction de la taille de l'échantillon d'apprentissage du jeu de données *Letter*.

Jeu	Préc. E (%)	Préc. R (%)	Taille R (%)
Letter	90.59 ± 0.20	92.82 ± 0.34	38.9 ± 3.26
Connect-4	73.28 ± 0.26	75.93 ± 0.43	42.9 ± 3.85
Waveform	82.04 ± 0.42	84.37 ± 0.32	46.6 ± 5.41
Optdigits	94.22 ± 0.43	96.60 ± 0.50	36.1 ± 13.53
Pendigit	97.01 ± 0.07	98.34 ± 0.25	19.1 ± 3.94
Statlog	87.70 ± 0.34	89.58 ± 0.21	43.0 ± 13.68
Segment	97.15 ± 0.31	98.36 ± 0.19	17.5 ± 4.53

TABLE 2 – Précision (*moyenne* $\pm$ *écart type*) des ensembles de classifieurs avec échantillons entier (E) et réduit d'apprentissage (R), ainsi que la taille de ce dernier

La table 3 présente la précision de l'ensemble optimal pour la classe la plus difficile sur la base de test. Les résultats du *bagging* classique sont également montrés sur cette table.

Ainsi, notre méthode améliore non seulement la précision globale de classification (de 2% en moyenne) par rapport au *bagging* classique, mais augmente aussi de façon significative la précision sur la classe la plus difficile (jusqu'à 15%), et ce, tout en utilisant un échantillon d'ap-

Jeu	Préc. E (%)	Préc. R (%)
Letter	83.63 ± 1.83	87.72 ± 0.93
Connect-4	46.25 ± 1.22	61.11 ± 2.40
Waveform	76.50 ± 0.70	78.78 ± 0.95
Optdigits	87.58 ± 1.64	96.24 ± 2.06
Pendigit	90.00 ± 0.78	96.68 ± 1.16
Statlog	52.53 ± 1.42	55.90 ± 1.61
Segment	91.07 ± 1.97	94.54 ± 0.69

TABLE 3 – Précision (*moyenne* $\pm$ *écart type*) de la classe la plus difficile des ensembles de classifieurs avec échantillons entier (E) et réduit d'apprentissage (R)

prentissage réduit d'au moins 50% (de plus de 80% pour 2 des 7 jeux de données). Notre algorithme d'élagage d'instances s'avère donc une technique efficace de sélection de données d'apprentissage pour *booster* la performance de l'ensemble de classifieurs sur les classes mineures, avec un impact notable sur la précision globale de classification.

4 Conclusion

Nous avons développé une nouvelle approche pour mesurer l'importance de chaque instance dans le processus d'apprentissage. Cette méthode révèle l'influence majeure des instances de faible marge sur la construction de classifieurs fiables. Cette mesure est à la base d'une nouvelle méthode d'apprentissage qui fournit un échantillon d'apprentissage pertinent pour les méthodes d'ensemble. Les résultats expérimentaux montrent que notre méthode permet non seulement de réduire de façon significative la taille de l'échantillon d'apprentissage mais d'améliorer aussi la précision de classification qui s'avère particulièrement avantageuse pour les classes difficiles.

Références

[1] L. Breiman. Bagging predictors. *Machine Learning*, 24(2) :123–140, Août 1996.

[2] L. Breiman. Random forests. *Machine Learning*, 45(1) :5–32, Octobre 2001.

[3] L. Breiman et al. Classification and Regression Trees. *Publisher : Wadsworth*, 1984.

[4] T.G. Dietterich. Ensemble methods in machine learning. *Proceedings of the First International Workshop on Multiple Classifier Systems*, Springer-Verlag, pages 1-15, 2000.

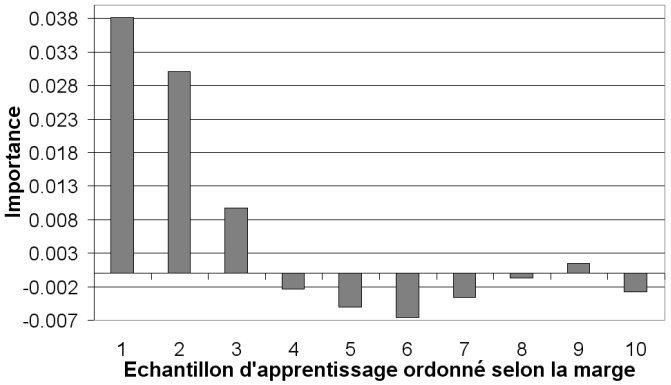
[5] Y. Freund and R. E. Schapire. A Short Introduction to Boosting. *Journal of Japanese Society for Artificial Intelligence*, 14(5) :771–780, 1999.

[6] P.O. Gislason, J.A. Benediktsson and J.R. Sveinsson. Random Forests for land cover classification. *Pattern Recognition Letters*, 27(4) :294-300, 2006.

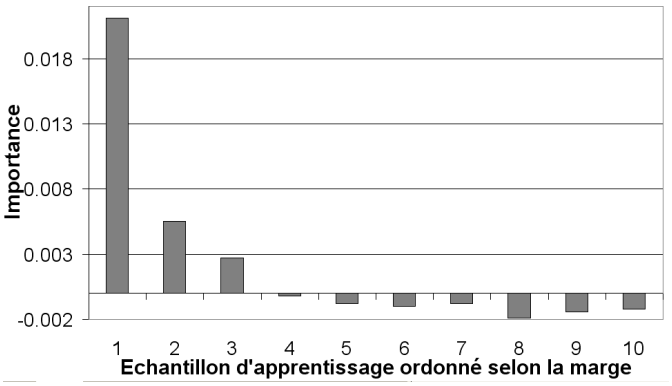
[7] L. Guo, S. Boukir, and N. Chehata. Support vectors selection for supervised learning using an ensemble approach. In *ICPR'2010, 20th IAPR International Conference on Pattern Recognition*, pages 37–40, 2010.

[8] R. E. Schapire, Y. Freund, P. Bartlett, and W.S. Lee. Boosting the margin : a new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5) :1651–1686, 1998.

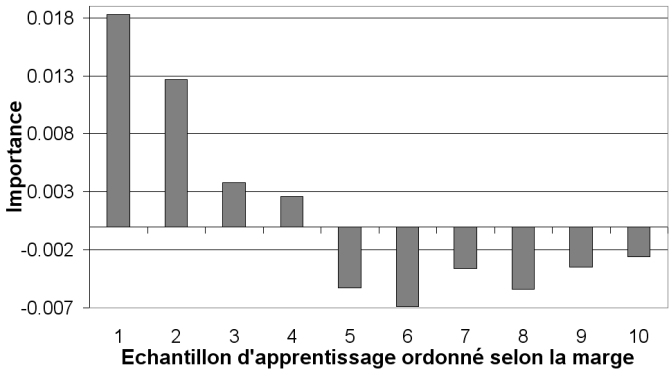
[9] V. Vapnik. *The Nature of Statistical Learning Theory*. Number ISBN 0-387-98780-0. Springer, New York, 1995.



(a) Jeu de données *Letter*.



(b) Jeu de données *Segment*.



(c) Jeu de données $Waveform$.

FIGURE 2 – Importance de sous-échantillon d'apprentissage basée sur la marge