Classifieurs Probabilistes Parcimonieux

Romain HÉRAULT[1], Yves GRANVALET[2]

[1]HEUDIASYC, UMR CNRS 6599
Université de Technologie de Compiègne
BP 20529, 60205 Compiègne cedex, France

[2]IDIAP
Rue du Simplon 4, Case Postale 592
CH-1920 Martigny, Switzerland
romain.herault@hds.utc.fr, yves.grandvalet@idiap.ch

Résumé – Les résultats retournés par les séparateurs à vaste marge sont souvent utilisés comme mesures de confiance pour la classification de nouveaux exemples. Cependant, il n'y a pas de fondement théorique à cette pratique. C'est pourquoi, lorsque l'incertitude de classification doit être estimée, il est plus sûr d'avoir recours à des classifieurs qui estiment les probabilités conditionnelles des classes. Ici, nous nous concentrons sur l'ambiguïté à proximité de la frontière de décision. Nous proposons une adaptation de l'estimation par maximum de vraisemblance, appliquée à la régression logistique. Le critère proposé vise à estimer les probabilités conditionnelles, de manière précise à l'intérieur d'un intervalle défini par l'utilisateur, et moins précise ailleurs. Le modèle est aussi parcimonieux, dans le sens où peu d'exemples contribuent à la solution. L'efficacité du calcul est ainsi améliorée par rapport à la régression logistique. De plus, nos premières expériences montrent une amélioration des performances par rapport à la régression logistique standard, avec des performances similaires à celles des séparateurs à vaste marge.

Abstract – The scores returned by support vector machines are often used as a confidence measures in the classification of new examples. However, there is no theoretical grounds sustaining this practice. Thus, when classification uncertainty has to be assessed, it is safer to resort to classifiers estimating conditional probabilities of class labels. Here, we focus on the ambiguity in the vicinity of the boundary decision. We propose an adaptation of maximum likelihood estimation, instantiated on logistic regression. The model outputs proper conditional probabilities into a user-defined interval and is less precise elsewhere. The model is also sparse, in the sense that few examples contribute to the solution. The computational efficiency is thus improved compared to logistic regression. Furthermore, preliminary experiments show improvements over standard logistic regression with performances similar to support vector machines.

1 Introduction

Lorsqu'il existe une vaste majorité d'exemples négatifs "non intéressants", et seulement peu d'exemples appartenant à la classe positive, l'apprentissage a tendance à biaiser ses résultats en faveur de la classe dominante. Ce biais peut être traité en rééquilibrant la distribution d'apprentissage [10, 2] : les exemples de la classe minoritaire peuvent être répliqués ou générés artificiellement, un certain nombre d'exemples de la classe majoritaire peuvent être éliminés. Cependant, d'une part, l'augmentation du nombre d'exemples de la classe minoritaire donne un calcul inefficace, d'autre part, la réduction de la classe majoritaire peut amener à l'élimination d'informations importantes pour la classification. Ce problème des classes déséquilibrées, notre motivation originale pour ce travail, doit pouvoir tirer profit d'un classifieur parcimonieux qui permet l'estimation précise des probabilités sur un intervalle d'intérêt.

Les séparateurs à vaste marge (SVM) sont les modèles parcimonieux les plus répandus. Plusieurs tentatives ont eu pour but de transformer le score retourné par ces derniers en une estimation de probabilité [8, 3]. Cependant, il n'existe aucune garantie théorique que les scores générés par les SVM représentent une mesure de confiance. De plus, [1] a démontré que les étiquettes des classes ne peuvent être retrouvées sans ambiguïté que sur la frontière de décision. Si la connaissance de ces probabilités est nécessaire, il est donc préférable d'utiliser des classifieurs probabilistes, comme la régression logistique, qui estiment, directement, les probabilités conditionnelles de façon consistante.

Nous nous proposons de construire un classifieur probabiliste qui est précis sur une "zone grise", là où les étiquettes des classes changent. L'obtention de l'incertitude de classification est garantie dans cette zone, où les probabilités sont bien calibrées. Ce classifieur permet aussi d'obtenir des règles de décision appropriées pour l'ensemble du domaine, règles correspondant aux pertes asymétriques de mauvaise classification. Se concentrer sur un petit intervalle de probabilités conditionnelles au lieu d'effectuer une estimation sur l'ensemble du domaine possède deux avantages : premièrement, l'objectif de l'apprentissage se rapproche de la minimisation du risque de mauvaise classification qui est l'objectif final; deuxièmement, [1] a prouvé que, si les probabilités conditionnelles peuvent être estimées partout sans ambiguïté, alors, les modèles à noyau ne peuvent être parcimonieux. L'imprécision des probabilités conditionnelles en dehors de l'intervalle d'intérêt est donc un élément clé de l'efficacité des méthodes à noyau.

Une méthode à noyau est parcimonieuse si un nombre limité d'éléments est pris en compte, par exemple, si un nombre important d'exemples d'apprentissage n'ont pas d'influence sur les procédures d'apprentissage et de test. La parcimonie est source d'efficacité de calcul.

Notre méthode est proche des méthodes adaptant l'objectif

d'apprentissage par l'utilisation de coûts différents pour les exemples positifs et négatifs [6, 11]. Néanmoins, elle diffère de ces derniers qui consistent à l'application de poids différents pour chaque catégorie.

2 Critère d'apprentissage

Nous avons un ensemble d'apprentissage $\mathcal{L} = \{\mathbf{x}_i, y_i\}_{i=1}^n$, où chaque exemple est décrit par les caractéristiques $\mathbf{x}_i$ et l'étiquette associée $y_i \in \{-1, 1\}$. En supposant l'indépendance des exemples, l'estimation de $p(y|\mathbf{x})$ peut être réalisée par la maximisation de la log-vraisemblance conditionnelle :

$$\sum_{i:y_i=1} \log(\widehat{p}(y=1|\mathbf{x}_i)) + \sum_{i:y_i=-1} \log(1-\widehat{p}(y=1|\mathbf{x}_i)) \ , \quad (1)$$

où $\widehat{p}(y|\mathbf{x})$ est l'estimateur de $p(y|\mathbf{x})$.

La règle de décision de Bayes est définie par les vraies probabilités conditionnelles $p(y|\mathbf{x})$, et les coûts de mauvaise classification. Les deux types d'erreurs sont: les faux positifs occasionnant une perte C^-; les faux négatifs occasionnant une perte C^+.

Bien que la règle de décision de Bayes soit définie par $p(y|\mathbf{x})$, elle n'a pas besoin d'un estimateur précis sur l'ensemble du domaine des probabilités: il suffit d'estimer les probabilités conditionnelles en $\frac{C^-}{C^+ + C^-}$, qui définit la frontière de décision . Asymptotiquement, c'est ce que réalisent les SVM [1] pour $p(y|\mathbf{x}) = 0.5$, ou pour d'autres probabilités lorsque le critère est asymétrique [6, 11].

Notre approche vise à donner une estimation des probabilités conditionnelles sur un intervalle $[p_{\min}, p_{\max}]$. Au-delà de cet intervalle, nous voulons juste savoir si $p(y|\mathbf{x})$ est plus petit que $p_{\min}$ ou plus grand que $p_{\max}$. Ceci peut-être formalisé par:

$$\sum_{i:y_i=1} \log\left(\min\left(\widehat{p}(y=1|\mathbf{x}_i), p_{\max}\right)\right) + \sum_{i:y_i=-1} \log\left(\min\left(1-\widehat{p}(y=1|\mathbf{x}_i), 1-p_{\min}\right)\right) \quad (2)$$

qui est un critère concave en $\widehat{p}(y=1|\mathbf{x}_i)$. Il est calibré pour la classification si $\frac{C^-}{C^+ + C^-} \in [p_{\min}, p_{\max}]$ [1].

3 Application à la régression logistique

La régression logistique est un modèle probabiliste classique qui considère que le log-ratio des probabilités conditionnelles est linéaire: $\log \frac{\widehat{p}(y=1|\mathbf{x})}{1-\widehat{p}(y=1|\mathbf{x})} = \mathbf{w}^T\mathbf{x} + b$, où $(\mathbf{w}, b)$ sont estimés en maximisant la vraisemblance (éq. 1) ou la vraisemblance pénalisée. Nous pouvons utiliser la notion de noyau dans la régression logistique en modélisant le log-ratio par $f(\mathbf{x}_i) + b$ où f appartient à un espace de Hilbert $\mathcal{H}$. Nous devons alors introduire dans le critère d'apprentissage une pénalisation afin d'éviter le sur-apprentissage [9, 13]. Maximiser la vraisemblance (1) pénalisée par la norme de f revient à minimiser:

$$\sum_{i=1}^n \log\left(1 + e^{-y_i(f(\mathbf{x}_i)+b)}\right) + \frac{\lambda}{2}||f||^2_{\mathcal{H}} \ , \quad (3)$$

Où λ est un hyper-paramètre ajusté par validation croisée. Contrairement aux SVMs, la régression logistique n'est pas parcimonieuse car tous les exemples influent sur la solution. Notre méthode consiste à appliquer le critère (2) au lieu de la log-vraisemblance. La régression logistique parcimonieuse à noyau minimise:

$$\sum_{i=1}^n \log\left(1 + e^{\max(-y_i(f(\mathbf{x}_i)+b), F_i)}\right) + \frac{\lambda}{2}||f||^2_{\mathcal{H}} \ , \quad (4)$$

Où $F_i = -\log\frac{p_{\max}}{1-p_{\max}}$ si $y_i = 1$ et $F_i = \log\frac{p_{\min}}{1-p_{\min}}$ si $y_i = -1$. La parcimonie est la conséquence de la saturation de la perte. Dans le critère d'apprentissage, c'est le terme d'ajustement, au lieu du terme de pénalisation, qui cause la parcimonie. Les exemples d'apprentissage avec une grande valeur $y_i f(\mathbf{x}_i)$ ne participeront pas au classifieur final.

La régression logistique à noyau peut être entrainée dans le domaine primal en utilisant la méthode de Newton [9], ou dans le domaine dual [4]. Pour notre approche, nous proposons un algorithme de contraintes actives, suivant en cela l'algorithme SimpleSVM [12, 5] qui a prouvé être efficace pour les SVMs [5]. Les exemples de l'ensemble d'apprentissage sont scindés en exemples informatifs (exemples ayant une valeur $y_i f(\mathbf{x}_i)$ petite) et en exemples non-informatifs. En suite, le critère d'apprentissage est optimisé considérant cette partition fixe et le résultat de cette optimisation permet une nouvelle partition entre exemples informatifs et non-informatifs. Ces deux étapes sont répétées jusqu'à ce que la précision souhaitée soit atteinte [5].

4 Expérimentations

Pour étudier le problème de deux classes déséquilibrées, nous avons choisi la base de donnée Forest, la plus grande base de données de l'UCI. [1] Les exemples sont décrits par 54 caractéristiques, 10 sont quantitatives et 44 sont binaires. Originellement, il y a 7 classes , mais nous considérons la discrimination de la classe positive *Krummholz* (20 510 exemples) de la classe négative classe *Épicéa/Sapin* (211 840 exemples). La proportion de la classe positive est de 8.8%, et les classes sont relativement bien séparées. Comme il n'y a pas de matrice de coût liée à ces données, nous avons arbitrairement choisit les coûts pour les faux positifs et les faux négatifs, C^+ et C^-, de façon à encourager un taux d'erreur équivalent pour les deux catégories, c'est à dire $\frac{C^-}{C^+ + C^-} = \pi^+$, où $\pi^+ = 8.8\%$ est la proportion d'exemples positifs. Les pertes sont alors définies à un facteur près, et nous choisissons $C^- = \pi^+$ et $C^+ = 1 - \pi^+$.

4.1 Cadre d'expérimentation

Pour assurer la représentativité des résultats, les données sont réparties en 10 sous-ensembles. Chaque sous-ensemble est itérativement utilisé an tant qu'ensemble d'apprentissage alors que les sous-ensembles restants sont utilisés comme ensembles de test. Ainsi, les ensembles d'apprentissage comprennent 23 235 exemples. La proportion d'exemples positifs (minoritaires) est identique pour tous les sous-ensembles. Les carac-

[1] Disponible sur `kdd.ics.uci.edu/databases/covertype`.

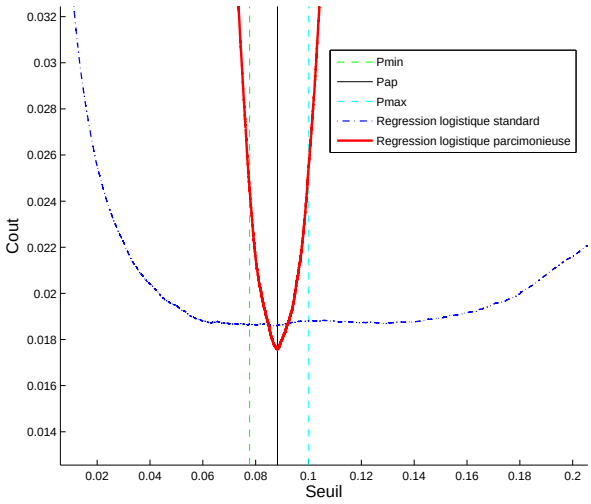

Figure 1: Coût de test en fonction du seuil de décision.

téristiques sont normalisées (centrées et réduites) avant chaque session d'apprentissage.

Les expériences rapportées ici ont été effectuées par des classifieurs linéaires. Nous avons optimisé le paramètre de pénalisation λ pour la régression logistique (3) et la régression logistique parcimonieuse (4) par une validation croisée sur 5 blocs. Nous avons optimisé conjointement le seuil de décision, cette procédure est souvent appliquée aux classifieurs dans le but de corriger le biais des probabilités estimées. La correction du biais doit favoriser la régression logistique.

L'intervalle $[p_{\min}, p_{\max}]$ des probabilités conditionnelles, qui est supposé être défini par l'utilisateur, n'est pas optimisé. De meilleurs résultats d'optimisation sont attendus pour de petits intervalles, mais l'intervalle des probabilités conditionnelles fiables se réduit alors. Nous rapportons les résultats pour diverses longueurs d'intervalles centrés sur π^+ sur l'échelle logarithmique, c'est à dire $\sqrt{p_{\min} p_{\max}} = \pi^+$.

4.2 Résultats

Nous rapportons les performances moyennes de la régression logistique parcimonieuse, ainsi que leur écart-type dans la table 1. Comme attendu, la moyenne du coût de test décroît lorsque l'intervalle $[p_{\min}, p_{\max}]$ décroît, $p_{\max} - p_{\min} = 1$ représente la régression logistique standard. Nous montrons aussi le seuil de décision moyen sur les probabilités conditionnelles estimées, seuil estimé par validation croisée. Ce dernier est juste au dessus de $\pi^+ = 8.8\%$ pour la régression logistique standard, mais la différence n'est peut-être pas significative (les tests d'hypothèses usuels ne peuvent être appliqués car les expériences ne sont pas indépendantes). Le seuil de décision correct est toujours choisi pour la régression logistique parcimonieuse sur de petit intervalles $[p_{\min}, p_{\max}]$. Les classifieurs sont donc bien calibrés en décision. La proportion d'exemples de l'ensemble actif (noté SV par identification aux vecteurs supports) est rapportée sur la dernière ligne. Cette proportion diminue aussi lorsque l'intervalle $[p_{\min}, p_{\max}]$ décroît.

La figure 1 compare, pour un essai, la sensibilité au seuil de détection du coût de test moyen des régressions logistiques standard et parcimonieuse (avec $p_{\max} - p_{\min} = 2.2\%$). Les figures obtenues pour les autres essais sont similaires. A savoir, la régression logistique possède un minimum plat et large, reflétant le fait que la proportion de décisions correctes ne change pas beaucoup autour du seuil de décision. Cela veut dire que les vraies probabilités conditionnelles fluctuent de façon non-monotone dans cette région. La régression logistique parcimonieuse se comporte beaucoup mieux avec un minimum plus étroit et plus bas centré sur π^+, reflétant des probabilités conditionnelles bien calibrées dans la région ciblée.

La table 2 résume les résultats obtenus avec les séparateurs à vaste marge. Les résultats des SVM standards sont mauvais parce que le coût optimisé, avec $C^+ = C^-$, n'est pas le bon. Ceci peut être compensé en déplaçant le seuil de décision. Les performances correspondantes sont montrées dans la colonne "Avec correction du biais". Cependant, un meilleur choix consiste à modifier la perte *hinge* pour faire en sorte que $C^+ \neq C^-$ [7]. Le résultat correspondant, donné dans la colonne C^+/C^-, atteint les performances de la régression logistique parcimonieuse sur des petits intervalles d'intérêt. Le nombre de vecteurs supports (notés SV) pour le cas C^+/C^- et le nombre d'exemples dans l'ensemble actifs pour les petits intervalles $[p_{\min}, p_{\max}]$ de la régression logistique parcimonieuse sont du même ordre de grandeur.

5 Discussion

Nous avons proposé un nouveau critère d'apprentissage, qui consiste à tronquer la log-vraisemblance binomiale. Ce critère produit un classifieur probabiliste parcimonieux, qui fournit des probabilités conditionnelles fiables au voisinage de la frontière de décision. Nous avons examiné en détail comment la régression logistique est modifiée par la maximisation de "la vraisemblance locale", mais ce principe peut être appliqué sur d'autres modèles de probabilités conditionnelles comme les réseaux de neurones. Bien que nous ayons uniquement discuté du problème de classification binaire, le principe est par essence multi-classe et peut être appliqué à une log-vraisemblance multinomiale. Le problème d'optimisation résultant reste un problème convexe pourvu que l'intervalle "intéressant" des probabilités conditionnelles soit un ensemble convexe.

Des expériences sont en cours pour confirmer l'intérêt pratique des classifieurs probabilistes parcimonieux, mais ils offrent déjà des résultats prometteurs pour le problème des classes déséquilibrées. Le critère d'apprentissage ignore les exemples bien classés hors de la "zone grise", définie par un intervalle sur les probabilités conditionnelles. Les exemples actifs sont des exemples ambigus et mal classés, ce qui permet d'ignorer bon nombre d'exemples de la classe majoritaire. Il y a donc un sous-échantillonage virtuel et ciblé de la classe majoritaire.

Nos premières expériences sur des classifieurs linéaires montrent que les classifieurs probabilistes tirent profit de la concentration du critère sur la "zone grise" près de la frontière de décision. La régression logistique parcimonieuse fournit de meilleures règles de décision que la régression logistique standard. Non seulement nous gagnons en erreur de test mais aussi elle est plus rapide à entraîner, grâce à sa capacité d'ignorer

Table 1: Coûts moyens de test pour les régression logistiques parcimonieuse et standard ($p_{\max} - p_{\min} = 100\%$)

$p_{\min}$ (%)	0	0.4	1.0	2.9	4.8	7.8
$p_{\max}$ (%)	100	72.0	47.5	24.1	15.8	10.0
$p_{\max} - p_{\min}$ (%)	100	71.6	46.4	21.2	11.0	2.2
Coût moyen de test ($\times 10^{-2}$)	1.86	1.86	1.85	1.85	1.83	1.78
Écart-type	± 0.01	± 0.01	± 0.01	± 0.01	± 0.02	± 0.02
Seuil moyen de décision (%)	9.5	9.0	9.0	9.0	8.8	8.8
Écart-type	± 1.3	± 1.1	± 0.9	± 0.6	± 0.2	± 0.0
Prop. moyenne de SV (%)	100	65.5	53.5	40.5	34.0	27.9

Table 2: Coûts moyens de test obtenus pour les SVMs

SVM	Standard	avec correction du biais	C^+/C^-
Coût moyen de test ($\times 10^{-2}$)	3.75 ± 0.23	2.31 ± 0.12	1.79 ± 0.02
Prop. moyenne de SV (%)	12.84 ± 0.79	13.16 ± 1.13	26.19 ± 0.60

les données non pertinentes. Les performances et les temps d'apprentissage sont comparables aux SVM entrainés avec des coûts asymétriques C^+/C^-, et nous pouvons profiter en plus de probabilités bien calibrées dans le voisinage de la frontière de décision.

References

[1] P. L. Bartlett and A. Tewari. Sparseness vs estimating conditional probabilities: Some asymptotic results. In *Proceedings of the 17th Annual Conference on Learning Theory*, volume 3120 of *Lecture Notes in Computer Science*, pages 564–578. Springer, 2004.

[2] Charles Elkan. The foundations of cost-sensitive learning. In *IJCAI*, pages 973–978, 2001.

[3] Y. Grandvalet, J. Mariéthoz, and S. Bengio. A probabilistic interpretation of SVMs with an application to unbalanced classification. In Y. Weiss, B. Schölkopf, and J. C. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 467–474. MIT Press, 2006.

[4] S. S. Keerthi, K. B. Duan, S. K. Shevade, and A. N. Poo. A fast dual algorithm for kernel logistic regression. *Mach. Learn.*, 61(1-3):151–165, 2005.

[5] G. Loosli and S. Canu. Comments on the "core vector machines: Fast SVM training on very large data sets". *Journal of Machine Learning Research*, to appear.

[6] K. Morik, P. Brockhausen, and T. Joachims. Combining statistical learning with a knowledge-based approach - a case study in intensive care monitoring. In *Proceedings of the Sixteenth International Conference on Machine Learning (ICML-99)*, 1999.

[7] E. Osuna, R. Freund, and F. Girosi. Support vector machines: Training and applications. Technical Report A.I. Memo No. 1602, M.I.T. AI Laboratory, 1997.

[8] J. C. Platt. Probabilities for SV machines. In A. J. Smola, P. L. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 2000.

[9] Volker Roth. Probabilistic discriminative kernel classifiers for multi-class problems. In *Proceedings of the 23rd DAGM-Symposium on Pattern Recognition*, pages 246–253, London, UK, 2001. Springer-Verlag.

[10] Kai Ming Ting. A comparative study of cost-sensitive boosting algorithms. In *Proc. 17th International Conf. on Machine Learning*, pages 983–990. Morgan Kaufmann, 2000.

[11] K. Veropoulos, C. Campbell, and N. Cristianini. Controlling the sensitivity of support vector machines. In T. Dean, editor, *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 55–60. Morgan Kaufmann, 1999.

[12] S. V. N. Vishwanathan, Alex J. Smola, and M. Narasimha Murty. SimpleSVM. In *Proceedings of the Twentieth International Conference on Machine Learning*, pages 760–767, 2003.

[13] J. Zhu and T. Hastie. Kernel logistic regression and the import vector machine. In *Advances in Neural Information Processing Systems 13*, 2001.