La discrimination à plus de deux classes Comparaison de plusieurs approches issues des *Support Vector Machines* (SVM)

Emmanuel Jakobowicz

Ecole Polytechnique Fédérale de Lausanne
Institut de mathématiques
CH-1015 Lausanne (Suisse)
emmanuel.jakobowicz@epfl.ch

RÉSUMÉ. *L'apprentissage automatique est aujourd'hui en pleine expansion, suite aux découvertes de V. Vapnik, la méthode de discrimination issue de ces théories obtient des résultats extrêmement bons. Les Support Vector Machines (SVM) ont été très bien développés dans le cas de deux classes, mais pour le cas de plus de deux classes plusieurs méthodes existent mais elles n'ont pas été clairement comparées. Dans ce travail, nous avons rassemblé les principales méthodes existantes et les avons comparées sur un certain nombre de jeux de données tests. D'autre part, comme le pourcentage de bien classées n'est bien souvent pas suffisamment sensible pour juger de la qualité d'un modèle, nous avons étudié de nouveaux critères de validation.*

MOTS-CLÉS : *Support Vector Machines (SVM), multi-classes, critère de validation, apprentissage automatique*

1. Introduction

Les méthodes de Support Vector Machines ou Séparateurs à Vaste Marges (SVM) font parties aujourd'hui des méthodes de discrimination les plus efficaces. Ce sont des méthodes d'apprentissage automatique, c'est-à-dire à partir d'un échantillon d'apprentissage, on va créer un modèle applicable sur des échantillons à tester. Elles ne sont par contre pas adaptées au cas d'une discrimination à plus de deux classes. Dans ce travail, nous avons cherché à rassembler les techniques utilisées pour passer à plus de deux classes.
Tout d'abord nous présenterons les SVM bi-classes. Suivra la description des différentes méthodes à plus de deux classes. J'introduirai ensuite rapidement quelques nouveaux critères de validation des modèles. Cet article se terminera sur des remarques et conclusions issues d'un certain nombre d'applications de ces différentes techniques.

2. Les Séparateurs à vaste marges (SVM)

Les SVM sont des méthodes issues de la théorie de minimisation structurelle du risque développée par V. Vapnik [VAPNIK 98]. Leur but est de séparer linéairement les observations de l'échantillon d'apprentissage en les projetant dans un espace de très grande dimension. La séparation se fera par une maximisation de la marge. Pour des observations $\mathbf{x}_i$ de la classe $y_i \in \{-1, 1\}$, il y aura deux étapes :

1. Projection des observations dans un espace de Hilbert à noyau reproduisant (de très grande dimension)
2. Séparation linéaire des données ainsi transformée en tentant de maximiser la marge.

Ceci revient à résoudre le problème :

$$\begin{array}{ll} & \min_{\mathbf{w}} \frac{1}{2} \parallel \mathbf{w} \parallel^2 \\ s.c. & y_i(\mathbf{w}\,\Phi(\mathbf{x}_i) + b) \geq 1 \quad \forall i \end{array}$$

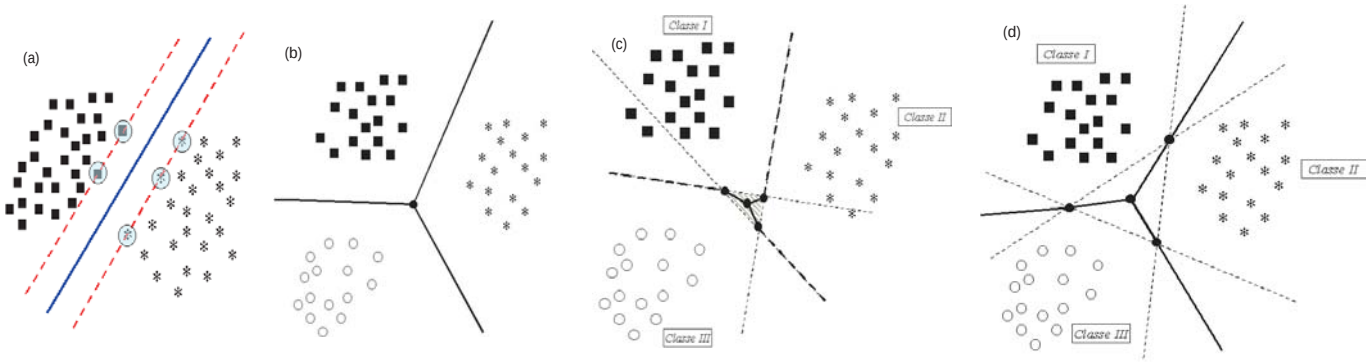

FIG. 1. *(a) Séparation par maximisation de la marge ; Méthodes M-SVM : (b) Weston & Watkins et Crammer & Singer ; (c) "un contre un" ; (d) "un contre le reste"*

La solution de son dual ne dépendant que du produit scalaire $\Phi(\mathbf{x}_i)\Phi(\mathbf{x}_j)$, nous utiliserons des fonctions appelées noyaux notées $k(\mathbf{x}_i, \mathbf{x}_j)$ qui devront satisfaire quelques propriétés (propriété de Mercer) et qui permettront de ne pas avoir à calculer de produit scalaire dans un espace de très grande dimension.

Le cas non séparable :
Bien souvent, malgré la projection, les données restent non séparables, on devra donc modifier le problème primal qui prendra la forme suivante :

$$\begin{array}{ll} & \min_{\mathbf{w},\xi_i,C} \|\mathbf{w}\|^2 + C \sum \xi_i \\ s.c. & y_i(\mathbf{w}\,\Phi(\mathbf{x}_i) + b) \geq 1 - \xi_i \quad \forall i \end{array}$$

On a ajouté des variables d'écart et un paramètre C qu'il faudra régler afin d'équilibrer d'un côté la maximisation de la marge et de l'autre le nombre d'observations que l'on accepte de mal classer.
Ces méthodes sont adaptées pour deux classes mais leur généralisation à plus de 2 classes n'est pas directe.

3. Les SVM à plus de deux classes

3.1. *Méthodes issues directement des SVM bi-classes [*FRIEDMAN *96]*

Méthode "un contre un" : On va tester toutes les paires de classes. On aura donc $\frac{k(k-1)}{2}$ problèmes du type SVM. [FRIEDMAN 96]
Pour appliquer cette méthode, pour chaque observation, on crée une nouvelle observation à k catégories, si entre les classes i et j, l'observation $\mathbf{x}$ choisit i, alors on augmente de 1 la i^{eme} catégorie de la nouvelle observation.
On choisit pour l'observation $\mathbf{x}$ la classe j correspondant à la catégorie avec le plus de vote. En cas d'égalité, on choisit le plus petit index de classe.

Méthode "un contre le reste" : On va tester chaque classe contre les $k-1$ autres classes. On choisira la classe obtenant les meilleurs résultats.

3.2. *Méthodes multi-classes*

Plusieurs méthodes ont été mises au point afin de passer à plus de deux classes en traitant simultanément toutes les classes. Les méthodes de Weston & Watkins et de Crammer & Singer seront développées.

Méthode de Weston & Watkins[WESTON 98] : Elle traite toutes les classes simultanément en résolvant le problème suivant :

$$\begin{array}{ll} \min_{\mathbf{w}} \frac{1}{2}\sum_{j=1}^{k} \|\mathbf{w}_j\|^2 + C\sum_{i=1}^{n}\sum_{j=1}^{k}\xi_i^j \\ s.c. & \begin{cases} (\mathbf{w}_{C(\mathbf{x}_i)} - \mathbf{w}_j)^T\mathbf{x}_i + b_{C(\mathbf{x}_i)} \geq 1 - \xi_i^j \\ \xi_i^j \geq 0, \quad (1 \leq i \leq n),\ (1 \leq j \neq C(\mathbf{x}_i) \leq k) \end{cases} \end{array}$$

Méthode de Crammer & Singer[CRAMMER 01] : Elle est directement issue de la méthode précédente, quelques simplifications ont été effectuées. En effet, on a retiré le biais et les variables d'écart ont été rassemblées de manière à obtenir un dual plus simple.

4. Les critères de validation

L'utilisation d'un modèle de discrimination est motivé par la qualité de celui-ci. Or, cette qualité est estimée par des critères de validation. J'ai donc consacré une partie de mon travail à l'étude de critères qui seraient plus appropriés que le simple pourcentage de bien classés qui n'est plus suffisant.

Le Ki de KXEN : Cette notion développée par l'entreprise $KXEN$ et présentée dans [MARKADE 03], obtient de bons résultats dans le cas binaire.

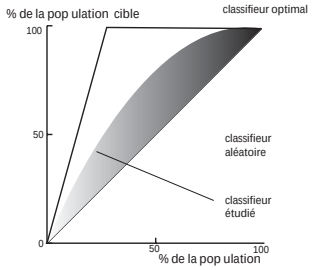

FIG. 2. *Ki : rapport de l'aire de la surface grise et de celle de la surface entre le classifieur aléatoire et le classifieur optimal*

Ce critère a l'avantage de valoir 1 pour un modèle parfait et 0 pour un modèle aléatoire. De plus, il permet de souligner des erreurs même dans le cas d'une distribution des classes très inégale. Malheureusement, le Ki n'est pas directement extensible à plus de deux classes, jusqu'alors, on a donc utilisé la moyenne ou le minimum des Ki pour chaque paire de classes. On aura donc $\frac{k(k-1)}{2}$ Ki à calculer.

Le κ de Cohen : C'est une mesure non paramétrique d'accord entre deux variables qualitatives pour des données appariées. Elle mesure l'écart à la diagonale dans la matrice de confusion. On peut l'écrire :

$$\kappa = \frac{N \sum_{i=1}^{k} z_{ii} - \sum_{i=1}^{k} z_{i.}\, z_{.i}}{N^2 - \sum_{i=1}^{k} z_{i.}\, z_{.i}}$$

où (z_{ij}) est la matrice de confusion associée au classifieur et N le nombre total d'observations.

5. Résultats et comparaisons

J'ai modifié et utilisé des logiciels mis au point par C.J. Lin, C.C. Chang et C.W. Hsu pour tester ces méthodes. [HSU]
J'ai procédé par validation croisée à 10 parts et calculé le pourcentage de bien classés ainsi que la moyenne et le minimum des Ki et le κ. J'ai d'autre part comparé ces méthodes à des discriminations par k plus proches voisins (k-NN) et par arbre de décision (ID3). J'ai utilisé un noyau radial car, par l'expérience, il obtient les meilleurs résultats et des temps de travail très courts. La recherche des paramètres optimaux (le C et l'inverse de la variance dans le noyau) a été faite par un test de chaque combinaison de paramètres possible dans des intervalles choisis. D'autre part, nous n'avons étudié que des jeux de données avec un faible nombre de classes (entre 3 et 5).
Les différentes applications donnent des résultats assez proches, on le voit dans le tableau suivant où apparaissent les κ pour chaque méthode. On obtient un léger avantage pour la méthode "un contre un" sur les trois autres. Par

contre, la comparaison faite avec les deux méthodes non SVM donne un avantage significatif aux méthodes de SVM multi-classes. D'autre part, le temps de travail est aussi une notion essentielle dans ces méthodes. En effet, ces méthodes doivent être rapides car lors de l'optimisation des paramètres, il faut appliquer un grand nombre de fois chaque méthode. Il ressort de même que la méthode "un contre un" est la plus rapide, ceci peut s'expliquer par le fait que nous n'avons traité que des cas avec peu de classes, ce qui fait peu de problèmes bi-classes à résoudre pour cette méthode. D'autre part, lorsque le C augmente, il arrive que la méthode de Crammer & Singer obtienne des temps de convergence extrêmement longs. Ceci peut être expliqué par le fait qu'on a utilisé un algorithme de décomposition du dual qui peut poser des problèmes de convergence.
Les nouveaux critères de validation des modèles ressortent comme plus sensibles que le pourcentage de bien classés. La moyenne des Ki et le κ semblent spécialement adaptés. Le fait qu'ils soient reliés à l'aire sous la courbe ROC (receiver operating characteristic) leur donne des propriétés que le pourcentage de bien classés ne possède pas.

Données	"un contre un"	"un contre le reste"	W. & W.	C. & S.	ID3	k-NN	# classes	# obs.	# var.
dna	**0.93**	0.92	0.89	0.92	0.88	–	3	2000	180
car	**0.97**	0.81	0.97	0.94	0.91	0.91	4	1728	6
wine	0.73	**0.89**	0.66	0.86	0.53	0.74	3	178	13
iris	0.89	**0.94**	0.89	0.88	0.84	0.83	3	120	4

TAB. 1. *κ pour un certain nombre de jeux de données*

6. Conclusions et ouvertures

Toutes ces observations amènent à un certain nombre de remarques. Tout d'abord, les méthodes SVM à plus de deux classes obtiennent de très bons résultats, elles surpassent les méthodes de discrimination classique aussi bien au niveau des résultats qu'au niveau du temps d'exécution. De plus, elles sont très simple d'utilisation et un processus automatisé peut assez facilement être mis en place. La méthodes qui, dans le cas d'un nombre de classes faible, semble la plus adaptée du fait de son efficacité mais aussi de sa simplicité est la méthode "un contre un". D'autre part, les critères de validation étudiés permettent d'éviter des erreurs et il serait intéressant dans bien des cas d'associer au pourcentage de bien classés soit la moyenne des Ki, soit le κ de Cohen.

Bien sûr, des approfondissements seraient possibles, ainsi on pourrait modifier la méthode de vote dans la méthode "un contre un", améliorer les décompositions proposées pour les méthodes traitant toutes les classes simultanément ou tester des jeux de données avec plus de classes (par exemple 26 pour la reconnaissance de caractères).

7. Bibliographie

[CRAMMER 01] CRAMMER K., SINGER Y., On the algorithmic implementation of multiclass kernel-based vector machines, *School of Computer Science and Engineering, Hebrew University*, , 2001.

[FRIEDMAN 96] FRIEDMAN J., Another Approach to Polychotomous Classification, *Stanford University*, , 1996.

[HSU] HSU C., LIN C., CHANG C., Logiciels LIBSVM et BSVM, `URL : www.csie.ntu.edu.tw/~cjlin/`.

[MARKADE 03] MARKADE E., Evaluating Modeling Techniques, KXEN Inc. (Knowledge Extraction Engines), 2003.

[VAPNIK 98] VAPNIK V., *Statistical learning theory*, John Wiley & Sons, Inc., N.Y., 1998.

[WESTON 98] WESTON J., WATKINS T., Multi-class Support Vector Machines, *Technical Report - Royal Holloway University of London*, , 1998.