
La classification des textes d'opinion par les Séparateurs à Vaste Marge (SVM) avec sorties probabilistes

Anh-Phuc TRINH

Laboratoire d'Informatique de Paris 6
104, avenue du Président Kennedy
75016, Paris
anh-phuc.trinh@lip6.fr

RÉSUMÉ. Nous avons étudié le problème de classification multi-classes par les Séparateurs à Vaste Marge (SVM) avec sorties probabilistes. La stratégie «un-contre-un» décompose ce problème en plusieurs problèmes locaux dans lesquels une fonction de décision nous donne une information discriminante. Il s'agit de combiner ces informations discriminantes afin de construire une sortie probabiliste du problème. Pour réaliser cette tâche, on emploie modèle exponentiel. Nous avons également utilisé cette méthode afin de classifier des textes suivant leurs opinions à partir du corpus d'apprentissage fourni lors de DEFT07.

ABSTRACT. We review the problem of multi-class classification by the SVMs, a strategy one-against-one decompose this problem into a series of the binary classifications, therefore we need to combine these discriminate information to construct its probabilistic output. A combining method based on an exponential model was proposed. Our experiences realize on a learning corpus of DEFT07, a recent challenge that our team have participated.

MOTS-CLÉS : Sortie probabiliste, Séparateurs à Vaste Marge, SVM, Modèle exponentiel, Détection d'opinions.

KEYWORDS: Probabilistic Output, SVM, Exponential Model, Sentiment Analysis, Opinion Detection .

1. Introduction

Ces derniers temps, la classification multi-classes porte une attention particulière sur l'utilisation efficace de Séparateurs à Vaste Marge (SVM). N'étant pas un modèle probabiliste, SVM détermine la classification selon la méthode « winner-take-all » [KER 90], ainsi on constate une perte d'information concernant la *corrélation* entre les classes dans le processus de classification. Après l'étude des articles tels que [PLA 00][HER 07], au sujet de la classification binaire, ou les articles [KER 90] [HAS 98] [Wu 04], au sujet de la classification multi-classes, nous avons constaté des améliorations possibles pour ces travaux. Le modèle exponentiel [WAL 04] a été utilisé pour construire une sortie probabiliste de la classification multi-classes.

Les améliorations sont réalisées sur les données réelles provenant du Défi Fouille de Textes 2007 auquel nous avons participé. Le sujet du défi est la détermination de l'opinion publique à travers des avis d'utilisateurs sur des produits, des jeux et des articles envoyés à des conférences. Nous avons représenté les textes sous la forme de sac de mots ayant chacun un poids Tf-Idf [SAL 88]. En outre, nous avons consultés les méthodes de modélisation des opinions proposées par les équipes participant au défi [WIL 05][RIL 06][CRE 07].

Nous présentons cet article de la manière suivante : dans la partie 2 nous décrivons le problème de classification multi-classes par les Séparateurs à Vaste Marge (SVM) et notre idée d'amélioration, ensuite dans la partie 3, nous décrivons les expériences menées sur le corpus d'apprentissage DEFT07, et nous comparons notre méthode avec les anciennes méthodes [KER 90][Wu 04] en utilisant les différents codages textuels [SAL 88][CRE 07].

2. Classifieur

Dans cette section, nous décrivons le problème de classification multi-classes par les SVM dans lequel sa sortie est une présentation probabiliste.

Définition 1 : (La probabilité a posteriori de la classification multi-classes)

Soit $E = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_m, y_m)\}$ est un ensemble d'apprentissage de m exemples. Suppose que chaque exemple $\mathbf{x}_i \in \mathcal{R}^n$ et l'étiquette associée y_i est un entier de l'ensemble $Y = \{1, 2, 3, \dots, k\}$. La probabilité a posteriori de classification multi-classes ($k > 2$) est une probabilité conditionnelle, sachant l'exemple \mathbf{x} , de multi-classes y

$$P(y/x) = p_i \text{ avec } \sum_{i=1}^k p_i = 1 \quad (1)$$

Notre travail est de construire cette probabilité conditionnelle à partir du modèle discriminant SVM. Le nombre de classes, noté par k , supérieur à deux nous pose un nouveau problème. En principe les SVMs sont des classificateur binaires, celui ne s'adapte pas à la tâche de classification multi-classes. Donc nous pouvons appliquer deux stratégies dans ce cas, soit «un-contre-un», soit «un-contre-les-autres». Selon les deux stratégies, notre problème s'est décomposé en plusieurs problèmes locaux de la classification binaire. La stratégie «un-contre-un» a été choisie [HSU 02].

Définition 2 : (Le problème local de la classification multi-classes)

Soit $E_{i,j} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_s, y_s)\}$ est un sous-ensemble de E , correspondant à deux classes différents $i \neq j$, c'est-à-dire, chaque y_1 prend une des ses deux valeurs $\{i, j\}$. La classifieur binaire à apprendre sur l'ensemble de $E_{i,j}$, nous notons $cl_{ij}(\mathbf{x})$.

$$cl_{ij}(\mathbf{x}) = \begin{cases} i & \text{si } f_{ij}(\mathbf{x}) = -1 \\ j & \text{si } f_{ij}(\mathbf{x}) = +1 \end{cases} \quad (2)$$

Où i et j sont les indices des classes et $f_{ij}(x)$ est la fonction de décision du SVM.

2.1. En cas de la classification binaire

[PLA 00] nous propose l'utilisation de la fonction de sigmoïde afin de créer la sortie probabiliste

$$P(y = i / y = i \text{ ou } j, \mathbf{x}) = r_{ij} = \frac{1}{(1 + \exp(A \times f_{ij}(\mathbf{x}) + B))} \quad (3)$$

Où A et B sont estimés en maximisant la log-vraisemblance conditionnelle sur l'ensemble d'apprentissage $E_{i,j}$

2.2. En cas de la classification multi-classes

L'idée la plus simple [KER 90] pour construire la probabilité a posteriori (1) à partir des classifieurs binaires locaux $cl_{ij}(\mathbf{x})$ est que nous utilisons la règle de vote. La probabilité $P(y=i|\mathbf{x})$ est égal au nombre de vote favorite pour la classe i divisé par le nombre total de vote. Soit la fonction de vote $V\{x\} = 1$ si x est correct, et zéro autrement.

$$p_i = \frac{2}{k(k-1)} \sum_{j:j \neq i} V(cl_{ij}(\mathbf{x}) = i), \quad i = 1, 2, \dots, k \quad (4)$$

[HAS 98] ont proposé une idée à minimiser la distance de Kullback-Leiber (KL) entre les probabilités locales r_{ij} et $\mu_{ij} = p_i/(p_i+p_j)$ donc la distance de KL est come suite

$$\min_{\mathbf{p}} KL(r_{ij} \parallel \mu_{ij}) = \sum_{i \neq j} |E_{i,j}| \left(r_{ij} \log \frac{r_{ij}}{\mu_{ij}} + (1 - r_{ij}) \log \frac{1 - r_{ij}}{1 - \mu_{ij}} \right) \quad (5)$$

$$s.c. \sum_{j:j \neq i} |E_{i,j}| \mu_{ij} = \sum_{j:j \neq i} |E_{i,j}| r_{ij}, \quad \sum_{i=1}^k p_i = 1, p_i > 0, i = 1, 2, \dots, k \quad (6)$$

[WU 04] ont proposé pour calculer la probabilité a posteriori \mathbf{p} , ils ont minimisé l'écart au carré vient de l'égalité $r_{ij} p_j = r_{ji} p_i$ en introduisant de multiplicateurs de Lagrange à la contraintes d'égalité (8).

$$\min_{\mathbf{p}} \frac{1}{2} \sum_{i=1}^k \sum_{j:j \neq i} (r_{ji} p_i - r_{ij} p_j)^2 \quad (7)$$

$$s.c. \sum_{i=1}^k p_i = 1, p_i \geq 0, i = 1, 2, 3, \dots, k. \quad (8)$$

La minimisation de la distance de Kullback-Leiber (KL) dans la formule (5) nous donne une nouvelle idée à créer un modèle probabiliste à apprendre. Celui-ci prendre des informations discriminantes provenant de l'ensemble de classifieurs SVM comme des données d'entrée, en maximisant l'entropie $H(y/x)$. Le modèle exponentiel [WAL 04] a été choisi.

3. Expériences

Pour évaluer notre méthode, nous l'avons comparer aux deux autres modèles présentés précédemment, celui basé sur une règle de vote (4) et celui proposé par Wu et al. (7).

Pour ce faire, nous avons réutilisé les corpus d'apprentissage provenant du Défi Fouille de Textes 2007¹ auquel nous avons participé. Le thème concernait la détection automatique d'opinions dans des textes présentant des avis d'utilisateurs sur des produits, des jeux et des articles envoyés à des conférences. Cette tâche peut être considérée comme un problème de classification multi-classes dans lequel chaque classe correspond à une valeur d'opinion (négative, neutre, positive).

¹ <http://defi07.limsi.fr/>

Une brève présentation des corpus sera effectuée, nous présenterons différentes méthodes de transformer les textes en des caractéristiques pour les SVM et enfin les résultats obtenus par les trois méthodes.

3.1. CORPUS D'APPRENTISSAGE (DEFT 2007)

Il y a trois corpus différents dans le cadre du défi : « Critiques de cinéma... », « Tests de jeux vidéo », « Relectures d'articles ». Nous pouvons en extraire un sac de mots pour chaque corpus.

Corpus d'apprentissage	Taille	Nb de phrases	Nb total de mots	Taille du sac de mots	Nb de documents
Critiques de cinéma...	4Mb	34622	792214	48489	2074
Test de jeux vidéo	17Mb	145543	3084878	56185	2537
Relectures d'articles	1,4Mb	11473	218588	13395	881

Tableau 1 – Corpus d'apprentissage

3.2. CODAGES TEXTUELS

Il existe plusieurs méthodes possibles pour représenter l'information textuelles, nous allons en énoncer certaines. Dans un premier temps, on peut considérer un document comme un sac de mots indépendants. Dans un deuxième temps, on peut convertir des séquences de mots (traits) en des codages d'opinions. Enfin, on peut ramener un texte en un ensemble de phrases afin de garder la relation existante entre les mots.

3.2.1. Au niveau des mots

Nous transformons chaque sac de mots en un espace vectoriel, avec les indices des mots, chacun associé à un poids. Le poids Tf-Idf [SAL 88] a été largement utilisé dans le domaine de la Recherche d'Information et de la Fouille de données Textuelles. Il existe plusieurs formulations possibles de Tf-Idf, nous avons décidé d'employer celle qui suit :

$$\text{Tf - Idf} = \text{fréquence_du_mot}_j \times \ln\left(\frac{D}{d_i \subset \text{mot}_j}\right) \quad (9)$$

Où D est le nombre total de documents, d_i est le nombre de documents contenant le mot j

3.2.2. Au niveau des traits

[WIL 05][RIL 06] suggèrent un autre codage afin de représenter les émotions ou les sentiments. Ils annotent *manuellement* un ensemble de traits (séquence de quelques mots) en leur donnant une opinion (très positive, positive, neutre, négative et très négative). Malheureusement, cette tâche nécessite un temps considérable et nous avons donc décidé de ne pas la suivre. Une solution automatique serait alors de prendre toutes les combinaisons possibles de traits dans un texte. Cependant, ceci nous menerait à une explosion combinatoire pour de très longs textes. Néanmoins, on peut limiter ce risque en sélectionnant statistiquement les traits. [CRE 07] ont proposé une méthode basée sur la distance de Kullback-Leibler donnant un *score de saillance* pour chaque trait, défini comme suit :

$$S(t, y = i) = [P(y = i/t) - P(t)] \times \log\left(\frac{P(y = i/t)}{P(t)}\right) \quad (10)$$

Le score de saillance peut donc être calculé pour chaque trait t appartenant à la classe i . Le corpus d'apprentissage est constitué de vecteurs de traits pour chaque document, ces vecteurs sont binaires selon la présence ou non du trait t .

3.2.3. Au niveau des phrases

On peut décrire un document comme une série de phrases. Chaque phrase est transformée en un vecteur binaire identifiant la présence ou non d'un mot. On reprend alors l'hypothèse de sac de mot mais au niveau des phrases. On calcule la probabilité conditionnelle qu'une phrase appartienne à un classe de la manière suivante :

$$P(y/\mathbf{x}) = \sum_{phrase \in x} P(y/phrase) \quad (11)$$

3.3. RESULTATS

On peut retrouver en Figure 1, les résultats obtenus par les trois méthodes sur les trois corpus d'apprentissage fournis par DEFT07. Dans le tableau 2, nous avons utilisé un modèle à base de règle de vote et nous avons cherché à déterminer le codage textuel donnant les meilleures performances (proportion maximale de documents bien classés).

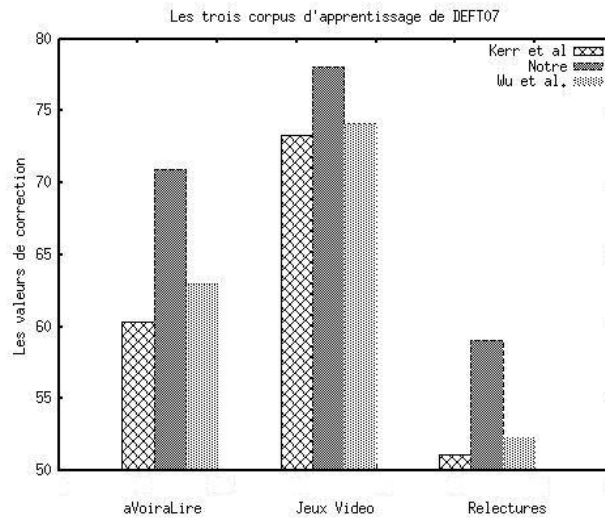


Figure 1: Performances des méthodes sur l'ensemble de test

La sélection des traits (10) réduit considérablement la dimensionnalité ce qui engendre généralement une perte d'information. Cependant, elle permet de réduire considérablement le temps d'apprentissage et de test ce qui est un avantage non négligeable. Pour un système avec règle de vote, le codage sous forme de traits donne les meilleures performances. Toutefois, ce codage ne donne pas les meilleures performances avec notre modèle.

Diviser le corpus d'apprentissage 60% et de test 40%			
Corpus d'apprentissage	Codages textuels		
	Mots	Traits	Phrases
Critiques de cinéma...	60,336	68,149	55,528
Test de jeux vidéo	73,326	76,377	56,102
Relectures d'articles	51,129	59,039	45,197

Tableau 2 : Valeurs de correction pourcentage pour différents codages

4. Conclusion

Nous avons proposé une nouvelle méthode basée sur le modèle exponentiel afin de créer une sortie probabiliste par les SVM dans le cadre d'une classification multi-classes. Nous représentons les textes sous la forme de sac de mots avec pour chacun une score Tf-Idf. Ensuite, nous appliquons des SVM suivant la stratégie du « un contre un » résultant la classe d'appartenance du texte. Enfin, notre modèle convertit l'ensemble des sorties des classifieurs en une estimation probabiliste basée sur la minimisation de la distance de Kullback-Leiber (KL). D'après les expériences réalisées sur les corpus de DEFT07, notre méthode donne de meilleures performances que d'autres systèmes estimant une sortie multi-classes.

5. Bibliographie

- [KER 90] S. Knerr, L. Personnaz and G. Dreyfus, Single-layer training revisited : a stepwise procedure for building and training a neural network. In J. Fogelman, editor, *Neurocomputing: Algorithm, Architectures and Applications*. Springer-Verlag (1990).
- [HAS 98] Hastie, T., and Tibshirani, R. "Classification by Pairwise Coupling". *Advances in Neural Information Processing Systems 10*, M. I. Jordan, M. J. Kearns, S. A. Solla, eds., MIT Press, 1998.
- [PLA 00] J.C. Platt, in *Advances in Large Margin Classifiers*, A. Smola, P. Bartlett, B. Schölkopf, D. Schuurmans, *Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods*. eds., pp. 61-74, MIT Press, (1999).
- [WU 04] Ting-Fan Wu, Chih-Jen Lin, Ruby C. Weng. Probability Estimates for Multi-class Classification by Pairwise Coupling. *Journal of Machine Learning Research* (2004), pp. 975-1005
- [HSU 02] Chih-Wei Hsu; Chih-Jen Lin, A comparison of methods for multiclass support vector machines, *IEEE transactions on Neural Networks*, (2002).
- [WAL 04] Hanna M. Wallach. *Conditional Random Fields: An Introduction*. Technical Report MS-CIS-04-21. Department of Computer and Information Science, University of Pennsylvania, 2004.
- [WIL 05] Theresa Wilson, Janyce Wiebe and Paul Hoffmann (2005). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. *Proceedings of Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, Vancouver, Canada.
- [RIL 06] Ellen Riloff, Siddharth Patwardhan and Janyce Wiebe (2006). Feature Subsumption for Opinion Analysis. *Conference on Empirical Methods in Natural Language Processing (EMNLP-2006)*
- [CRE 07] Eric Crestan, Stéphane Gigandet et Romain Vinot. Approche naïves à l'analyse d'opinion (2007). *Actes de l'atelier du 3ème Défi Fouille de Texte*, pp. 47-56.