SVM et k-ppv pour la reconnaissance d'émotions

Vincent GUIGUE, Alain RAKOTOMAMONJY, Stéphane CANU

PSI - CNRS FRE 2645 - INSA de Rouen
Avenue de l'Université 76801 Saint Étienne du Rouvray Cedex, France
{Vincent.Guigue, Alain.Rakotomamonjy, Stephane.Canu}@insa-rouen.fr

Résumé – La faculté de reconnaître automatiquement les émotions peut s'avérer utile dans le développement du dialogue homme/machine. Nous avons pris la suite des travaux de J. Healey [Hea00] et tenté de relever le défi d'identifier 8 émotions basiques dans des signaux issus de capteurs physiologiques. Le premier obstacle à franchir est la grande dimension du problème : la sélection de variables et l'analyse discriminante linéaire ont permis de réduire la dimensionalité du problème. Le second problème concerne le caractère multi-classes de la discrimination (8 émotions à identifier). Les *Support Vector Machines* (SVM) ont déjà fait leurs preuves en discrimination bi-classes mais ils sont encore peu utilisés dans les autres cas. 3 types de SVM multi-classes seront abordés dans cet article et comparés à 2 méthodes classiques : les k plus proches voisins (kppv) et la discrimination par maximum *a posteriori* sur une modélisation gaussienne (utilisée par J. Healey). Avec un taux de reconnaissance supérieur à 90%, cette expérience est un succès et permet d'envisager des applications dans les domaines du *context aware* et de l'*ubiquitous computing*[CK00].

Abstract – At the present time, one of the great challenge in human and computer interaction is emotion recognition. It would allow computer to have friendly reactions with regards to human behaviour. In this context, we have followed up the J. Healey's work and we aim at recognizing 8 emotions based on physiological signals. After having extracted a large amount of cues, we have to face a high dimensional classification problem. We use variable selection algorithm and linear discriminant analysis to reduce the number of variables. Then, we deal with the multi-class discrimination problem. The Support Vector Machine (SVM) algorithm is acknowledged to be a powerfull method for 2-class discrimination but it is not often used to tackle multi-class problem. 3 implementations of multi-class SVM will be presented in this paper and are compared to 2 classical methods : the k nearest neighbours (knn) and a maximum *a posteriori* algorithm based on a gaussian model (used by J. Healey). Experiments show that using our algorithm we can achieve a recognition rate higher than 90 %. That allows us to consider applications in context aware and ubiquitous computing fields [CK00].

1 Introduction

L'identification automatique des émotions basiques pourrait de nouveaux horizons dans les relations entre hommes et ordinateurs. Au niveau de la reconnaissance vocale par exemple, les progrès réalisés ces dernières années sont considérables. Cependant, le taux de reconnaissance de tels systèmes n'est probant que dans le cas de dictées monotones. En introduisant la connaissance de l'état émotionnel de l'utilisateur, il devient possible de construire un dialogue homme/machine plus naturel. Dans le secteur automobile, l'ordinateur de bord d'une voiture pourrait être informé de l'endormissement du conducteur ou de son niveau de stress, il agirait alors en conséquence. De manière plus générale, la connaissance approfondie de l'utilisateur permet à l'ordinateur de proposer une interface personnalisée et adaptée aux circonstances, de réagir plus intelligemment et même d'anticiper les besoins de cet utilisateur.

Mais le problème d'identification d'émotions est complexe. Le but de cet article est de proposer une méthodologie efficace pour y faire face. La première étape consiste à extraire les variables explicatives à partir de mesures de signaux physiologiques. Le problème obtenu est de très grande dimension. Dans un deuxième temps, l'enjeu est donc de réduire la dimension du problème grâce à différents algorithmes de sélection de variables et d'analyse discriminante linéaire. Enfin, plusieurs méthodes de classification ont été comparées pour étiqueter les données (SVM, k plus proches voisins et maximum *a posteriori* sur une modélisation gaussienne).

Nous étudierons successivement ces 3 phases puis nous verrons les résultats obtenus en fonction des divers paramètres. Le but est de quantifier l'apport lié aux méthodes mises en œuvre dans chacune de ces phases.

2 Méthode

2.1 Données

Les signaux physiologiques correspondant à des états émotionnels sont ceux utilisés par J. Healey dans sa thèse [PH02]. Ils sont issus de 4 capteurs (résistance de la peau, respiration, pression sanguine et electromyogramme) auxquels s'ajoute le rythme cardiaque (déduit des variations de pression sanguine). Ces données représentent 25 minutes d'enregistrement par jour sur une période de 20 jours et sur une personne unique. Chaque jour contient 5 signaux qui représentent les 8 émotions basiques à identifier (pas d'émotion, énervement, haine, peine, amour, amitié, joie, révérence). Nous avons construit 56 variables explicatives pour chaque portion de signal représentant une émotion. Ces 56 variables contiennent les moyennes, écarts-types et densités spectrales (sur une bande de 0 à 0,6 Hz) des cinq signaux issus des capteurs. Finalement, la base d'exemples comporte 160 points repartis dans 8 classes.

Notons que les étiquettes des données représentent des émo-

tions : il s'agit donc de valeurs non objectives. Cela implique que l'expression physiologique d'une même émotion peut varier légèrement d'un jour à l'autre. De plus, les émotions peuvent se mélanger les unes aux autres. Enfin, les capteurs sont sensibles au contexte. L'épaisseur de la couche de gel sous l'électrode de mesure de la résistance de la peau a par exemple une influence importante sur les mesures. Tous ces éléments contribuent à rendre le problème encore plus délicat.

2.2 Réduction de la dimensionalité

2.2.1 Critère de Lambda Wilks

La sélection de variables consiste en une suppression séquentielle de variables (*backward selection*) visant à minimiser le critère de Lambda Wilks [Sap90]. Ce critère permet de mesurer la séparabilité des données :

$$CritLW = \frac{\det(W)}{\det(V)} \tag{1}$$

où W est la matrice de covariance intra-classe et V la matrice de covariance globale. L'objectif est de choisir un sous-ensemble de variables qui minimise le volume moyen de chaque classe et maximise le volume total : les classes sont ainsi mieux séparées.

2.2.2 Critère de maximisation de la marge SVM

L'algorithme complet des SVM est décrit section 2.3.3. Ce critère propose de regarder la sensibilité de chaque variable sur la marge du classifieur [Rak03]. L'introduction des variables unitaires ν_i au niveau de l'expression du noyau gaussien donne :

$$\begin{gathered} K(\nu \cdot x, \nu \cdot y) = exp\left(-\frac{\sum_{i=1}^{d}(\nu_i(x_i - y_i))^2}{2\sigma^2}\right) \\ x, y \in \mathbb{R}^d \end{gathered} \tag{2}$$

où $\cdot$ est le produit terme à terme de 2 vecteurs. La marge est de la forme [Vap98] :

$$\begin{gathered} M = 2/\|\omega\|, \\ \|\omega\|^2 = \sum_i (\alpha_i \alpha_j y_i y_j K(\nu x, \nu y)) \end{gathered} \tag{3}$$

Finalement, la dérivée de $\|\omega\|^2$ par rapport aux ν_i est proportionelle à l'influence de chaque variable i sur la marge. Les variables les moins sensibles sont éliminées en *backward selection*. Ce critère peut être aisément élargi aux SVM *un-contre-un* et *un-contre-tous* en considérant :

$$\sum_{k=1}^{N} \frac{\partial \|\omega_k\|^2}{\partial \nu_i} \tag{4}$$

où N est le nombre de classifieurs.

2.2.3 Analyse discriminante linéaire (ADL)

L'ADL vise à réduire la dimension du problème tout en séparant au mieux les différentes classes [DHS01]. De manière formelle, le problème est de trouver la matrice de projection A qui maximise :

$$\max_A \left(\frac{A^T B A}{A^T V A}\right) \tag{5}$$

où B est la matrice de covariance inter-classes et V la matrice de covariance globale. Le but est de favoriser le regroupement des points d'une même classe et de maximiser la distance entre les *clusters* obtenus. En transformant A en un ensemble de vecteurs $[a_1, (\ldots), a_n]$, le problème devient :

$$V^{-1} B a_i = \lambda a_i, \forall i \in \{1, (\ldots), n\} \tag{6}$$

A se compose donc des vecteurs propres associés aux n premières valeurs propres triées par ordre décroissant de $V^{-1}B$, n étant fixé par l'utilisateur et inférieur au nombre de classes du problème.

2.3 Algorithmes de discrimination

Soit une base de données étiquetée de n_{pts} points répartis en n_{Cl} classes :

$$\{x_i, y_i\}, i \in \{1, (\ldots), n_{pts}\}\ y_i \in \{1, (\ldots), n_{Cl}\}$$

L'enjeu est de trouver une règle de décision :

$$\begin{array}{rccc} f: & \mathbb{R}^d & \rightarrow & \mathcal{A} \\ & x & \rightarrow & f(x) \end{array} \tag{7}$$

telle que $f(x)$ soit une bonne prédiction de l'étiquette de x au sens d'un critère donné.

La base de données sera divisée en un ensemble d'apprentissage de n_{app} points et un ensemble de test de n_t points.

2.3.1 K plus proches voisins

Le classifieur des k plus proches voisins (kppv) est un classifieur universel de référence. Pour un problème n_{Cl}-classes, chaque point de test prend l'étiquette de la classe dominante parmi ses kppv :

$$\begin{gathered} f(x) = y_j \text{ avec} \\ \text{occ}(y_j) > \text{occ}(y_i),\ \forall i \neq j,\ i \in \{1, (\ldots), n_{Cl}\} \end{gathered} \tag{8}$$

où $\text{occ}(y_i)$ est l'opérateur définissant le nombre d'occurrences de y_i parmi les étiquettes des kppv de x.

2.3.2 Maximum *a posteriori* sur une modélisation gaussienne (MAP gaussien)

Dans cette méthode (retenue par J. Healey [PVH01]), chaque classe k est assimilée à une gaussienne dont la moyenne μ_k et la variance Σ_k sont déterminées sur les points d'apprentissage. La classe des points de test est obtenue de la manière suivante (en supposant que les classes sont équiprobables) [DHS01] :

$$\begin{gathered} f(x) = \arg\max_k \left(\exp(-(x-\mu_k)\Sigma_k^{-1}(x-\mu_k))\right), \\ k \in \{1, (\ldots), n_{Cl}\} \end{gathered} \tag{9}$$

2.3.3 SVM multi-classes

L'algorithme des *Support Vector Machines* (SVM) décrit par V. Vapnik [Vap98] propose de projeter les points dans un nouvel espace $\mathcal{H}$ muni d'un produit scalaire $\langle ., . \rangle$ grâce à une fonction $\phi(x)$. En pratique, cette transformation est implicite dans le noyau $K(x, y) = \langle \phi(x), \phi(y) \rangle$. La frontière de décision est de la forme :

$$f(x) = \langle \omega, \phi(x) \rangle_{\mathcal{H}} + b = \sum_{i=1}^{n_{app}} \alpha_i K(x_i, x) + b \tag{10}$$

Dans le cas où les données sont séparables, la frontière $f(x)$ optimale (qui maximise la marge entre les classes) est obtenue en résolvant le problème quadratique suivant :

$$\begin{array}{c} \min_{\omega} \frac{1}{2}\|\omega\|^2 \\ \text{sous les contraintes :} \\ y_i(\langle \omega, \phi(x_i)\rangle + b) \geq 1 \; \forall i \in \{1, (\ldots), n_{app}\} \end{array} \quad (11)$$

Pour pouvoir traiter plus de deux classes, il convient d'apporter les modifications nécessaires. L'article de C. W. Hsu [HL02] compare trois approches du problème SVM multi-classes. Les deux premières méthodes sont basées sur une multiplication des classifieurs bi-classes tandis que la dernière propose une résolution globale.

- **Un-contre-tous.** Le un-contre-tous a été la première réponse proposée pour faire face aux problèmes multi-classes. Chaque classe est opposée à toutes les autres. Il faut donc poser n_{Cl} problèmes binaires. La fusion des résultats est très simple :

$$\begin{array}{c} Classe\ de\ x = \arg\max_k(f_k(x)), \\ k \in \{1, (\ldots), n_{Cl}\} \end{array} \quad (12)$$

- **Un-contre-un.** Cette solution consiste à créer tous les classifieurs bi-classes envisageables du problème, c'est à dire $n_{Cl}(n_{Cl}-1)/2$ classifieurs binaires. Chaque classifieur vote pour tous les points et chaque point se voit attribuer la classe qui a reçu le plus de suffrages.
- **Méthode globale.** Cet algorithme est décrit très clairement par J. Weston [WW98]. Le problème est de trouver n_{Cl} classifieurs

$$f_k(x) = \langle \omega_k, \phi(x)\rangle + b_k, \forall k \in \{1, (\ldots), n_{Cl}\} \quad (13)$$

qui minimisent :

$$\begin{array}{c} \min_{\omega} \frac{1}{2}\|\omega_k\|^2 \\ \text{sous les contraintes :} \\ f_{y_i}(x_i) \geq f_k(x_i), \quad k \in \{1, (\ldots), n_{Cl}\} \\ k \neq y_i \\ i = 1, (\ldots), n_{app} \end{array} \quad (14)$$

L'introduction de variables de relâchement ξ permet de traiter les problèmes non séparables. C représente la pénalité pour les points d'apprentissage se trouvant au delà de la marge. Dans le cas bi-classes, il est nécessaire de minimiser :

$$\min_{\omega,\xi} \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n_{app}} \xi_i \quad (15)$$

et de transformer les contraintes en :

$$\begin{array}{l} y_i(\langle \omega, \phi(x_i)\rangle + b) \geq 1 - \xi_i \\ \xi_i \geq 0, \; \forall i \in \{1, (\ldots), n_{app}\} \end{array} \quad (16)$$

3 Résultats

L'évaluation des performances a été effectuée par estimation non biaisée de l'erreur en généralisation (méthode du *leave-one-out* [HTF01]). Les résultats présentés dans les tableaux 1 et 2 sont des résultats optimisés. Les paramètres ont été échantillonnés puis toutes les combinaisons ont été évaluées (toujours en *leave-one-out*) afin de trouver les paramètres optimaux.

TAB. 1 – Meilleurs résultats obtenus (en pourcentage de reconnaissance) - Pré-traitements simples.

	sans trait.	SV LW	SV SVM1/1	ADL seule
SVM 1/1	53,12%	66,87%	**68.75%**	80,63%
SVM 1/tous	52,50%	66,25%	-	79,37%
SVM k-class	**55,00%**	**67,50%**	-	80,63%
kppv	37,50%	39,37%	-	**81,25%**
MAP gauss.	41,88%	41,25%	-	77,50%

TAB. 2 – Meilleurs résultats obtenus (en pourcentage de reconnaissance) - Pré-traitements combinés.

	sans trait.	ADL+ SV (LW)	ADL+ SV (SVM 1/1)	ADL+ SC+ SV (LW)
SVM 1/1	53,12%	87,50%	**88,13%**	**90,62%**
SVM 1/tous	52,50%	85,00%	-	88,12%
SVM k-class	**55,00%**	86,68%	-	88,75%
kppv	37,50%	**88,75%**	-	**90,62%**
MAP gauss.	41,88%	83,75%	-	85,00%

3.1 Paramètres utilisés

Les algorithmes de réduction de la dimensionalité nécessitent différents réglages : il faut leur donner le nombre de variables à éliminer (pour la sélection de variables) et le nombre d'axes de projection pour l'analyse discriminante linéaire. Les résultats optimaux ont été obtenus en sélectionnant entre 20 et 30 variables puis en projetant les points sur 3 à 5 axes discriminants. Sur ces 2 plages de valeurs, les performances sont constantes. Le réglage optimal est le même lorsque plusieurs algorithmes sont conjugués.

L'algorithme noté SC (sélection de capteur) dans le tableau 2 consiste en une élimination manuelle du capteur de pression sanguine.

Les SVM utilisés sont basés sur des noyaux gaussiens. Le couple de paramètres $(\sigma = 0.08, C = 1000)$ est optimal lorsque les données sont traitées par ADL. Dans le cas contraire, c'est le couple $(\sigma = 2, C = 200)$ qui a donné les meilleurs résultats. Le paramètre σ du noyau est beaucoup plus sensible que C.

Les kppv se sont montrés particulièrement robustes. Les performances sont optimales pour k appartenant à $\{5, (\ldots), 15\}$, elles faiblissent lentement en s'éloignant de cet intervalle.

L'algorithme du MAP gaussien est adaptatif, il ne nécessite aucun réglage.

3.2 Bilan des Résultats

Les meilleurs résultats sont issus du couplage de plusieurs algorithmes de réduction de la dimensionalité. Ces pré-traitements effectuent la majeure partie du travail de discrimination. Cela explique le fait que les SVM perdent leur avantage par rapport aux algorithmes plus simples sur les données prétraitées. Cependant, ces méthodes sont encore perfectibles comme en témoignent les résultats obtenus en supprimant manuellement les données liées au capteur de pression sanguine. Ce phénomène met en évidence une faiblesse des algorithmes de sélection de variables. En effet, ils auraient du éliminer les variables qui apportent du bruit.

Une partie du gain de performance par rapport au travail de

J. Healey [PVH01] (90,62% contre 81% de bonne classification) est du à l'augmentation du nombre de variables explicatives (densités spectrales effectuées sur une plus large bande de fréquences). En utilisant les méthodes décrites dans cet articles sur les variables explicatives de J. Healey, le taux de reconnaissance optimal est de 86,88%.

4 Conclusion

L'application traitée dans cet article reste particulière, cependant, une méthodologie globale de traitement du signal peut être extraite des travaux présentés.

La première phase repose sur l'extraction d'un maximum de variables explicatives : cela permet de limiter les pertes d'information. Néanmoins, augmenter le nombre de variables de manière arbitraire ne garantit pas une bonne discrimination. Cette augmentation doit être contrôlée de manière appropriée par une méthode de sélection de variables. La sélection de variables est un problème d'actualité ([GE03]) et les algorithmes sont maintenant capables de sélectionner un petit sous-ensemble de variables (de l'ordre de 20) parmi un grand nombre de variables explicatives (de l'ordre de 5000). Une fois, les variables pertinentes sélectionnées, une analyse discriminante linéaire permet de trouver une transformation linéaire de l'espace augmentant la séparabilité des classes en réduisant encore la dimension du problème. La combinaison de ces 2 techniques donne les résultats les plus pertinents.

L'étape suivante est de choisir un algorithme de classification. Ici les résultats paraissent moins clairs : les SVM et les kppv font à peu près jeu égal, mais les SVM demandent beaucoup plus de temps de calcul. Il semble clair que la plus grande partie du travail de discrimination est effectuée au niveau de l'ADL, une fois les pré-traitements réalisés, les SVM perdent leur avantage sur les kppv.

Cette méthodologie modulaire est simple et efficace. Cependant, il reste un certain nombre de perspectives pour améliorer les résultats obtenus. Il faudrait construire un noyau particulièrement adapté à la reconnaissance d'émotions pour profiter pleinement de la puissance des SVM. Des noyaux asymétriques permettraient par exemple de bénéficier de la multi-résolution (comme pour les kppv) tout en maximisant la marge entre les classes [Tsu98]. Un autre axe de recherche concerne l'intégration d'algorithme de sélection de variables à l'intérieur même de techniques d'apprentissage comme les SVM. Cependant, la méthode décrite par Y. Grandvalet [GC02] ne s'est pas montrée satisfaisante pour ce problème.

Finalement, il est important de garder à l'esprit le caractère subjectif des données. Les émotions ne sont pas des *valeurs exactes* de plus, elles sont souvent combinées les unes aux autres. En ajoutant la variabilité du protocole de mesure et le faible nombre de points d'apprentissage il semble évident que le taux de reconnaissance ne sera jamais parfait.

Références

[CK00] Guanling Chen and David Kotz. A survey of context-aware mobile computing research. Technical Report TR2000-381, Dept. of Computer Science, Dartmouth College, November 2000.

[DHS01] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. Wiley Interscience, 2001.

[GC02] Yves Grandvalet and Stéphane Canu. Adaptive scaling for feature selection in svms. Vancouver, 2002. NIPS.

[GE03] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3 :1157–1182, 2003.

[Hea00] Jennifer Healey. *Wearable and Automotive Systems for the Recognition of Affect from Physiology*. PhD thesis, MIT, 2000.

[HL02] Chih-Wei Hsu and Chih-Jen Lin. A comparison of methods for multi-class support vector machines. *IEEE Transactions on Neural Networks*, 13 :415–425, 2002.

[HTF01] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, 2001.

[PH02] Rosalind W. Picard and Jennifer Healey. Eight-emotion sentics data. MIT Affective Computing Group `http ://affect.media.mit.edu`, 2002.

[PVH01] Rosalind W. Picard, Elias Vyzas, and Jennifer Healey. Toward machine emotional intelligence : Analysis of affective physiological state. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 23(10) :1175–1191, 2001.

[Rak03] Alain Rakotomamonjy. Variable selection using svm-based criteria. *Journal of Machine Learning Research*, 3 :1357–1370, 2003.

[Sap90] Gilbert Saporta. *Probabilités, Analyse de données et Statistiques*. Editions Technip, 1990.

[Tsu98] Koji Tsuda. Support vector classifier with asymmetric kernel functions. Technical Report ETL-TR-98-31, Machine Understanding Division, Electrotechnical Laboratory, Japan, 1998.

[Vap98] Vladimir N. Vapnik. *The Statistical Learning Theory*. Springer, 1998.

[WW98] Jason Weston and Chris Watkins. Multi-class support vector machines. Technical Report CSD-TR-98-04, Department of Computer Science, Royal Holloway, University of London, UK, 1998.