HAL open science

Informatique Affective : Classification de l'Etat Emotionel

Stephane Cholet, Helene Paugam-Moisy, Sebastien Regis, Lionel Prevost

To cite this version:

Stephane Cholet, Helene Paugam-Moisy, Sebastien Regis, Lionel Prevost. Informatique Affective : Classification de l'Etat Emotionel. Extraction et Gestion des Connaissances, Université Grenoble Alpes, Jan 2017, Grenoble, France. hal-01513329

HAL Id: hal-01513329

https://hal.science/hal-01513329v1

Submitted on 24 Apr 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - ShareAlike 4.0 International License

Informatique Affective : Classification de l'Etat Emotionel

S. Cholet*, H. Paugam-Moisy*
S. Regis*, L. Prevost**

* LAMIA, LAboratoire de Mathématiques, Informatique et Applications(EA 4540),
Université des Antilles, BP 592, 97157 Pointe-à-Pitre,
Guadeloupe (French West Indies), France.
stephane.cholet@univ-antilles.fr,
**ESIEA, Ecole d'Ingénieurs du Monde Numérique,
9 rue Vesale, 75005 PARIS, France.

Résumé. *"L'informatique affective est l'informatique qui se réfère à, résulte de, ou influence délibérément les émotions"* (Picard (1997)). Singulièrement, l'équilibre psychosocial est désormais considéré comme une préoccupation majeure pour la santé, tant sur le plan individuel que public. Prédire les émotions humaines via des méthodes non-invasives est un challenge propulsé notamment par le développement des systèmes d'accompagnement intelligents. Afin de déterminer l'état émotionel d'un individu, ayant potentiellement un trouble psychosocial, il convient de traiter tant les émotions expressives que neutres. Une approche originale de classification de l'information émotionelle à partir de vidéos est proposée.

1 Introduction

La prédiction des émotions humaines est un challenge majeur, propulsé notamment par le développement des systèmes d'accompagnement intelligents. Afin de mettre au point un système de prévention des risques psychosociaux, nous cherchons à déterminer l'état émotionnel d'un individu, potentiellement dépressif, en utilisant des enregistrements non invasifs, récoltés dans des situations de vie courante. Ces grandes masses de données (*big data*) existent dans la base de données du challenge AVEC2014, qui propose des données images et audio d'individus en conversation avec un avatar, dans le cadre de séquences de dialogue semi-dirigé. Cet article présente les méthodes et procédés mis en œuvre pour classifier et prédire l'état émotionnel d'individus à partir d'enregistrements vidéo. Dans un premier temps, les données seront présentées, ainsi que les prétraitements opérés en vue de leur classification. Ensuite, nous expliquerons les méthodes de classification au moyen de *Support Vector Machines* (SVM). Les parties 4 et 5 proposeront, respectivement, une analyse des résultats et des pistes envisagées pour la poursuite des travaux.

2 Données et traitements préliminaires

2.1 Présentation des données

Les données sont issues de la base de données du challenge AVEC2014, organisé par Valstar et al. (2014) à Orlando (Floride, Etats-Unis) . Elles sont constituées de vidéos, associant un flux d'images et un flux sonore, ainsi que de descripteurs visuels et de descripteurs audio de la littérature, fournis par les auteurs. Les individus filmés interagissent avec un avatar, dans des conditions réelles d'utilisation d'un ordinateur (webcam et micro du commerce, luminosité variant d'une vidéo à l'autre). Certains d'entre eux sont dépressifs ou souffrent de divers troubles psychosociaux, ce qui constitue une source de données diversifiée pour nos travaux.

Pour une première étude, nous avons décidé de ne traiter que les données visuelles. Les descripteurs visuels sont des histogrammes LGBP (*Local Gabor Binary Pattern*, ou motifs locaux binaires de Gabor), tels que décrits par Chan et al. (2007), de la face du sujet. Les visages (un par séquence vidéo) sont détectés automatiquement en amont, grâce à l'outil de détection de Viola et Jones (2001). Chaque image est associée à un vecteur de descripteurs de dimension $d_{initial} = 16\ 992$. La base comprend environ $n_{initial} = 120\ 000$ images. Chaque image a été étiquetée par un ensemble d'individus non-experts, suivant la modélisation de Russell (1980), dans les trois dimensions affectives *arousal, valence* et *dominance*.

L'*arousal* définit dans quelle mesure l'émotion ressentie est associée à une sensation d'énergie, i.e. d'activité ou de passivité d'un individu. La *valence*, elle, exprime l'état d'esprit positif ou négatif du sujet (Gerber et al. (2008), et la *dominance* le degré de contrôle ressenti. Il convient de noter que ces trois concepts s'évaluent toujours vis-à-vis de la situation courante. Le modèle de Russell est repris dans la Figure 1.

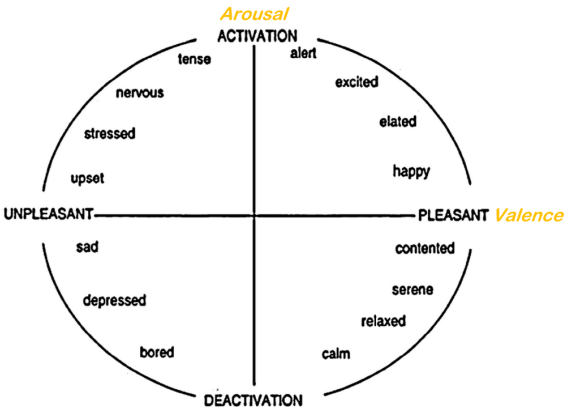

FIG. 1 – *Circumplex de Russell (1980).*

La complexité des données provient notamment de leur volume et de leur dimension qui sont tous deux importants. Cette complexité a nécessité la mise en œuvre de prétraitements en vue de leur classification.

2.2 Traitements préliminaires

Dans la mesure où les données seront fournies en entrée d'un classifieur neuronal, il a fallu les préparer afin d'éviter tout biais dans les résultats.

L'analyse des données a révélé qu'à certaines images étaient associés des descripteurs invalides. i.e. nuls. Ces données sont aberrantes, puisqu'elles associent des descripteurs nuls à des intensités émotionnelles valides. Elles ont donc été supprimées. Ces erreurs sont dues, par exemple, à l'échec de détection d'un visage dans une image. On passe ainsi à une base de $n_{final} = 100\,000$ images, soit 20 000 exemples ainsi supprimés.

Chacune des variables a été normalisée. Le fait de classifier un si grand nombre d'exemples, chacun ayant une très grande dimension, nécessite un temps de calcul conséquent. En appliquant une Analyse en Composante Principale, la dimension des données a été réduite à $d_{final} = 3\,000$. Cette réduction d'un facteur supérieur à 5 est justifiée puisque les dimensions retenues expriment 99% de l'information contenue dans les $16\,992$ descripteurs initiaux.

La dimension des données a donc été réduite d'un facteur 5 ($d_{final} = 3\,000$) en appliquant une Analyse en Composantes Principales. Cette réduction est d'autant plus justifiée que les dimensions retenues expriment 99% de l'information contenue dans les $16\,992$ descripteurs initiaux.

Les traitements de données ont été réalisés sur Wahoo, le cluster du Centre Commun de Calcul Intensif de l'Université des Antilles.

3 Expérimentations

Dans la littérature, la plupart des méthodes mises au point sur ces données étaient des méthodes de régression, comme Kächele et al. (2014) et Williamson et al. (2014). Adoptant une approche différente, nous avons choisi d'utiliser une classification basée sur les SVM, dans la mesure où cette méthode a permis d'obtenir de bonnes performances dans un contexte proche (voir Senechal et al. (2010)). Par ailleurs, seule la dimension affective *arousal* sera considérée pour cette première étude. Deux jeux de données sont utilisés : le premier jeu, le *trainset*, d'environ $50\,000$ images sert à construire les classifieurs, alors que le *development set*, tout aussi conséquent, permet de tester la performance en généralisation. Les étiquettes du *testset* ne sont pas fournies, ne permettant pas d'évaluer de performances sur ce jeu.

3.1 Classification

Un SVM, parfois appelé Séparateur à Vaste Marge en français (Canu (2014)), réalise des séparations non-linéaires entre deux classes (Cortes et Vapnik (1995)). La méthode est basée sur deux principes fondamentaux : d'une part, sur le plongement virtuel des données dans un espace de dimension supérieure, au moyen d'une fonction noyau, et d'autre part, sur une séparation de marge optimale par un perceptron dans l'espace de dimension supérieure (voir Guermeur et Paugam-Moisy (1999)).

Ici, la classification se pose comme une étape préliminaire à d'autres traitements prédictifs. Il s'agit donc de déterminer des classes d'intensité émotionelle, qui peuvent être à la base du déclenchement d'alertes dans un système de prévention des risques psychosociaux. On peut noter que dans de telles applications, la valeur précise de l'intensité émotionnelle peut être

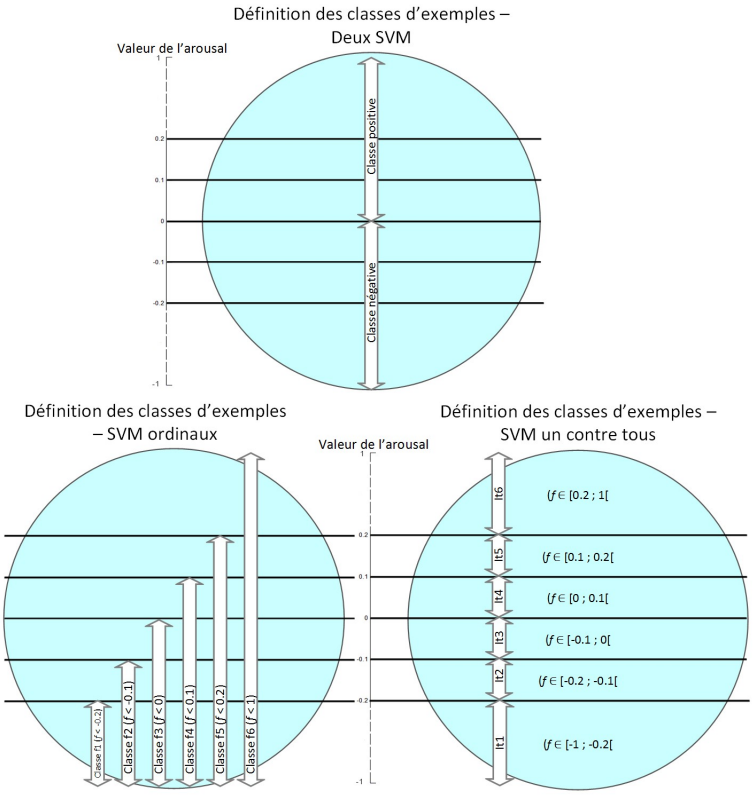

FIG. 2 – *En haut : construction des classes pour l'approche 1 (deux SVM) - A gauche : construction des classes pour l'approche 2 (SVM ordinaux) - A droite : construction des classes pour l'approche 3 (SVM un contre tous).*

moins pertinente ou plus difficilement exploitable que sa tendance ou sa classe. Trois approches sont présentées pour la classification des descripteurs visuels, comme le résument les schémas de la Figure 2.

Une première approche consiste à classifier le signe de l'émotion, i.e. à déterminer si un individu en test est passif (*arousal* négatif) ou actif (*arousal* positif). En effet, il peut exister une corrélation entre l'*arousal* d'un individu et son état dépressif (voir Valstar et al. (2014)). Un SVM est alors appris et spécialisé dans la classification des exemples dont l'étiquette k est positive $(k >= 0)$. On peut ainsi séparer les exemples positifs des exemples négatifs.

La deuxième approche prend d'avantage en compte la distribution des étiquettes : en effet, si théoriquement le cercle de Russell prévoit un *arousal* entre -1 et 1, les données utilisées présentent des étiquettes dans $I' = [-0.4; +0.3]$, distribués comme dans la Figure 3.

Ainsi, suivant le modèle des SVM ordinaux décrit par Frank et Hall (2001), six SVM ont été entrainés, chacun spécialisé pour la reconnaissance des exemples dont l'étiquette k est in-

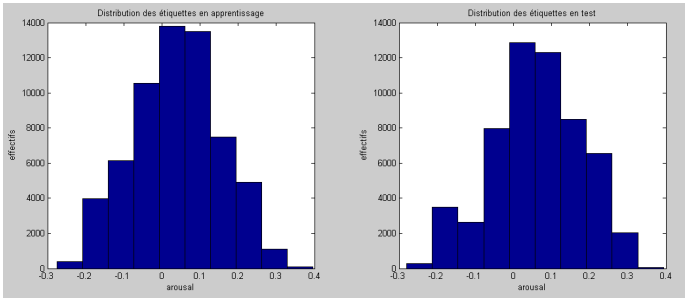

FIG. 3 – *Dans les données, la dimension arousal prend ses valeurs dans* $[-0.3; 0.4]$. *A gauche : distribution des étiquettes d'apprentissage - A droite : distribution des étiquettes de test.*

férieure à une valeur frontière $f_i \in \{-0.2, -0.1, 0, 0.1, 0.2, 1\}$. Cette méthode est inspirée des travaux de Prevost et al. (2014) qui visaient à déterminer l'âge d'un individu sur des images.

Une variante et troisième approche a consisté à spécialiser chacun des classifieurs dans la reconnaissance des exemples dont l'étiquette est comprise dans un intervalle donné $I_{ti} \in \{[-1; -0.2[, ..., [0.2; 1[\}$, i.e. à opposer chacune des classes contre toutes les autres. Il s'agit de l'implémentation de la méthode SVM Un Contre Tous, ou SVM One Against All expliquée par Milgram et al. (2006).

3.2 Résultats

Les résultats sont donnés dans le Tableau 1 pour la deuxième et la troisième approche, i.e. pour les SVM ordinaux et les SVM un contre tous.

TAB. 1 – *Comparaison des taux de succès pour les deux méthodes utilisées : à gauche les SVM ordinaux et à droite les SVM un contre tous.*

	Taux de succès (%) SVM ordinaux		Taux de succès (%) SVM un contre tous
f_1	**97.23**	I_{t1}	**97.23**
f_2	86.02	I_{t2}	**88.78**
f_3	63.16	I_{t3}	**75.2**
f_4	**71.92**	I_{t4}	64.91
f_5	**92.63**	I_{t5}	79.28
f_6	**99.93**	I_{t6}	92.63
Moyenne	**85.14**	Moyenne	83.01

La première approche n'a pas permis d'obtenir de bons résultats, qui variaient entre 60 et 70% de taux succès. Les taux de succès pour les deuxième et troisième approche sont donnés dans la Figure 4. On note que ces deux méthodes fournissent de bons taux de succès pour toutes les classes, bien que certaines peinent un peu plus à être reconnues que d'autres. En effet, les

meilleurs taux de succès sont obtenus via la méthode ordinale, sur les classes « extrêmes », avec 97.23% pour la frontière $f1$ et 99.93% pour la frontière $f6$. C'est la frontière $f3$, qui a le plus de mal à être reconnue pour les SVM ordinaux (63.16%). Pour les SVM un contre tous, c'est l'intervalle I_{t4} qui peine le plus à être reconnu (64.91%). Dans les deux cas, il s'agit de classes correspondant à des valeurs d'arousal proches de zéro.

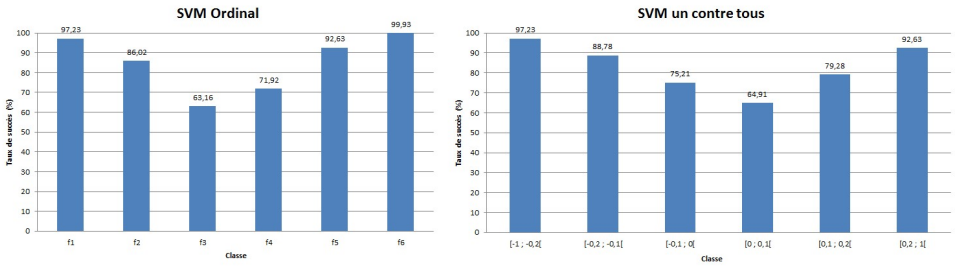

FIG. 4 – *Taux de succès des classifieurs SVM ordinaux et un contre tous.*

Le challenge AVEC2011, qui demandait de classifier les données (voir Schuller et al. (2011)), a utilisé des données différentes du challenge AVEC2014. Cependant, elles sont de même type (vidéo et bande son, ainsi que des descripteurs visuels et des descripteurs audio). Les étiquettes étaient également fournies dans la dimension affective *arousal*. Les vainqueurs (Ramirez et al. (2011)) ont mis en œuvre des *Latent-Dynamic Conditional Random Field* sur les données et ont obtenu un taux de succès de 81.7% sur le *development set* de 2011, soit environ 3% de moins que notre méthode.

A titre indicatif, les vainqueurs du challenge AVEC2014 ont obtenu un score (coefficient de Pearson) de $r = 0.59$. Ce résultat, n'est pas comparable aux performances obtenues ici, pour plusieurs raisons, et notamment parce-que nous faisons de la classification et non de la régression.

4 Conclusion

La classification des données visuelles par des SVM a permis de mettre en valeur des discriminations intéressantes.

En particulier, les résultats font apparaître un phénomène de zone neutre : les émotions expressives sont mieux reconnues que les émotions neutres, malgré un plus grand nombre d'exemples d'apprentissage pour ces dernières. Cependant, l'utilisation de la simple modalité visuelle ne peut pas discriminer à elle seule l'ensemble des intensités émotionnelles.

Dans un premier temps, on étudiera si la mise en place d'un taux de rejet peut améliorer les résultats. La poursuite des travaux consistera à intégrer dans la classification des informations sur les données provenant de la modalité auditive.

5 Perspectives

Ce papier décrit des travaux effectués dans le cadre d'un projet de création d'un système de prévention des risques psychosociaux, visant à déterminer l'état émotionnel d'un individu au moyen de données vidéo. Les SVM ont été utilisés afin de classifier l'état émotionnel, et ont donné des résultats très encourageants pour la suite. Toutefois, l'utilisation des SVM n'est pas adaptée à une utilisation dans le cadre d'une classification en temps réel. Pour prendre en compte cet aspect, le *reservoir computing* ou les *spiking neurons* pourront apporter des possibilités intéressantes. Enfin, la segmentation de l'image devra être repensée afin de considérer non plus une zone fixe englobant le visage, mais des zones d'intérêt sur la face du sujet. L'aspect temporel (i.e. la proximité émotionnelle entre deux images successives), ainsi que la spécialisation à un seul utilisateur sont des pistes également envisagées.

Références

Canu, S. (2014). Svm and kernel machines : linear and non-linear classification. Ocean's Big Data Mining. https ://oceandatamining.sciencesconf.org/.

Chan, C.-H., J. Kittler, et K. Messer (2007). Multi-scale local binary pattern histograms for face recognition. In *International Conference on Biometrics*, pp. 809–818. Springer.

Cortes, C. et V. Vapnik (1995). Support-vector networks. *Machine Learning 20*, 273–297.

Frank, E. et M. Hall (2001). A simple approach to ordinal classification. In *European Conference on Machine Learning*, pp. 145–156. Springer.

Gerber, A. J., J. Posner, D. Gorman, T. Colibazzi, S. Yu, Z. Wang, A. Kangarlu, H. Zhu, J. Russell, et B. S. Peterson (2008). An affective circumplex model of neural systems subserving valence, arousal, and cognitive overlay during the appraisal of emotional faces. *Neuropsychologia 46*(8), 2129–2139.

Guermeur, Y. et H. Paugam-Moisy (1999). *Théorie de l'apprentissage de Vapnik et SVM, Support Vector Machines*, pp. 109–138. Hermès.

Kächele, M., M. Schels, et F. Schwenker (2014). Inferring depression and affect from application dependent meta knowledge. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pp. 41–48. ACM.

Milgram, J., M. Cheriet, et R. Sabourin (2006). "one against one" or "one against all" : Which one is better for handwriting recognition with svms ? In *Tenth International Workshop on Frontiers in Handwriting Recognition*. Suvisoft.

Picard, R. W. (1997). *Affective computing*, Volume 252. MIT press Cambridge.

Prevost, L., P. Phothisane, et E. Bigorgne (2014). Live stream oriented age and gender estimation using boosted lbp histograms comparisons. In *Proceedings of the 3rd International Conference on Pattern Recognition Applications and Methods*, pp. 790–798. SCITEPRESS-Science and Technology Publications, Lda.

Ramirez, G. A., T. Baltrušaitis, et L.-P. Morency (2011). Modeling latent discriminative dynamic of multi-dimensional affective signals. In *Affective Computing and Intelligent Interaction*, pp. 396–406. Springer.

Russell, J. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology 39*(6), 1161–1178.

Schuller, B., M. Valstar, F. Eyben, G. McKeown, R. Cowie, et M. Pantic (2011). Avec 2011–the first international audio/visual emotion challenge. In *International Conference on Affective Computing and Intelligent Interaction*, pp. 415–424. Springer.

Senechal, T., K. Bailly, et L. Prevost (2010). Automatic facial action detection using histogram variation between emotional states. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pp. 3752–3755. IEEE.

Valstar, M., B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, et M. Pantic (2014). Avec 2014 : 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pp. 3–10. ACM.

Viola, P. et M. Jones (2001). Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, Volume 1, pp. I–511. IEEE.

Williamson, J. R., T. F. Quatieri, B. S. Helfer, G. Ciccarelli, et D. D. Mehta (2014). Vocal and facial biomarkers of depression based on motor incoordination and timing. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pp. 65–72. ACM.

Summary

"Affective Computing is computing that relates to, arises from, or deliberately influences emotions" (Picard (1997)). Specifically, psychosocial equilibrium is now considered as a major issue for both individual and public health. Predicting human emotions via non-intrusive methods is a great challenge triggered by the rise of intelligent assisting systems. In order to determine the emotional state of a subject having a potential psychosocial disorder, it is required to process both neutral and expressive emotions. A computer-engineered method aiming to provide emotional information from videos is proposed. An original approach for affect classification and its interpretation are presented.