HAL open science

Classification supervisée avec second étage optionnel pour variables de covariance conditionnelle hétérogène

Thomas Burger, Thierry Dhorne

To cite this version:

Thomas Burger, Thierry Dhorne. Classification supervisée avec second étage optionnel pour variables de covariance conditionnelle hétérogène. XVI ème rencontre de la Société Francophone de Classification, Sep 2009, Grenoble, France. pp.45-48. hal-00394979

HAL Id: hal-00394979

https://hal.science/hal-00394979v1

Submitted on 6 May 2021

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.

Classification supervisée avec second étage optionnel pour variables de covariance conditionnelle hétérogène

Thomas Burger, Thierry Dhorne

Université Européenne de Bretagne, Université de Bretagne-Sud, CNRS, Lab-STICC, Centre de Recherche Yves Coppens, BP 573, F-56017 Vannes cedex, FRANCE
{prenom}.{nom}@univ-ubs.fr – http ://www-labsticc.univ-ubs.fr/˜burger/

RÉSUMÉ. *En informatique, il est fréquent qu'un problème de classification fasse intervenir des variables de nature hétérogène, dont les différentes covariances (conditionnellement à la classe) n'ont pas le même ordre de grandeur. Si la procédure d'apprentissage supervisé est basée sur des modèles génératifs, certaines variables, de faible covariance, mais discriminantes peuvent ne pas être suffisamment considérées. Nous présentons un schéma de classification à deux étages permettant de ne pas perdre l'information discriminante issue de ces variables sous-représentées durant l'apprentissage. Le second étage est optionnel : il n'est utilisé que lorsqu'une information pertinente est manquante pour la classification, et que celle-ci peut provenir de variables sous-représentées. La difficulté du problème est de déterminer automatiquement dans quels cas le second étage doit être utilisé. Pour résoudre ce problème, nous proposons de travailler dans le formalisme des fonctions de croyance (FC), et d'utiliser une procédure de décision basée sur une généralisation de la transformée pignistique.*

MOTS-CLÉS : *Classification de données multimodales/multimédia, apprentissage automatique, modèles génératifs, fonctions de croyance, transformée pignistique*

1. Introduction

En informatique, les problèmes de classification supervisée ont des spécificités [DUD 01] qui rendent parfois l'utilisation des méthodes statistiques classiques inadaptées. Ainsi, en reconnaissance de forme (vision par ordinateur, compression MPG7), en reconnaissance de la parole [RAB 89], en indexation automatique de contenus multimédia [MUL 07], nous sommes confrontés aux problèmes suivants :

- bruit des données (dû à la compression de l'image, à la qualité des capteurs sonores/vidéos, etc).
- besoin de généricité par rapport au nombre de classes.
- contraintes machines : sensibilité des conditions d'acquisitions, calcul temps-réel, bufferisation, etc.
- nature hétérogène des variables de classification (il est en effet fréquent de mélanger des descripteurs de couleur, de texture, de forme, etc.), dont la prise en compte de manière conjointe est difficile.

En raison de tout cela, il est classique d'utiliser des méthodes génératives plutôt que discriminantes [ARA 06] : HMM avec observation gaussiennes, ACP et modèles des classes basés sur des mélanges gaussiens, etc..

Cependant, l'intérêt des méthodes génératives pour la réduction du bruit peut aussi devenir un inconvénient. Les modèles génératifs ne considèrent que les grandes tendances des échantillons d'apprentissage, et "lissent" le reste en perdant une partie de l'information. Or, dans certains cas, c'est cette information perdue qui permet la séparation entre les classes. En pratique, cela peut en particulier arriver si, conditionnellement à chaque classe, les matrices de covariance sont différentes. Et cela arrive de manière fréquente lorsque les variables sont issus de différents jeux de descripteurs de nature hétérogène.

Dans [ARA 07, ARA 08, ARA 09], nous nous sommes intéressés à la reconnaissance automatique de gestes multimodaux de la Langue des Signes Américaine (ASL). Tous les problèmes énumérés ci-dessus ont été rencontrés. Dans ces articles, nous avons apporté plusieurs éléments de solutions, et plusieurs variantes. Néanmoins :

– La méthode de classification proprement dite n'est ni isolée de son application, ni des autres traitements informatiques n'ayant pas de rapport avec la classification. Elle n'est pas isolée non plus des prétraitements que nous avons dû utiliser pour standardiser le problème vis-à-vis des autres méthodes de l'état de l'art, et ainsi proposer un protocole de comparaison rigoureux.
– Les deux transformations mathématiques appliquées aux structures d'information mise en jeu dans l'algorithme de classification ne sont pas justifiées d'un point de vue théorique. Depuis, la première a même été améliorée, et cette amélioration correspond au travail publié en parallèle par une autre équipe de recherche [DUB 08]. La seconde a été formalisée, justifiée et publiée depuis [BUR 09].

Cet article a donc pour objectif de détailler les aspects méthodologiques qui manquent à nos travaux sur l'ASL, en s'appuyant sur les résultats expérimentaux de [ARA 07, ARA 08, ARA 09], les résultats théoriques de [DUB 08, BUR 09], et de replacer le tout dans un contexte statistique appliquée à l'informatique. Dans la deuxième partie, nous fournissons les éléments théoriques de manipulation des fonctions de croyance nécessaires à la méthode. Dans la troisième partie, nous décrivons la méthode de classification et nous résumons les résultats expérimentaux.

2. conversions entre probabilités et fonctions de croyance

Nous supposons le lecteur familier de la théorie de Dempster-Shafer et des fonctions de croyance (FC). Ainsi, nous ne reprenons pas les définitions de base, et celles-ci peuvent être trouvées dans [DEM 68, SHA 76, SME 94].

Dans de nombreuses situations, il est nécessaire de passer du formalisme probabiliste au formalisme des FC, et vice-versa. En fonction de ce que les probabilités et les FC modélisent, il y a plusieurs manières d'effectuer ces conversions [COB 03], et le choix de certaines plutôt que d'autres fait débat. Dans notre cas, nous retiendrons la conversion d'une probabilité en une fonction de croyance consonante décrite dans [DUB 08]. Soient p une fonction de probabilité sur Ω, $\{h_1, \ldots, h_N\}$ l'ensemble des éventualités de Ω ordonnés par valeurs de probabilité décroissante, et m la FC correspondant au résultat de la conversion. On a m nulle partout, sauf pour les éléments focaux du type $\{h_1, \ldots, h_k\}$, pour lesquels, on a : $m(\{h_1, \ldots, h_k\}) = k \times [p(h_k) - p(h_{k+1})]$.

A l'inverse, nous considérons la conversion d'une FC en une probabilité grâce à la transformée pignistique [SME 94]. Dans [BUR 09], nous proposons une généralisation de la transformée pignistique, dont le résultat n'est pas nécessairement une probabilité. Elle dépend d'un paramètre γ, mais dans le cas où $\gamma = 1$ nous retombons sur la transformée pignistique originale. Nous avons :

$$\mathbb{B}_\gamma(B) = m(B) + \sum_{\substack{B \subset A \subseteq \Omega \\ A \notin \Delta_\gamma}} \frac{m(A) \cdot |B|}{N(|A|, \gamma)} \quad \forall B \in \Delta_\gamma \qquad \text{avec} \qquad N(|A|, \gamma) = \sum_{k=1}^{\gamma} \frac{|A|!}{k!(|A| - k)!} \cdot k \tag{1}$$

avec $\mathbb{B}_\gamma$ désignant le résultat de la transformée de m, $|A|$ désignant le cardinal de A, et Δ_γ désignant l'ensemble des éléments de 2^Ω de cardinal inférieur ou égale à γ. L'intérêt de la transformée pignistique est classiquement de convertir une FC en une probabilité juste avant la prise de décision, de manière à prendre une décision conforme à la notion classique de pari en théorie des jeux. Dans le cas où $\gamma \neq 1$, la généralisation proposée entraîne une décision qui ne sera pas forcement focalisée sur une seule éventualité, mais sur plusieurs (au maximum γ), parmi lesquelles l'hésitation est justifiée. Par contre, le nombre d'éventualités ainsi retenu ne sera pas forcément de γ si le problème ne le justifie pas. Ainsi, on sort du cadre de la théorie des jeux, comme cela est justifié dans [BUR 09].

3. Schéma de classification à second étage optionnel

Considérons d'abord le schéma classique d'une classification basée sur des modèles génératifs. Notons $C_1, \ldots, C_K$ les K classes du problème. Pour chaque classe C_q, un modèle génératif, noté G_q, a été appris sur un échantillon représentatif. Ensuite, chaque individu I_i à classer est considéré : on calcule sa vraisemblance pour chacun des modèles G_q. Cela permet d'obtenir un ensemble de K vraisemblances $\mathcal{L}^i(G_q), \forall q \leq K$. De manière clas-

sique, I_i est classée selon la méthode du maximum a posteriori (MAP), c'est à dire qu'il est associé à la classe $C_* = \text{argmax}_{(C_q)}(\mathcal{L}^i(G_q))$.

Nous proposons de raffiner cette méthode : dans un premier temps, une étude sommaire des corpus de données doit permettre d'évaluer quelles sont les variables susceptibles d'être perdues durant la génération des modèles. Par exemple, il est classique d'avoir un premier jeu de descripteurs d'une nature donnée (de formes, par exemple) dont la prise en compte dans les modèles génératifs est minimisée par un autre jeu de descripteurs (de couleurs par exemple). Néanmoins, il est à première vue difficile de déterminer dans quelles proportions les informations issues du premier jeu de descripteurs vont être perdues. Par souci de notations simples, notons V_p l'ensemble des variables qui sont potentiellement **p**répondérantes, et V_l, l'ensemble de celles dont l'influence risque d'être **l**issée dans les modèles. Ensuite, pour chaque classe C_q, non plus un, mais deux modèles génératifs sont appris : le premier, $G_q^{p,l}$ sur l'ensemble $\{V_p, V_l\}$, et le second G_q^l, sur V_l uniquement. Le premier étage de classification consiste à calculer pour tout individu I_i les vraisemblances $\mathcal{L}^i(G_q^{p,l})$ des modèles $G_q^{p,l}$. Deux situations sont possibles :

- L'une des vraisemblances est clairement plus élevée que les autres. La décision est donc facile à prendre. Cela signifie que l'information était suffisante pour mener à bien la tâche de classification.
- Plusieurs vraisemblances ont le même ordre de grandeur, et le choix de l'une par rapport à l'autre peut sembler arbitraire. Il est donc nécessaire de vérifier si une information provenant de V_l, déterminante pour la classification, n'a pas été perdue durant la génération des modèles $G_q^{p,l}$. Dans un tel cas, il est donc judicieux de ne considérer que les quelques classes dont les vraisemblances sont suffisamment élevées, et de procéder à un second tour parmi celles-ci, en n'utilisant cette fois-ci, que V_l.

Cette stratégie, relativement simple sur le principe a le mérite de permettre de récupérer l'information manquante pour finaliser la classification, comme dans les méthodes de boosting. Par contre, à l'inverse du boosting, elle est générique par rapport au nombre de classes : Ainsi, quand une nouvelle classe (la $K+1$ième) doit être ajoutée au problème (un mot supplémentaire en reconnaissance de la parole, un nouveau visage dans un système d'identification, etc.), il suffit de fournir les deux modèles $G_{K+1}^{p,l}$, G_{K+1}^l, et cela ne remet pas en cause les modèles des autres classes. Cependant, quelle stratégie de décision adopter pour pouvoir automatiquement arbitrer entre les deux situations décrites plus haut ? En effet, il est difficile de décider automatiquement dans quels cas on peut considérer qu'une unique classe ressort des comparaisons de vraisemblance, et dans quels cas, un second étage est nécessaire.

A la fin du premier étage, plutôt qu'une stratégie du MAP, nous proposons la stratégie suivante : Nous normalisons les K vraisemblances en les divisant par la somme des vraisemblances, afin que leur somme valle 1. Ainsi, après cette normalisation, nous avons une distribution de probabilité subjective modélisant l'appartenance à chacune des classes. Ensuite, cette probabilité est convertie en une FC par la méthode indiquée dans la partie 2 [DUB 08]. Enfin, la généralisation de la transformée pignistique est utilisée [BUR 09]. Si la quantité d'information est suffisante pour qu'une seule classe ressorte, alors la décision prise à l'issue du premier étage sera la même qu'avec le MAP, et la procédure se terminera ainsi. En revanche, pour les cas limites, un sous-ensemble restreint de K' classes (avec $K' \leq \gamma$), parmi lesquelles il est difficile d'opérer une discrimination, est retenu. Dès lors, le second étage de classification est utilisé : Comparaison des K' modèles G_q^l, et prise de décision avec la stratégie du MAP.

Le dernier point à discuter est le choix de la valeur de γ. Il y a plusieurs stratégies possibles : Celle qui consiste à bien étudier le jeu de données, afin d'en connaître les spécificités, et déterminer le nombre maximal de classes parmi lesquelles une hésitation est possible (nous appelons un tel regroupement de classe, un cluster). Cette stratégie exige une connaissance précise de la topologie des données qu'il est en pratique difficile d'avoir dans certains problèmes en informatique. C'est pourquoi, une méthode plus automatique peut être préférée. En théorie, il est envisageable d'effectuer une validation croisée sur toutes les valeurs possibles de γ, et de choisir la valeur donnant le meilleur résultat. En pratique, cela n'est ni très élégant, ni robuste à l'augmentation du nombre de classes dans le problème. Dès lors, une solution intermédiaire consiste à définir de manière automatique des clusters au sein desquels une hésitation a un sens. Dès qu'une décision incomplète surgie au premier étage de classification, le second étage doit discriminer entre toutes les classes du cluster contenant l'hésitation. Ainsi, la valeur de γ n'a en pratique plus d'importance, et la valeur $\gamma = 2$ fait parfaitement l'affaire, quelque soit la taille du cluster.

C'est cette dernière solution que nous préconisons (si le problème le permet) et que nous avons utilisée pour la reconnaissance de l'ASL. Dans ce problème, la reconnaissance de chaque signe de l'ASL est effectuée au moyen d'un HMM, V_p représente les descripteurs des mouvements des mains, et V_l représente les descripteurs des mouvements des autres parties du corps (mouvements de têtes, expression faciale, etc.). Cette méthode nous a permis d'obtenir des taux de classifications meilleurs que les méthodes classiques :

- 23.7% des erreurs sont évitées par rapport au schéma de base indiqué plus haut, correspondant à la classification systématique par MAP à l'issu du premier étage.
- 31.1% des erreurs sont évitées par rapport à l'utilisation conjointe des deux modèles $G^{p,l}$ et G^l et de leur mise en cascade systématique (second étage utilisé de manière systématique). En effet, dans un tel cas, certaines décisions correctes à l'issu du premier étage sont remises en cause au niveau du second.
- 37.4% des erreurs sont évitées par rapport à l'utilisation conjointe de deux modèles G^p et G^l en parallèle (et non en cascade) et de la fusion systèmatique des vraisemblances.

4. Conclusion

Nous proposons dans cet article une méthode de classification ayant un second étage optionnel, permettant de s'affranchir de nombreux défauts de la classification supervisée par modèles génératifs. Cette méthode est basée sur une méthode de décision originale nécessitant de transformer l'information probabiliste issue des modèles de classes en une fonction de croyance. La justification de la méthode est basée sur la validation théorique de transformations mises en jeu et de son utilisation dans le cas d'un problème réel, la reconnaissance de gestes de la Langue des Signes. La suite de ce travail consiste en application de cette méthode à de nouveaux problèmes.

5. Bibliographie

[ARA 06] ARAN A., AKARUN L., Recognizing Two Handed Gestures with Generative, Discriminative and Ensemble Methods via Fisher Kernels, *Lecture Notes in Computer Science : Multimedia Content Representation, Classification and Security International Workshop, MRCS 2006*, , 2006, p. 159–166.

[ARA 07] ARAN O., BURGER T., CAPLIER A., AKARUN L., Sequential Belief-Based Fusion of Manual and Non-Manual Signs, *Gesture Workshop (GW'07)*, 2007.

[ARA 08] ARAN O., BURGER T., CAPLIER A., AKARUN L., A Belief-Based Sequential Fusion Approach for Fusing Manual and Non-Manual Signs, *Pattern Recognition*, vol. 42(5), 2008, p. 812–822.

[ARA 09] ARAN O., BURGER T., CAPLIER A., AKARUN L., Sequential Belief-Based Fusion of Manual and Non-Manual Information for Recognizing Isolated Signs, *Gesture-Based Human-Computer Interaction and Simulation, Sales Dias, M. ; Gibet, S. ; Wanderley, M.M. ; Bastos, R. (Eds.), LNCS/LNAI*, vol. 5085, 2009, Springer.

[BUR 09] BURGER T., CAPLIER A., A Generalization of the Pignistic Transform for Partial Bet, *accepted to ECSQARU'09*, July 2009.

[COB 03] COBB B. R., SHENOY P., A Comparison of Methods for Transforming Belief Functions Models to Probability Models, *Lecture Notes in Artificial Intelligence*, vol. 2711, 2003, p. 255–266.

[DEM 68] DEMPSTER A., A generalization of Bayesian inference, *Journal of the Royal Statistical Society, Series B*, vol. 30(2), 1968, p. 205–247.

[DUB 08] DUBOIS D., PRADE H., SMETS P., A definition of subjective possibility, *International Journal of Approximate Reasoning*, vol. 48(2), 2008, p. 352–364.

[DUD 01] DUDA R., HART P., STORK D., *Pattern Classification*, Wiley, 2001.

[MUL 07] MULHEM P., QUÉNOT G., BERRUT C., Recherche d'information multimédia, *Patrick Gros, éditeur, L'indexation multimédia - Descriptions et recherche automatiques*, , 2007, p. 25–49.

[RAB 89] RABINER L., A tutorial on hidden Markov models and selected applications in speech recognition, *proceedings of the IEEE*, vol. 7(2), 1989, p. 257–286.

[SHA 76] SHAFER G., *A Mathematical Theory of Evidence*, Princeton University Press, 1976.

[SME 94] SMETS P. K. R., The transferable belief model, *Artificial Intelligence*, vol. 66(2), 1994, p. 191–234.