HAL
archives-ouvertes.fr

Discrimination et Classification supervisée en référence à des prototypes

Stéphane Verdun, Véronique Cariou, El Mostafa Qannari

▶ **To cite this version:**

Stéphane Verdun, Véronique Cariou, El Mostafa Qannari. Discrimination et Classification supervisée en référence à des prototypes. 42èmes Journées de Statistique, 2010, Marseille, France, France. 2010. <inria-00494776>

HAL Id: inria-00494776
https://hal.inria.fr/inria-00494776

Submitted on 24 Jun 2010

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.

Discrimination et classification supervisée en référence à des prototypes

Stéphane Verdun & Véronique Cariou & El Mostafa Qannari

Laboratoire de Sensométrie et Chimiométrie,
Ecole Nationale Vétérinaire, Agroalimentaire et de l'Alimentation Nantes Atlantique - ONIRIS
rue de la Géraudière, BP 82225, 44322 Nantes cedex 03, France

Abstract

L'objectif de la classification supervisée est d'affecter des individus à des groupes définis a priori à partir des mesures effectuées sur des variables. Dans ce contexte, les analyses discriminantes linéaire et quadratique sont parmi les méthodes les plus populaires. Elles sont fondées sur des hypothèses de multinormalité. Dans certaines situations, cette règle s'avère inappropriée en particulier dans le cas d'un groupe multimodal ou en présence d'éléments atypiques. Pour pallier ce problème, la méthode proposée consiste à pondérer les individus de manière à déterminer des statistiques robustes. Dans une communication précédente, nous avons introduit une méthode de classification (non supervisée) basée sur la détermination d'une matrice stochastique. Un des intérêts de cette approche est d'exhiber des barycentres pondérés (appelés prototypes) au sein des différents groupes. Cette démarche est étendue au cadre de la classification supervisée et au cadre de la discrimination. Pour la classification supervisée, nous adoptons une démarche similaire à celle préconisée dans le cadre des réseaux de neurones probabilistes. Pour l'analyse factorielle discriminante, nous utilisons le système de pondération pour l'estimation des paramètres de localisation des matrices de variance covariance à l'intérieur de chaque groupe, ainsi que de la matrice de variance covariance totale.

Supervised classification is concerned with the problem of assigning individuals to one of several predefined groups from their measurements with respect to a set of variables. Linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) are particularly appropriate for the setting where the variables are assumed to have a multivariate normal distribution. However, a pitfall linked to these methods relates to the fact that each group is represented by a single centroid and the assignment rule is based on how each new individual is far removed from the centroids associated with the various groups. Obviously, this rule may not be relevant in situations where a group presents a multimodal pattern or in presence of outliers in one or several groups. We circumvent this problem by proposing a general and versatile strategy of assigning weights to the individuals. In practice, these weights are used to compute robust statistics (mean and variance-covariance matrix within each group) by down-weighting those observations which are deemed

to be abnormal in that sense that they do not follow the general pattern of the data within each group. We also show how these weights can be used within a strategy of classification pertaining to Probabilistic Neural Network (PNN) which is known to be an efficient strategy of classification based on non parametric estimations of the probability density functions and a Bayesian decision rule. The effectiveness of the methods of analysis is illustrated on the basis of real data sets.

keywords classification supervisée, classement, matrice stochastique, réseaux de neurones probabilistes.

1 Introduction

Dans le cadre classique de la discrimination où p variables quantitatives sont mesurées sur n individus qui sont eux même répartis en K groupes connus a priori, nous proposons une nouvelle démarche pour l'affectation d'un individu à un des K groupes à partir des mesures obtenues à l'aide des p variables. Cette méthode de classement s'inscrit dans le cadre général de la règle de décision Bayésienne où une nouvelle observation x est affectée au groupe pour lequel sa probabilité d'appartenance est la plus grande. Notons par h_k la probabilité donnée a-priori d'appartenir au groupe G_k. L'observation x est affectée a postériori au groupe G_k si:

$$P(G_k/x) = max_i\ P(G_i/x) \text{ for } i = 1, \ldots, K$$

où $P(G_k/x)$ correspond à la probabilité d'appartenir au groupe G_k sachant l'observation x. Notons $f(x/G_k)$ la fonction de densité de probabilité associée au groupe G_k. D'après le théorème de Bayes, la règle d'affectation revient à affecter x au groupe pour lequel le produit $h_k * f(x/G_k)$ est maximum. Le problème de classement revient donc à un problème d'estimation de densité au sein de chacun des groupes G_i. Si l'on se place dans le cadre de l'analyse discriminante linéaire et quadratique, la fonction de densité est supposée être une distribution multinormale.

Dans certaines situations, cette règle s'avère inappropriée en particulier dans le cas d'un groupe multimodal ou en présence d'éléments atypiques. Pour pallier ce problème, la méthode proposée vise à pondérer les individus de manière à produire des statistiques robustes. Ce système de poids est pris en compte pour l'élaboration d'une règle de classement sur la base d'une estimation de la fonction de densité par la méthode des noyaux à l'instar de la démarche préconisée dans le cadre des neurones probabilistes (Specht (1990)). De même, nous utilisons un système de poids pour l'estimation des paramètres de localisation et de dispersion au sein de chaque groupe permettant ainsi de conduire à une analyse factorielle discriminante pondérée.

2 Pondération des individus

Dans une communication précédente (Verdun et al. (2009)), nous avons proposé une nouvelle démarche de classification (non supervisée) basée sur une matrice stochastique. De manière pratique, la matrice stochastique est définie à partir d'une matrice de similarités entre les individus. Nous avons, en particulier, souligné qu'un des intérêts de cette méthode était de déterminer des prototypes à l'intérieur de chaque groupe. Ce sont des barycentres pondérés des individus de la classe considérée. Le système de pondération reflète le degré de centralité des individus dans la classe. Cette démarche est étendue au cadre de la classification supervisée et de l'analyse factorielle discriminante.

Par la suite, la similarité au sein de chaque groupe est évaluée à l'aide de la fonction gaussienne

$$W(x_i, x_j) = e^{-\frac{1}{2\sigma^2}||x_i - x_j||^2}$$

Ceci permet de définir à l'intérieur du groupe G_k une matrice stochastique P_k en normalisant de manière appropriée la matrice des similarités entre les individus du groupe considéré. Enfin, un système de poids associés aux individus est extrait à partir du vecteur de probabilité stationnaire de la matrice P_k.

Cette démarche peut être justifiée à l'aide de la théorie des graphes. Les individus du groupe G_k sont alors assimilés aux sommets d'un graphe. Une marche aléatoire dont la matrice de passage est la matrice stochastique P_k peut être définie. Les coefficients p_{ij} correspondent ainsi à la probabilité que la marche aléatoire atteigne le sommet j au pas suivant en partant du sommet i. Le vecteur de probabilité stationnaire $w^{(k)}$, c'est à dire le vecteur propre à gauche associé à la valeur propre 1, donne la loi asymptotique de la marche aléatoire. De manière plus concrète, la probabilité associée à l'individu i (sommet i) reflète le degré de centralité de cet individu dans le groupe auquel il appartient.

3 Règle de classement et Représentation factorielle

3.1 Méthode de classement par prototypes

La méthode de classement en référence à des prototypes consiste à affecter une observation x au groupe G_k si

$$\sum_{i \in G_k} w_i^{(k)} W(x_i, x) = \max_l \sum_{i \in G_l} w_i^{(l)} W(x_i, x)$$

où $w_i^{(l)}$ est le poids associé à l'individu x_i du groupe G_l. Cette règle s'apparente à celle proposée dans le cadre des réseaux de neurones probabilistes pondérés (Ramakrishnan et Selvan (2006), Ramakrishnan et Elmary (2009)).

3.2 Utilisation de la pondération dans l'AFD

La pondération peut aussi être utilisée dans un but descriptif dans le cadre de l'analyse factorielle discriminante. Il est ainsi possible de calculer les axes factoriels discriminants issus de l'analyse factorielle discriminante des données en tenant compte des pondérations. En effet, les axes sont les vecteurs propres de la matrice $T^{(-1)}B$, où la matrice T est la matrice de variance covariance totale et B la matrice de variance covariance interclasses. Si le calcul de ces deux matrices tient compte des pondérations des individus, cela permet de limiter l'influence d'individus atypiques au sein de leur classe dans la détermination des axes.

4 Application

La méthode de classement a été appliquée sur des données de spectrométrie. Des spectres proche infra-rouge de pommes ont été mesurés dans l'objectif de prédire la maturité des pommes. Les données comportent 550 variables (différentes longueurs d'ondes) mesurées sur 1066 individus (pommes) dont l'appartenance à l'un des 6 groupes (la maturité) est connue. Une analyse en composantes principales a été effectuée sur ce tableau de données et les 20 premières composantes principales ont été retenues pour la suite de l'étude. Les individus ont été séparés en différents échantillons d'étalonnage (70% des individus) et de validation. Les méthodes de classement par prototypes (DS), par réseaux de neurones (PNN) et par analyse discriminante linéaire (ADL) ont été appliquées. Pour les deux premières, la valeur du paramètre σ a été obtenue par validation croisée. Les résultats correspondant à une simulation consistant à diviser $N = 10$ fois les individus en un ensemble d'étalonnage et de validation sont représentés dans la Figure 1. Il ressort que les méthodes de classement par prototypes et des réseaux de neurones présentent une meilleure performance que l'ADL. Par ailleurs, la méthode de classement par prototypes présentent un léger avantage par rapport aux réseaux de neurones probabilistes. Il est cependant intéressant de noter que la méthode de classement par prototypes s'avère moins sensible que les réseaux de neurones probabilistes à l'introduction de bruit (généré par permutation des étiquettes de classes pour certains individus) dans les données. La prise en compte d'un système de poids permet ainsi d'améliorer la robustesse de la règle de classement.

5 Conclusions et perspectives

Une méthode visant à pondérer les individus a été présentée. D'une part, ces pondérations permettent de mieux cerner les individus centraux d'un groupe. D'autre part, ces poids peuvent être utilisés dans diverses applications telles que de la discrimination (méthode de classement par prototypes) ou lors du calcul des axes de l'analyse factorielle discrim-

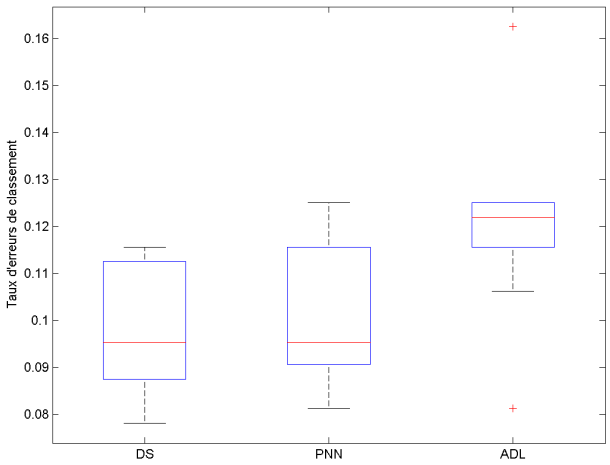

Figure 1: Taux d'erreurs de classement (échantillons de validation)

inante (AFD). De manière plus générale, nous avons proposé une démarche allant de la classification non supervisée (Verdun et al. (2009)) à l'analyse factorielle discriminante. Le principe est de déterminer une matrice stochastique reflétant la proximité des points ou leur densité et d'en déduire un système de poids associés aux individus au sein de chacune des classes. Ce système de poids traduit le degré de centralité des individus et peut être utilisé dans un cadre descriptif (classification non supervisée, AFD) ou prédictif (classification supervisée).

Étant donné le caractère général de la démarche de pondération, il est tout à fait possible de l'adapter à d'autres contextes tels que la régression linéaire ou l'analyse en composantes principales, par exemple.

Une autre direction d'investigation serait de considérer plusieurs prototypes par groupe plutôt qu'un prototype unique, fut-il pondéré. Ceci pourrait être plus approprié à des situations où les groupes seraient constitués de plusieurs sous-groupes différents.

Bibliographie

[1] Ramakrishnan, S. and El Emary, I.M.M. (2009) Comparative study between traditional and modified probabilistic neural networks. *Telecommunication Systems*, 40(1), 67–74.

[2] Ramakrishnan, S. and Selvan, S. (2006) Classification of Brain Tissues Using Multiwavelet Transformation and Probabilistic Neural Network. *International Journal of Simulation: Systems, Science and Technology*, 7(9), 9–25.

[3] Specht, D.F. (1990) Probabilistic neural networks. *Neural Networks*, 3, 109–118.

[4] Verdun, S. et Cariou, V. et Qannari, E.M. (2009) Classification en référence à une matrice stochastique, *41èmes Journées de Statistique, SFdS*, Bordeaux, 25-29 mai 2009.