Vers une classification non supervisée basée sur un nouvel indice de connectivité

Frédéric BLANCHARD, Michel HERBIN, Philippe VAUTROT

CReSTIC-LERI
IUT, rue des Crayères, BP 1035,
F-51687 Reims Cedex 2, France
frederic.blanchard@univ-reims.fr,
michel.herbin@univ-reims.fr,
philippe.vautrot@univ-reims.fr

Résumé – Cet article présente une nouvelle approche pour la classification non-supervisée de données. Cette nouvelle méthode repose sur la construction d'un indice de connectivité et ne fait aucune hypothèse sur la forme des classes ni sur leurs effectifs. L'approche présentée constitue une alternative aux méthodes basées sur l'estimation de densité de probabilité.

Abstract – This paper expose a new approach for data clustering. This new method leads on the definition of a new connectivity index. This approach is an alternative technique of density estimation based methods.

1 Introduction

La classification non supervisée est un domaine important de l'analyse exploratoire des données. Les développements des méthodes de classification doivent beaucoup à l'évolution des capacités informatiques. L'objectif général de la classification est de déterminer une partition des données (ou une répartition des données en groupements) de sorte que deux données sont soit regroupées, si elles sont très semblables, soit séparées si elles sont assez différentes. Dans l'approche non supervisée, l'interprétation dépend naturellement du domaine d'application, et ces domaines sont nombreux : biologie, médecine, reconnaissance de formes etc... On peut distinguer deux types de méthodes de classification : les méthodes hiérarchiques et les méthodes de partitionnement.[CDG+89]
Parmi les méthodes de partitionnement, de nombreuses techniques reposent sur l'estimation de densité de probabilité (par exemple [CM99][EKSX96][HBV96]). Dans [HBV96], Herbin et al. proposent une méthode proche de celle présentée par Hinneburg et Keim. dans [HK98]. Comme toutes les méthodes basées sur ce principe, on y suppose que les classes sont définies à l'aide des zones ayant localement une densité qui présente un maximum. Les maxima locaux de l'estimation de la fonction de densité de probabilité (ou *modes*) jouent un role similaire aux *centres* dans les algorithmes de type "centres mobiles" mais contrairement à ce genre de méthodes, aucune hypothèse sur les formes des classes n'est émise. Malheureusement, en dépit de leur efficacité globale, il existe des situations dans lesquelles ces méthodes échouent. En effet, l'estimation de la densité de probabilité requiert l'utilisation d'un paramètre de lissage. Si ce paramètre de lissage (fenêtre de lissage ou *bandwidth*) est trop petit, l'estimation de la densité de probabilité est bruitée, particulièrement dans les queues de distribution [Sil86]. Ceci conduit à une surestimation du nombre de modes de la densité [HBV01], tandis qu'un lissage trop fort masquera certaines classes. La nécessité d'utiliser une alternative à cette fonction de densité de probabilité nous a donc amené à construire un indice de connectivité paliant les faiblesses sus-citées, tout en conservant l'avantage de non-hypothèse sur les formes des classes.
Notre indice de connectivité est une nouvelle fonction dont l'"esprit" est proche de la fonction de densité de probabilité. Le principe est le suivant : après avoir déterminé les k plus proches voisins de chaque donnée, on attribue à chaque individu de l'échantillon, le nombre de données pour lesquelles il est un des k plus proches voisins. Autrement dit, les $k - ppv$ déterminent des voisinages pour chaque donnée de l'échantillon, et l'indice de connectivité d'une données est le nombre de voisinages auxquels elle appartient. On peut alors utiliser cette indice dans un processus de classification, à l'instar de la fonction de densité de probabilité. Cette nouvelle fonction est plus robuste (peu sensible à des points isolés très éloignés) mais ne néglige pas les classes de faible densité. Elle présente donc un grand intéret dans les situations ou l'estimation de la densité conduit à des résultats décevants.
Dans la section 2 nous présenterons dans le détail la méthode de construction de l'indice de connectivité. Nous présentons ensuite, dans la partie 3, l'estimation de la densité de probabilité dans un contexte de classification. Dans la partie 4, nous comparons l'indice de connectivité avec la de densité de probabilité. Enfin, dans la section 5, une discussion est proposée avant de conclure.

2 Indice de connectivité

Définissons maintenant l'indice de connectivité. Nos données sont dans un espace métrique F^p de dimension p. On défi -

nit le voisinage d'une donnée x_i de l'échantillon $\Omega = \{x_1, x_2, ..., x_n\}$ dans F^p comme l'ensemble des k plus proches voisins dans Ω de x_i, en utilisant une distance quelconque dans F^p. On le note $V_k(x_j)$. On défi nit alors l'indice de connectivité d'une donnée dans l'échantillon comme une fonction "quantifi ant" son appartenance aux voisinnages des autres données de l'échantillon. L'indice de connectivité d'une donnée dans l'échantillon est d'autant plus élevé que cette données appartient à un nombre important de voisinnages des autres données. Son calcul en une donnée $x_i \in \Omega$ dépend donc directement de l'appartenance de x_i aux voisinnages des x_j $(j = 1..n)$.

On défi nit donc l'indice de connectivité dans Ω, d'une donnée x_i de Ω comme :

$$IC(x_i) = \sum_{j=1}^{n} M(x_i, V_k(x_j)) \quad (1)$$

où $M(x_i, V_k(x_j))$ est une fonction d'appartenance à $V_k(x_j)$ évaluée au point x_i. Dans le cadre de ce travail, on choisit M de sorte que : $M(x_i, V_k(x_j))$ vaut 1 si $x_i \in V_k(x_j)$ et 0 sinon. Autrement dit, la connectivité d'une donnée x_i correspond dans ce cas au nombre de données auxquelles x_i est connectée.

La fonction ainsi construite nous permet de quantifi er la liaison d'une donnée dans l'échantillon et révéler ainsi les structures de classes de cet échantillon. Cette notion se substitue alors à celle de densité de probabilité.

3 Estimation de la densité de probabilité

Rappelons tout d'abord le principe de l'estimation de densité de probabilité par la méthode de Parzen. Considérons un échantillon $\{x_i\}_{i=1..n}$ de taille n, d'un espace F^p de dimension p (i.e. $\forall i, x_i = (x_i^1, .., x_i^p)$). L'estimation de la densité de probabilité, selon la méthode de Parzen, au point $x \in F^d$ est donné par :

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{h^p} K \left(\frac{x - X_i}{h} \right) \right) \quad (2)$$

où K est la fonction noyau et h le paramètre de lissage. Dans un contexte de classifi cation peut par exemple utiliser un noyau gaussien :

$$K \left(\frac{x - X_i}{h} \right) = (2\pi)^{-p/2} \, exp \left(-\frac{1}{2} \left(\frac{d(x, X_i)}{h} \right)^2 \right) \quad (3)$$

où (d est une distance sur F^p.
L'observation de l'estimation de densité de probabilité et la détection de ses modes (maxima locaux) permet ensuite de déterminer des classes dans l'échantillon de données. On peut par exemple utiliser des algorithmes de "lignes de partage des eaux"[HBV01] ou des méthodes de "mean shift"[CM99].
Les principaux inconvénients de la fonction de densité de probabilité utilisée en classifi cation proviennent du paramètre de lissage et de la dimensionnalité du problème. En effet, si l'on choisit un paramètre de lissage trop petit, la fonction obtenue contient trop de maxima locaux et donc de modes. A l'inverse un paramètre de lissage trop fort risque de masquer des classes de faible densité. Par ailleurs, lorsque le nombre de données est faible (sous-echantillonnage) il est nécessaire d'effectuer un lissage plus large. Et lorsque la dimensionnalité du problème augmente, le phénomène d'espace creux ("curse of dimensionality") conduit à devoir effectuer un lissage plus important déteriorant ainsi la qualité de l'estimation et modifi ant la forme des classes obtenues. Ces éléments sont ilustrés par les Figures 3 et 4 (calculées à partir de l'échantillon de la Figure 1).

4 Densité de probabilité Vs. Indice de connectivité

La valeur de la densité n'est pas informative sur l'appartenance d'une donnée à une classe. Comme on peut le voir sur la Figure 3, seules les variations locales de densité permettent d'informer sur l'appartenance de cette donnée à une classe. Or l'estimation de la densité est généralement bruitée -parfois fortement- [Sil86], il est donc diffi cile d'extraire une information pertinente sur les variations locales de densité (Figures 3 et 4). La plupart des auteurs utilisent alors des estimations de la densité à différentes échelles (i.e. différents lissages) pour essayer de classer les données.
Notre critère de connexité est directement informatif sur l'appartenance d'une donnée à une classe. Sur la Figure 2, on constate en effet que notre indice fait apparaitre parfaitement la troisième classe, de densité beaucoup plus faible que les deux autres, contrairement à la densité de probabilité (Figures 3 et 4). Cet indice nous parait donc plus informatif que la densité pour classer les données. La classifi cation utilisant cet indice conserve les avantages des méthodes basées sur la densité (pas d'hypothèse sur la forme et le nombre des classes) mais il permet de plus de différencier sans diffi culté des classes de faible densité et/ou de forte densité.

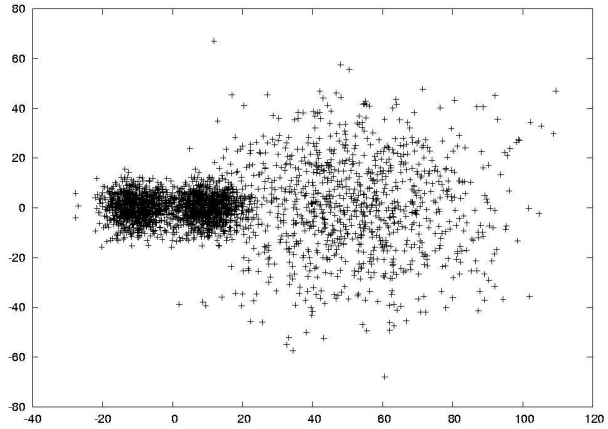

FIG. 1 – Echantillon de données initial (3 ensembles de 1000 données chacun)

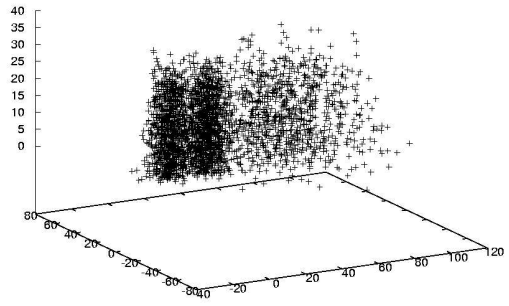

FIG. 2 – Indice de connectivité, révèlant la présence des classes

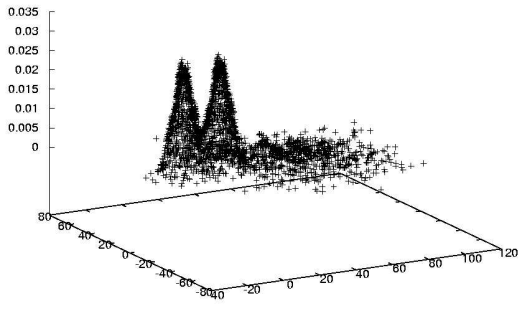

FIG. 3 – Estimation de la densité de probabilité masquant par du bruit les classes de faible densité

5 Discussion et conclusion

Notre indice de connectivité présente de nombreux avantages sur la fonction de densité de probabilité. Il permet un meilleure prise en compte des classes ayant des densités différentes dans un échantillon de données et palie les faiblesses de la fonction de densité qui masque les classes de faible densité. Sa mise en place est aisée et son calcul peu couteux.
Cependant notre fonction n'est pas exempte d'inconvénients. Tout d'abord, le choix du paramètre k lors de la détermination des $k - ppv$ peut augmenter les temps de calcul, tout comme l'estimation du k optimal $\hat{k}$ par diverse techniques (maximisation de la variance interclasse, bootstrap etc...). Par ailleurs l'efficacité d'une méthode utilisant l'indice de connectivité repose naturellement sur le choix de l'algorithme de clustering utilisé. Nos travaux et développements actuels consistent à concevoir un nouvel algorithme de classification basé sur notre indice et utilisant des représentations floues afin d'obtenir un outil plus flexible et plus efficace.

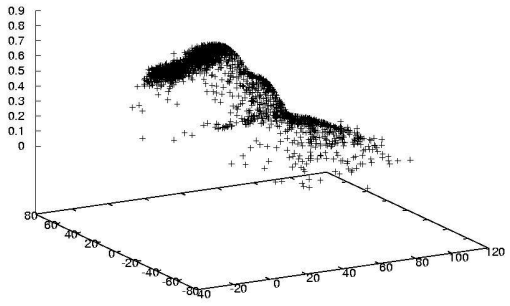

FIG. 4 – Lissage de la densité de probabilité entrainant l'impossibilité de faire apparaitre les classes

Références

[CDG+89] G. Celeux, E. Diday, G. Govaert, Y. Lechevallier, and H. Ralambondrainy. *Classification automatique des données.* Paris, 1989.

[CM99] D. Comaniciu and P. Meer. Mean shift analysis and applications. In *IEEE Int. Conf. Computer Vision(ICCV'99)*, pages 1197–1203, Kerkyra, Greece, 1999.

[EKSX96] M. Ester, H. P. Kriegel, J. Sander, and X. Xiaowei. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. 2nd Int. Conf. on Knwoledge Discovery and Data Mining (KDD 96)*, pages 226–231, Portland, 1996. AAAI Press.

[HBV96] M. Herbin, N. Bonnet, and P. Vautrot. A clustering method based on the estimation of the probability density function and on the skeleton by influence zones, application to image processing. *Pattern Recognition Letters*, (17) :1141–1150, 1996.

[HBV01] M. Herbin, N. Bonnet, and P. Vautrot. Estimation of the number of clusters and influence zones. *Pattern Recognition Letters*, (22) :1557–1568, 2001.

[HK98] A. Hinneburg and D. A. Keim. An efficient approach to clustering in multimedia databases with noise. In *Proc. 4th Int. Conf. on Knowledge Discovery and Data Mining (KDD98)*, pages 58–65, New York, 1998. AAAI Press.

[Kow95] F. Kowalewski. A gradient procedure for determining clusters of relatively high point density. *Pattern Recognition*, (28) :1973–1984, 1995.

[Sil86] B. W. Silverman. *Density Estimation for Statistics and Data Analysis.* Monographs on Statistics and Applied Probability, 1986.