Pondération locale des variables en apprentissage numérique non-supervisé [1]

Nistor Grozavu, Younès Bennani, Mustapha Lebbah

LIPN - CNRS UMR 7030 - Université Paris 13
99, avenue J-B. Clément, 93430 Villetaneuse

{Prenom.Nom}@lipn.univ-paris13.fr

Résumé. Dans cet article, nous proposons une nouvelle approche de pondérations des variables durant un processus d'apprentissage non supervisé. Cette méthode se base sur l'algorithme « batch » des cartes auto-organisatrices. L'estimation des coefficients de pondération se fait en parallèle avec la classification automatique. Ces pondérations sont locales et associées à chaque référent de la carte auto-organisatrice. Elles reflètent l'importance locale de chaque variable pour la classification. Les pondérations locales sont utilisées pour la segmentation de la carte topologique permettant ainsi un découpage plus riche tenant compte des pertinences des variables. Les résultats de l'évaluation montrent que l'approche proposée, comparée à d'autres méthodes de classification, offre une segmentation plus fine de la carte et de meilleure qualité.

1 Introduction

La taille des données peut être mesurée selon deux dimensions, le nombre de variables et le nombre d'observations. Ces deux dimensions peuvent prendre des valeurs très élevées, ce qui peut poser un problème lors de l'exploration et l'analyse de ces données. Pour cela, il est fondamental de mettre en place des outils de traitement de données permettant une meilleure compréhension des données. La réduction des dimensions est l'une des plus vieilles approches permettant d'apporter des éléments de réponse à ce problème. Les méthodes qui nous intéressent dans ce papier sont celles qui permettent de faire à la fois de la réduction de dimension et la classification non supervisée de données en utilisant les cartes auto-organisatrices (SOM : Self-organizing Map). Celles-ci sont souvent utilisées parce qu'elles sont considérées à la fois comme outils de visualisation et de partitionnement non supervisé de différents types de données. Elles permettent de projeter les données sur des espaces discrets qui sont généralement en deux dimensions. Plusieurs extensions des cartes auto-organisées ont été dérivées du premier modèle original proposé par Kohonen (Kohonen,

[1] Ce travail a été réalisé dans le cadre du projet Infom@gic du Pôle de Compétitivité Cap Digital (Image, Multimedia and Vie numérique).

1989) et introduites en littérature : *μ-SOM* (Guérif et al., 2005), *gr-SOM* (Kaburlasos et Papadakis, 2004), *w-SOM* (Guérif et Bennani, 2007), *BeSOM* (Lebbah et al., 2007). Ces modèles sont différents les uns des autres, mais partagent la même idée de présenter les données de différents types de grande dimension en une simple carte à deux dimensions. Un intérêt majeur sera donné à l'algorithme *w*-SOM qui est une extension de l'algorithme *w-k*-moyennes. Ce modèle permet en même temps de construire une carte topologique et d'estimer des pondérations globales de chacune des variables constituant la base d'apprentissage.

La pondération de variable consiste à associer des valeurs numériques (poids de pondération) à chaque variable. Elle permet de nous donner une information sur l'importance de la variable. Ainsi une variable possédant une pondération forte explicite le fait qu'elle est pertinente et qu'elle a participé activement au processus de classification.

Dans ce papier, nous proposons une méthode de pondération locale des variables basée sur l'approche w-SOM (Guérif et Bennani, 2007). Ces pondérations seront utilisées pour la segmentation de la carte topologique. En effet, contrairement à la méthode de pondération globale w-SOM qui estime un seul vecteur de pondérations pour tout l'ensemble des référents, la pondération locale associe un vecteur de pondérations des variables à chaque référent de la carte. Par conséquent nous pouvons utiliser ces pondérations pour regrouper les prototypes qui ont les mêmes variations du vecteur de pondérations.

La suite de cet article est organisée comme suit : nous présentons notre approche de pondération locale des variables dans la section 2, après l'introduction de l'algorithme w-SOM. Dans la section 3, nous présentons les différents résultats obtenus. Finalement on termine par la conclusion et les perspectives de la méthode proposée.

2 Pondération des variables

L'algorithme des w-k-moyennes proposé par Huang et al. (2005) utilise une pondération globale et il a été étendu aux cartes auto-organisatrices par Guérif et Bennani (2007). Dans cet article, nous proposons de remplacer la pondération globale utilisée par l'algorithme w-SOM par une pondération locale. Dans notre version locale de l'algorithme w-SOM, un vecteur de pondérations est associé à chaque référent. Ce vecteur est optimisé et estimé pendant le processus d'apprentissage.

Dans la version globale de w-SOM, le poids w_j associé à la *j*-eme variable est le même pour toutes les autres *j*-eme variables associées à l'ensemble des référents *Z* de la carte. En étendant ce modèle au cas d'une pondération locale, chaque référent *k* de la carte a alors son propre vecteur de poids w_k. En Notant w_{jk} le poids de la *j*-ème variable pour le référent *k*, la fonction de coût s'écrit de la manière suivante :

$$P(U,Z,W)=\sum_{i=1}^{N}\sum_{j=1}^{n}\sum_{k=1}^{C}u_{ik}\,w_{jk}^{\beta}\sum_{l=1}^{C}h_{kl}\left(x_{ij}-z_{lj}\right)^2 \qquad (1)$$

$$\text{contraintes} : \begin{cases} \sum_{i=1}^{C} u_{ik} = 1, 1 \le i \le N \\ u_{ik} \in \{0;1\}, 1 \le i \le N, 1 \le k \le C \\ \sum_{j=1}^{n} w_{jk} = 1, w_{jk} \in [0;1], 1 \le k \le C \end{cases}$$

où *C* est le nombre de référents, *N* le nombre d'exemples, et *n* la dimension de l'espace.

En notant *U* la matrice de partitionnement ou d'affectation des exemples *x* sur la carte topologique, l'optimisation de $P(U,Z,W)$, s'effectue en itérant l'optimisation suivante :

1. Optimiser $P(U,\hat{Z},\hat{W})$ en fixant Z et W ; chaque individu *x* est affecté au référent dont il est le plus proche au sens de la distance euclidienne pondérée :

$$u_{ik} = \begin{cases} 1, si\ k = \arg \min_{1 \le l \le C} \sum_{j=1}^{n} w_{jl}^{\beta} \left(x_{ij} - \hat{z}_{lj} \right)^2 \\ 0, \text{ sinon} \end{cases}$$

2. Optimiser $P(\hat{U},Z,\hat{W})$ en fixant *U* et *W* ; chacun des référents est remplacé par le barycentre des individus qui lui sont affectés et de ceux qui sont affectés à ces voisins déterminés par la fonction de voisinage *h*:

$$z_k = \frac{1}{\sum_{i=1}^{N} \hat{u}_{ik}} \times \sum_{i=1}^{N} \hat{u}_{ik} x_i h_{ik} \qquad (2)$$

3. Optimiser $P(\hat{U},\hat{Z},W)$ en fixant *U* et *Z* ; on utilise l'approche analytique proposée par Huang et al. (2005) en modifiant la définition de D_j. Dans le cas d'une pondération locale, chaque référent *k* a ses propres coefficients de pondérations w_{jk} et on remplace D_j par D_{jk}. Ainsi, on obtient alors la fonction de coût suivante :

$$P(\hat{U},\hat{Z},W) = \sum_{j=1}^{n} \sum_{k=1}^{C} w_{jk}^{\beta} \sum_{i=1}^{N} u_{ik} \sum_{l=1}^{C} h_{kl} \left(x_{lj} - z_{lj} \right)^2 \qquad (3)$$

$$= P(\hat{U},\hat{Z},W) = \sum_{j=1}^{n} \sum_{k=1}^{C} w_{jk}^{\beta} D_{jk}$$

où :

$$D_{jk} = \sum_{i=1}^{N} u_{ik} \sum_{l=1}^{C} h_{kl} \left(x_{ij} - z_{lj} \right)^2$$

Le poids w_{jk} de la variable *j* pour le référent *k* est défini de la manière suivante :

$$w_{jk} = \begin{cases} 0, \ \ si D_{jk} = 0 \\ \left(\sum_t \left[\frac{D_{jk}}{D_{jt}} \right] \right)^{\frac{1}{\beta - 1}} sinon \end{cases} \qquad (4)$$

Ainsi nous obtenons un vecteur des pondérations des variables pour tous les sous ensembles associés aux référents de la carte. La pertinence des variables dépend de ces pondérations, ainsi, si elles sont plus petites, on pourra les éliminer. Par conséquent, nous pourrons utiliser ces pondérations pour regrouper les référents qui ont des vecteurs de pondération les plus proches.

Algorithme w-SOM (version locale) :lw-SOM

Initialiser les vecteurs de pondération W et l'ensemble des référents de la carte Z aléatoirement.

Pour t = 1,…T_{max} **faire** // T_{max} indique le nombre d'itérations

-Optimiser $P(U,\hat{Z},\hat{W})$: à chaque référent, on lui affecte l'individu le plus proche en utilisant la distance pondérée d_w. Où $d_w = \sum_{j=1}^{n} w_{jl}^{\beta} \left(x_{ij} - \hat{z}_{lj} \right)^2$;

-Optimiser $P(\hat{U},Z,\hat{W})$: les référents sont remplacés par la moyenne pondérée des individus affectés à chaque référent et ceux des référents voisins déterminées par la fonction de voisinage. (Formule 2)

-Optimiser $R(\hat{U},\hat{Z},W)$: on estime les vecteurs de pondérations affectés à chaque référent de la carte. (Formule 4)

Fin de boucle

Généralement l'utilisation des cartes topologiques est suivie d'une segmentation des référents de la carte. Souvent ces méthodes de segmentation se résument à l'utilisation de l'algorithme de classification hiérarchique ou des K-moyennes combinés avec un indice de qualité pour déterminer la taille de la partition idéale de la carte. Ainsi dans ce qui suit nous proposons une application de lw-SOM consistant à utiliser ces pondérations locales, pour segmenter la carte en utilisant l'algorithme K-moyennes combiné avec l'indice de qualité interne de Davies-Bouldin, (Vesanto et Alhoniemi (2000)).

3 Résultats expérimentaux

3.1 Jeux de données

Nous avons utilisé différents jeux de données de taille et de complexité variable pour évaluer notre approche et nous présentons ici les résultats obtenus sur deux d'entre eux : le premier est le jeu de données Iris introduit par R.A. Fisher en 1988, le second est mis à la disposition de la communauté par l'Université d'Irvine (Breiman et al., 1984) :

- Le jeu de données d'Iris contient 3 classes de 50 individus chacun où chaque classe représente un type donné d'une fleur Iris. Une classe est linéairement séparable des deux autres qui se recouvrent fortement. La base Iris est composée de 4 attributs qui décrivent : la longueur, la largeur du sépale et respectivement du pétale. Les classes sont : Iris Setosa, Iris Versicolor, Iris Virginica.
- La base waveform est composée de 5000 individus divisés en 3 classes. La base originale comportait 21 variables, mais 19 variables additionnelles distribuées selon une loi normale ont été rajoutées sous forme de bruit. Chaque individu a été généré comme une combinaison de 2 sur 3 vagues.

3.2 Segmentation de la carte topologique

L'algorithme lw-SOM décrit dans la section précédente permet d'obtenir d'une part, une projection en deux dimensions des données et d'autre part, une pondération des variables spécifiques à chaque région de l'espace. Vesanto et Alhoniemi (2000) ont proposé de segmenter une carte topologique en combinant l'algorithme des k-moyennes à l'indice de Davies-Bouldin qui permet de déterminer automatiquement la taille de la partition après segmentation. Nous avons appliqué cette approche sur les référents et sur les pondérations.

En utilisant le jeu de données Iris, nous avons évalué le découpage par l'approche classique et en utilisant les k-moyennes sur les vecteurs de pondérations. La figure 1 indique une segmentation de la carte 6×4 en deux sous ensembles, utilisant seulement les référents de la carte. La figure 2 indique une segmentation de la même carte en utilisant dans ce cas les vecteurs de pondérations qui prennent en considération l'importance locale de chacune des

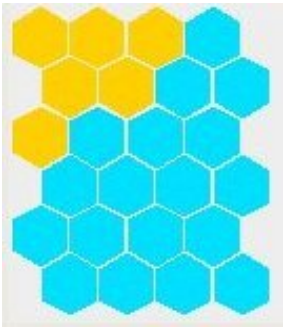

FIG. 1 - *Segmentation de la carte 6x4 (base IRIS) avec K-moyenne (approche classique – utilisation des référents). La valeur de l'indice de Davies Bouldin = 0.0776*

variables fournie par le vecteur *w* de dimension *n*. On observe clairement, sur la figure 2, que cette segmentation fournit trois sous ensembles correspondant aux trois classes de la base Iris. Cette amélioration est confirmée par la diminution de l'indice de qualité (Davies Bouldin) qui passe de *0.0776* à 0.0556

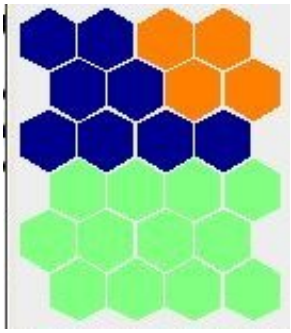

FIG. 2 – *Segmentation de la carte 6x4 avec k-moyennes utilisant les vecteurs de pondération. La valeur de l'indice de Davies Bouldin = 0.0556*

Nous allons maintenant comparer les différents découpages de la carte à l'aide des k-moyennes obtenues sur le jeu de données *waveform*. Les figures 3 et 4 indiquent le résultat de la segmentation après apprentissage de la carte topologique avec la méthode lw-SOM.

La figure 3 indique une segmentation de la carte utilisant l'approche classique sur les référents.

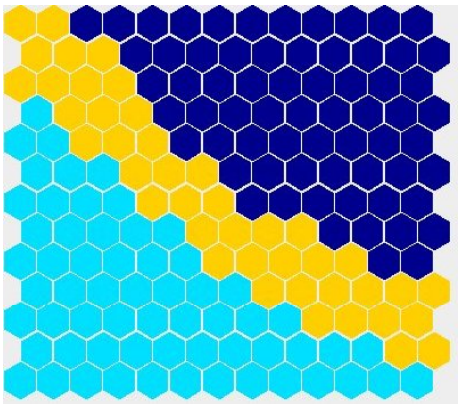

FIG. 3 – *Segmentation de la carte (base waveform) en appliquant, les k-moyennes sur les vecteurs référents. Valeur de l'indice de Davies Bouldin = 0,49.*

Celle-ci permet d'obtenir une segmentation de la carte en 3 sous ensembles avec un indice de qualité égale *0.49*. La figure 4 indique une segmentation de la carte en utilisant les k-

moyennes avec les pondérations locales. Celle-ci permet d'améliorer la qualité de la partition en retrouvant une partition contenant 3 sous ensembles avec un indice de qualité égale *0.007*.

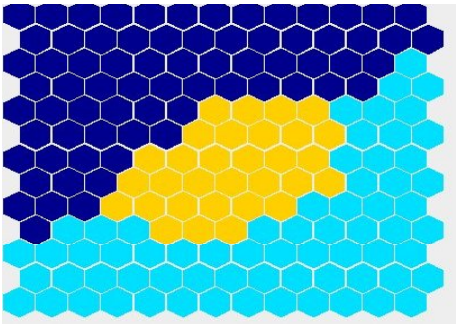

FIG. 4 - *Segmentation de la carte avec lw-SOM, en appliquant les k-moyennes sur les pondérations des neurones. Valeur de l'indice de Davies Bouldin = 0,007.*

En observant les résultats obtenus avec les deux bases, il est clair que l'utilisation des pondérations des variables locales permet de mieux segmenter la carte topologique.

Discussion

Les figures 5 et 6 représentent les différentes pondérations des variables associées aux référents de la carte topologique obtenue avec lw-SOM, sous forme de signal. En observant les cartes, il est clair que visuellement, on peut segmenter la carte par rapport aux différentes pondérations en regroupant les référents qui ont des pondérations proches. Contrairement à w-SOM globale, l'algorithme lw-SOM, permet de caractériser chaque sous ensemble associé à un référent de la carte, par un vecteur de pondérations indiquant la pertinence de chacune des variables. Par conséquent, les variables qui ont des pondérations proches permettent de distinguer les référents proches, ainsi d'obtenir la segmentation.

En observant la figure 5 représentant les pondérations locales de la base des Iris, on peut voir un sous ensemble de prototype sur le coin haut à gauche de la carte qui est caractérisé par les deuxièmes et quatrièmes variables associées à des pondération très forte. En observant aussi la figure 6 représentant les pondérations locales de la base *waveform*, on peut voir trois sous ensembles de référent avec des pondérations variables. On rappelle que les variables du bruit sont représentées dans la partie droite du signal. On distingue par exemple que dans le coin droit en bas de la carte ces pondérations représentent plus le bruit que les variables.

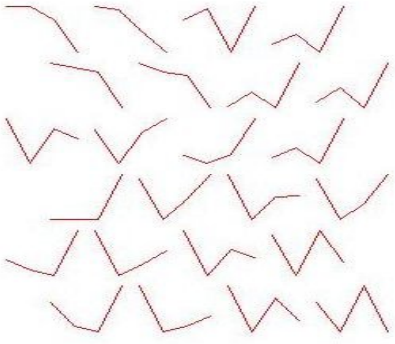

FIG. 5 – *Les pondérations des variables de la base Iris*

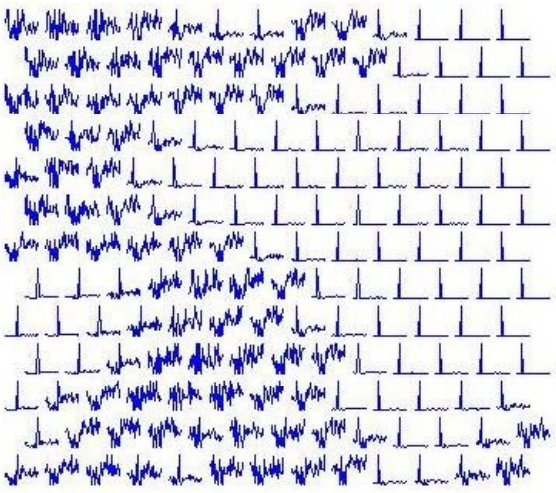

FIG. 6 – *Les pondérations de variables de la base Waveform*

4 Conclusions

Dans cette étude, nous avons introduit une nouvelle méthode de pondération des variables. L'algorithme lw-SOM propose d'apprendre les vecteurs de pondération associés à chaque référent durant le processus d'apprentissage en se basant sur l'algorithme batch de Kohonen. Contrairement à w-SOM, notre approche permet de caractériser chaque sous ensemble « cluster » par les variables les plus pertinentes. En obtenant un vecteur des pondérations pour chaque référent de la carte, on peut regrouper ceux qui ont des pondérations

similaires et éliminer ainsi les autres. Nous obtenons donc un découpage de la carte plus fin en comparaison aux autres algorithmes classiques. Comme perspectives, nous envisageons poursuivre ce travail pour un objectif de sélection de variables et d'étendre ce modèle aux variables binaires.

5 Références

[1] Blansche A., Gancarski P., Korczak J.J.(2006). *MACLAW: A modular approach for clustering with local attribute weighting*, Pattern Recognition Letters, 27 (11), 1299-1306.

[2] Blum L. A., Langley P. (1997). *Selection of relevant features and examples in machine learning*, in: Elsevier Science B.V.

[3] Breiman L., Friedman J.H., Olshen R.A., Stone C.J. (1984). *Classification and Regression Trees*, Wadsworth International Group: Belmont, California, 43-49.

[4] Chan E., Ching W., Ng M. et Huang Z. (2004). *An Optimization Algorithm for Clustering Using Weighted Dissimilarity Measures Optimization*, Pattern Recognition, V37, 943-952.

[5] Guérif S., Bennani Y. (2007). *Dimensionality Reduction Through Unsupervised Features Selection*, EANN'07, International Conference on Engineering Applications of Neural Networks, Thessaloniki, Hellas.

[6] Guérif S., Bennani Y. and Janvier E. (2005). *µ-SOM Weighting features during clustering*, Proceedings of the 5th Workshop On Self-Organizing Maps (WSOM'05), 397-404.

[7] Huang J. Z., Ng M. K., Rong H., Li Z. (2005). *Automated Variable Weighting in k-Means Type Clustering*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27(5): 657-668.

[8] Kaburlasos V.G., Papadakis S.E. (2004). *grSOM: a granular extension of the self-organizing map for structure identification applications*, Fuzzy Systems, Proceedings. 2004 IEEE, Volume 2, 789-794.

[9] Kohonen, T. (1989), *Self-Organization and associative memory*, Heidelberg: Springer-Verlag, Berlin, 3rd edition.

[10] Law H. C. (2006). *Clustering, Dimensionality Reduction, and Side Information*, PhD Thesis, Michigan State University - Department of Computer Science and Engineering.

[11] Lebbah M., Rogovschi N. and Bennani Y. (2007). *BeSOM : Bernoulli on Self Organizing Map*, International Joint Conferences on Neural Networks. IJCNN 2007, Orlando, Florida.

[12] Vesanto, J et E.Alhoniemi (2000), *Clustering of the self organizing map*, IEEE Transactions on Neural Networks 11 (3), 586-600.

[13] Wettschereck, Aha D.W., and Mohri T. (1997), *A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms*, Artif. Intell Rev., 11(1-5):273-314.

Pondération locale des variables en apprentissage numérique non-supervisé

Summary

In this paper, we present a new approach of variable characterization during non-supervising clustering process. Our method is based on Self-organizing Map and the estimation of weights is done in conjunction with the automatic classification. These weightings are local and associated with each referent of the self-organizing map. They reflect the local importance of each variable for the classification. The weights are used for a local segmentation of the topological map, giving also a richest cutting taking into account the relevance of the variables. The results of the evaluation show that the proposed approach, as compared to other methods of classification, offers a finer segmentation of the map and with a better quality.