Pondération et classification simultanée de données binaires et continues

Nicoleta Rogovschi*,Mustapha Lebbah**, Nistor Grozavu**

*LIPADE, Université Paris Descartes
45 rue des Saints Pères
75270 Paris Cedex 06, France
prénom.nom@parisdescartes.fr
**LIPN-UMR 7030 Université Paris 13 - CNRS
99, av. J-B Clément - F-93430 Villetaneuse France.
prénom.nom@lipn.univ-paris13.fr

Résumé. Dans cet article, nous proposons une nouvelle approche de classification topologique et de pondération des variables mixtes (qualitatives et quantitatives codées en binaire) durant un processus d'apprentissage non supervisé. Cette approche est basée sur le modèle des cartes auto-organisatrices. L'apprentissage est combiné à un mécanisme de pondération des différentes variables sous forme de poids d'influence sur la pertinence des variables. L'apprentissage des pondérations et des prototypes est réalisé d'une manière simultanée en favorisant une classification optimisée des données. L'approche proposée a été validée sur des données qualitatives codées en binaire et plusieurs bases de données mixtes.

1 Introduction

La carte topologique proposée par (Kohonen, 2001) utilise un algorithme d'auto-organisation (SOM) qui fournit la quantification et la classification de l'espace des observations. Récemment, de nouveaux modèles de cartes auto-organisatrices dédiés à des données spécifiques ont été proposés dans (Bishop et al., 1998; Lebbah et al., 2008). Quelques-uns de ces modèles sont basés sur un formalisme probabiliste et d'autres sont des méthodes de quantification. Dans la littérature on trouve des approches basées sur la pondération comme les travaux de (Huang et al., 2005; Blansche et al., 2006; Grozavu et al., 2009). Pour les données continues, un modèle de cartes auto-organisatrices a été déjà proposé pour la pondération locale des variables appelé lw-SOM, (Grozavu et al., 2009). Cet algorithme se présente comme l'adaptation aux cartes SOM de l'approche de pondération proposée pour les K-moyennes par (Huang et al., 2005). Le modèle lw-SOM est dédié uniquement au cas des variables continues et n'est pas directement applicable aux données catégorielles. A notre connaissance, parmi les approches de pondération qui existent nous n'avons pas rencontré des travaux qui portent sur la classification non supervisée pondérée basée sur les cartes auto-organisatrices qui traitent des données mixtes (qualitatives codées en binaire et quantitatives). Nous voulons dans ce papier présenter une version déterministe qui tient compte de la nature des données sans utiliser des versions

"kernelisées" ou probabilistes. Dans ce travail, nous proposons une carte topologique auto-organisatrice basée sur la pondération des variables pour analyser les données mixtes (qualitatives et quantitatives). Les pondérations des attributs indiquent à un utilisateur l'importance relative de chacun des attributs pour la discrimination des classes. Elles correspondent aux degrés d'utilisation des variables dans le processus de classification. C'est un modèle de quantification qui fournit un ensemble conséquent de prototypes qui possèdent la propriété d'être facilement interprétables (les prototypes et les données appartiennent au même espace). Notre article est structuré de la manière suivante : dans la section 2, nous introduisons notre modèle et l'algorithme d'apprentissage associé. Les résultats obtenus sur des bases binaires et mixtes sont décrits dans la section 3. Une conclusion et quelques perspectives de notre méthode sont présentées dans la section 4.

2 La carte topologique des données mixtes

Comme dans le cas des cartes auto-organisatrices classiques, nous supposons que la grille $\mathcal{C}$ a une topologie discrète (un espace de sortie discret) définie par un graphe non orienté. D'habitude, ce graphe est une grille régulière à une ou deux dimensions. On note par N_{cell} le nombre de cellules dans la grille $\mathcal{C}$. Cette structure de graphe permet de définir une distance $\delta(i,j)$, entre deux cellules i et j de $\mathcal{C}$, comme étant la longueur de la plus courte chaîne permettant de relier les cellules i et j. Le modèle que nous proposons lw-MTM (Local Weighted Mixed Topological Map) est basé sur le formalisme de quantification des cartes topologiques. Soit $\mathcal{A}$ l'ensemble de données $\mathbf{x}$ d'apprentissage où chaque observation $\mathbf{x} = (x^1, x^2, ..., x^k, ..., x^d)$ est composée de deux parties : une partie continue $\mathbf{x}^{r[.]} = (x^{r[1]}, x^{r[2]}, ..., x^{r[n]})$ ($\mathbf{x}^{r[.]} \in \mathcal{R}^n$) et une autre partie catégorielle $\mathbf{x}^{c[.]} = (x^{c[1]}, x^{c[2]}, ..., x^{c[l]}, ..., x^{c[k]})$ où la l^{ime} composante $x^{c[l]}$ a M_l modalités. Chaque variable catégorielle peut être codée avec une variable binaire, comme un vecteur $x^{b[.]} = (x^{b[1]}, ..., x^{b[Ml]})$ où $x^{b[l]} \in \{0,1\}$). La partie catégorielle peut être représentée par une partie binaire $\mathbf{x}^{b[.]} = (x^{b[1]}, x^{b[2]}, ..., x^{b[l]}, ..., x^{b[m]})$ d'une telle manière que chaque observation $\mathbf{x}$ est ainsi une réalisation d'une variable aléatoire qui appartient à $\mathcal{R}^n \times \beta^m (\beta = \{0,1\})$ après avoir appliqué le codage binaire. Avec ces notations une observation particulière $\mathbf{x} = (\mathbf{x}^{r[.]}, \mathbf{x}^{b[.]})$ est un vecteur mixte (variables continues et binaires) de dimensions $d = n + m$. A chaque cellule j de la carte, on associe un vecteur référent $\mathbf{w}_j = (\mathbf{w}_j^{r[.]}, \mathbf{w}_j^{b[.]})$ de dimension d, où $\mathbf{w}_j^r \in R^n$ et $\mathbf{w}_j^{b[.]} \in \beta^m$ qui représente un codage binaire de la variable catégorielle associée $\mathbf{w}_j^{c[.]}$. On note par $\mathcal{W}$ l'ensemble des vecteurs référents, par $\mathcal{W}^r$ l'ensemble de la partie continue et par $\mathcal{W}^b$ la partie binaire des vecteurs référents. Dans la section suivante, nous présentons un nouveau modèle de cartes topologiques dédiées aux données mixtes. L'algorithme d'apprentissage associé s'inspire de la version batch de l'algorithme de Kohonen dédié aux données continues (Kohonen, 2001) et l'algorithme BinBatch dédiés aux données binaires (Lebbah et al., 2000). Ces deux modèles sont améliorés de manière à tenir compte de la pondération associée à chaque variable catégorielle et chaque variable continue. Dans l'algorithme que nous proposons, la mesure de similarité et le vecteur référent sont spécifiques à chaque type de données : c'est la distance Euclidienne avec une "moyenne" dans le cas des variables continues et la distance de Hamming avec le centre médian dans le cas binaire.

2.1 La minimisation de la fonction de coût

Comme dans le cas des cartes topologiques classiques, nous proposons de minimiser la fonction de coût modifiée suivante :

$$\mathcal{G}(\phi, \mathcal{W}, \mathcal{Y}) = \sum_{\mathbf{x} \in \mathcal{A}} \sum_{j \in C} \mathcal{K}(\delta(\phi(\mathbf{x}), j)) \mathbf{y}_j{}^{\tau} ||\mathbf{x} - \mathbf{w}_j||^2 \tag{1}$$

où τ est un paramètre d'ajustement qui est nécessaire pour l'estimation de l'ensemble des pondérations $\mathcal{Y}$. On note par ϕ la fonction d'affectation qui attribut chaque observation $\mathbf{x}$ à une cellule de $\mathcal{C}$. $\mathcal{K}^{\mathcal{T}}$ est une fonction de voisinage qui dépend du paramètre T (appelé température) : $\mathcal{K}(\delta) = \mathcal{K}^T(\delta/T)$, où $\mathcal{K}$ est une fonction noyau particulière qui est positive et symétrique ($\lim_{|x| \to \infty} \mathcal{K}(x) = 0$). Ainsi $\mathcal{K}$ définit pour chaque cellule j une région de voisinage sur la carte $\mathcal{C}$. Le paramètre T permet de contrôler la taille du voisinage d'influence d'une cellule sur la carte, celle-ci décroît avec le paramètre T. Par analogie avec l'algorithme des cartes topologiques, on peut faire décroître la valeur de T entre deux valeurs T_{max} et T_{min}. Le vecteur $\mathbf{y}_j = (\mathbf{y}_r^{r[.]}, \mathbf{y}_j^{c[.]})$ est le vecteur de pondération, où $\mathbf{y}_j^{r[.]}$ est la pondération de la partie continue des observations et $\mathbf{y}_j^{c[.]}$ est le vecteur de pondération des variables catégorielles. Ainsi, dans le cas de la partie catégorielle, la pondération dépend de la variable catégorielle et non pas de la modalité. On notera par $\mathcal{Y}$ l'ensemble des vecteurs de pondération. Pour le codage binaire la distance euclidienne est remplacée par la distance de Hamming $\mathcal{H}$, ainsi nous pouvons réécrire la fonction de coût de la manière suivante :

$$\begin{aligned} \mathcal{G}(\phi, \mathcal{W}, \mathcal{Y}) \quad = \quad & \sum_{\mathbf{x} \in \mathcal{A}} \sum_{j \in C} \mathcal{K}(\delta(\phi(\mathbf{x}), j)) (\mathbf{y}_j^{r[.]})^{\tau} \mathcal{D}_{euc}(\mathbf{x}^{r[.]}, \mathbf{w}_j^{r[.]}) \\ & + \sum_{\mathbf{x} \in \mathcal{A}} \sum_{j \in C} \mathcal{K}(\delta(\phi(\mathbf{x}), j)) (\mathbf{y}_j^{c[.]})^{\tau} \mathcal{H}(\mathbf{x}^{b[.]}, \mathbf{w}_j^{b[.]})] \end{aligned} \tag{2}$$

La minimisation de la fonction de coût (2), est réalisée en utilisant une procédure itérative en trois étapes :

- **La phase d'affectation** : supposant que $\mathcal{W}$ et $\mathcal{Y}$ sont fixés, on doit minimiser $\mathcal{G}(\phi, \mathcal{W}, \mathcal{Y})$ par rapport à ϕ. Cela nous amène à l'utilisation de la fonction d'affectation suivante : $\phi(\mathbf{x}) = \arg\min_j ((\mathbf{y}_j^{r[.]})^{\tau} ||\mathbf{x}^{r[.]} - \mathbf{w}_j^{r[.]}||^2 + (\mathbf{y}_j^{c[.]})^{\tau} \mathcal{H}(\mathbf{x}^{b[.]}, \mathbf{w}_j^{b[.]}))$
- **La phase de quantification** : supposant que ϕ et $\mathcal{Y}$ sont fixés, cette étape minimise $\mathcal{G}(\phi, \mathcal{W}, \mathcal{Y})$ par rapport à $\mathcal{W}$ dans l'espace $R^n \times \beta^m$. La minimisation de la fonction de coût (2) nous mène à minimiser les deux termes de la fonction respectivement dans R^n et dans β^m. On observe facilement que ces deux minimisations nous permettent de définir :
  - **la partie continue** $\mathbf{w}_j^{r[.]}$ du vecteur référent $\mathbf{w}_j$ comme le vecteur "moyenne" de la manière suivante : $\mathbf{w}_j^{r[.]} = \dfrac{\sum_{i \in C} \mathcal{K}(\delta(i,j)) \sum_{\mathbf{x} \in \mathcal{A}, \phi(\mathbf{x}) = r} \mathbf{x}_i^{r[.]}}{\sum_{i \in C} \mathcal{K}^T(\delta(i,j)) n_i}$, où n_i représente le nombre correspondant d'observations affectées.
  - **la partie binaire** $\mathbf{w}_j^{b[.]}$ du vecteur référent $\mathbf{w}_j$ comme le centre médian de la partie binaire des observations $\mathbf{x} \in \mathcal{A}$ pondérées par $\mathcal{K}(\delta(j, \phi(\mathbf{x})))$. Chaque composante

$\mathbf{w}_j^{b[.]} = (w_j^{b[1]}, ..., w_j^{b[l]}, ..., w_j^{b[m]})$ est ensuite calculée de la manière suivante : $w_j^{b[l]} = \begin{cases} 0 & \text{si } \left[\sum_{\mathbf{x}\in\mathcal{A}} \mathcal{K}(\delta(c,\phi(\mathbf{x})))(1-\mathbf{x}^{b[l]})\right] \geq \\ & \left[\sum_{\mathbf{x}\in\mathcal{A}} \mathcal{K}(\delta(j,\phi(\mathbf{x})))\mathbf{x}^{b[l]}\right] \\ 1 & \text{sinon} \end{cases}$,

– **La phase de pondération** : supposant que ϕ et $\mathcal{W}$ sont fixés, cette étape minimise $\mathcal{G}(\phi, \mathcal{W}, \mathcal{Y})$ par rapport à $\mathcal{Y}$ dans l'espace R^{n+k} où k est la dimension de la partie catégorielles. Dans le cas des variables catégorielles, le poids dépend de la variable et non pas des modalités. Le calcul de la pondération se fait de la manière suivante : $y_j^l = \begin{cases} 0, \text{ si } D_j^l = 0 \\ \frac{1}{\sum_t \left[\frac{D_j^l}{D_t^l}\right]^{\frac{1}{\tau-1}}}, \text{ sinon} \end{cases}$, où $D_j^l = \sum_{\mathbf{x}\in\mathcal{A}} \sum_{i=1}^{C} K(\delta(i,j))(x^l - w_i^l)^2$

La minimisation de $\mathcal{G}(\phi, \mathcal{W}, \mathcal{Y})$ est effectuée d'une manière itérative en appliquant les trois étapes principales. A la fin de l'apprentissage $\mathbf{w}_c$ qui a le même codage que les observations peut être décodé. En pratique, comme dans le cas des cartes topologiques traditionnelles, on utilise une fonction de lissage pour contrôler la taille du voisinage.

3 Résultats expérimentaux

3.1 La base Zoo

C'est une base de données extraite du répertoire UCI, (Asuncion et Newman, 2007). Ce jeu de données contient l'information sur 101 animaux décrits par 16 variables qualitatives : dont 15 variables sont binaires (oui/non) et une est catégorielle avec 6 modalités. Chaque animal est étiquetté de 1 à 7 conformément à sa classe (son espèce). Utilisant le codage disjonctif pour les variables qualitatives avec 6 valeurs possible, on obtient une matrice binaire 101×21 ($individus \times variables$). Les résultats de notre approche sur la base Zoo sont représentés dans la figure 1. On peut visualiser les référents, ainsi que les variables qui caractérisent ces référents pour chaque cellule de la carte. Pour une meilleure et une simple analyse, on a sélectionné uniquement les variables associées à la modalité "oui" avec une pondération supérieure à 0.02. On constate qu'on a quasiment des regroupements "homogènes" et mieux séparés. On constate aussi que certains types de poisson sont regroupés autour de cellules voisines (cellule 12, 13,17, 18) avec certaines variables communes : "aquatic, "toothed", "backbone", " tail". La même analyse peut être réalisée sur le reste des cellules. On retrouve sur toute la carte une distribution homogène des modalités et un regroupement des animaux autour des modalités.

3.2 Autres bases de données

Dans cette section nous montrons les contributions de notre modèle lw-MTM par rapport à l'algorithme déterministe sans tenir compte de la pondération, appelé ici MTM et l'algorithme probabiliste PrMTM (Probabilistic Topological Map), (Rogovschi et al., 2008). Pour la suite on utilise trois autres bases mixtes obtenues du répertoire UCI (Asuncion et Newman, 2007). Pour évaluer la qualité de la classification, nous adoptons une approche d'évaluation qui utilise des étiquettes externes. Ainsi, on utilise la pureté de la classification pour mesurer les résultats de la classification. Nous avons comparé notre modèle lw-MTM avec l'algorithme MTM (qui

case 1 : boar,calf,cheetah,goat, leopard,lion,lynx,mongoose, polecat,pony,puma, pussycat,raccoon, reindeer, tortoise, wolf *hair, milk,predator,toothed,tail,catsize*	*case 2 :* aardvark,bear, cavy,hamster *hair,milk,toothed,backbone,breathe s, tail,*	*case 3* *eggs,toothed, backbone,breathes,tail*	*case 4 :* lark,pheasant,sparrow, *feathers,eggs,airborne backbone, breathes,tail*	*case 5:* *feathers,eggs,tail,backbone*
case 6: girl *hair,milk,predator,toothed,backbone tail,catsize*	*case 7:* *eggs,aquatic,predator,toothed, legs,tail,catsize,backbone,breathes, fins*	*Case8:* haddock,newt,penguin, seahorse, sole *feathers,eggs,aquatic, backbone,breathes,tail*	*Case9:* carp,chicken,crow,dove parakeet,rhea,skimmer, duck, flamingo,gull, hawk, skua,swan,vulture kiwi,ostrich, *eathers,airborne backbone,breathes tail*	*case 10:* *feathers,eggs ,airborne, predator,backbonebreathes, tail*
case 11: *milk,aquatic,predator,toothed backbone,breathes,tail, catsize*	*case 12:* dogfish,dolphin,pike platypus,porpoise,tuna *eggs,aquatic,predator toothed,backbone,fins legs,tail,catsize*	*case 13:* bass,catfish,chub,herring, piranha, scorpion,seasnake,stingray *eggs,aquatic, predator, toothed backbone, fins,legs,tail*	*case 14:* clam, gnat,octopus pitviper,seawasp slowworm,tuatara *eggs,predator,backbone breathes, tail*	*case 15:* lobster,starfish crab,crayfish *eggs,aquatic,predator*
case 16: hare, squirrel, vole *hair,milk,toothed,backbone tail,*	*case 17:* frog, fruitbat, vampire *hair,milk,predator,toothed, backbone,breathes,tail*	*case 18:* *eggs,aquatic, predator, toothed,backbone, breathes, tail*	*case 19:* honeybee, housefly, moth, slug, wasp, worm *eggs,breathes*	*case 20* *eggs, breathes*
case 21: antelope, buffalo, deer,elephant giraffe, gorilla, oryx, seal, wallaby *tail, catsize*	*Case 22:* mink,mole, opossum,sealion *hair, milk,tail,catsize*	*case 23:* frog,toad *eggs,aquatic,predator toothed,backbone, breathes*	*case 24:* flea, termite *eggs.*	*case 25:* ladybird *eggs*

FIG. 1 – *Carte lw-MTM 5×5 représentant l'ensemble des référents et des variables (ici présentées en rouge) par cellule.*

Pureté : %	MTM	PrMTM	lwMTM
Cleve (13×7)	83.39	84.45	85.76
Credit (13×10)	82.66	84.57	86.44
Thyroid (21×14)	95.38	97.41	97.53

TAB. 1 – *Comparaison entre lw-MTM, MTM et PrMTM utilisant l'indice de pureté sur 50 expérimentations. MTM : Carte topologique classique dédiée aux données mixtes. PrMTM : carte topologique probabiliste utilisant la loi Gaussienne et Bernoulli*

ne prend pas en compte les pondérations) et l'algorithme probabiliste PrMTM. Dans ces expérimentations, la comparaison des différents résultats est mesurée à l'aide du taux de pureté en utilisant l'étiquette connue de chaque observation. La comparaison est réalisée en calculant la moyenne des puretés sur 50 expériences. Le tableau 1 montre les performances atteintes avec notre modèle lw-MTM et les modèles PrMTM et MTM. Nous observons une amélioration des puretés sur toutes les bases. En examinant le tableau 1, nous observons par exemple, pour la base *Cleve* une amélioration de la pureté de 83.39% à 85.76%. En ce qui concerne la base de données *Credit* nous améliorons les résultats de 82.66% à 86.4%o. Finalement pour la base de données *Thyroid* on observe une amélioration de 95.38% à 97.53%.

4 Conclusion

Dans cet article nous avons proposé une approche de cartes auto-organisatrice pondérée pour les données catégorielles et mixtes. La pondération de la distance durant le processus d'apprentissage permet de détecter les degrés de participation des différents attributs durant

la classification. Plus une variable a un poids élevé, plus l'algorithme de classification tiendra compte des informations véhiculées par cette variable. La pondération de la distance a pour objectif l'adaptation de la mesure de (dis)similarité entre observations et l'amélioration des résultats de la classification en renforçant principalement les variables les plus importantes. La pondération de la distance est très utile surtout dans le cadre des données mixtes, puisque dans le cas ou la partie qualitative codée en binaire est beaucoup plus volumineuse que la partie quantitative (et vice-versa), ça nous permet lors de la phase d'apprentissage de régulariser les adaptations et tenir compte de l'importance de chaque variable. Comme perspectives de ce travail, nous envisageons d'utiliser les pondérations estimées pour effectuer une sélection des variables les plus pertinentes et les comparer avec d'autres techniques de sélection de variables.

Références

Asuncion, A. et D. Newman (2007). UCI machine learning repository. http ://www.ics.uci.edu/∼mlearn/MLRepository.html.

Bishop, C. M., M. Svensén, et C. K. I.Williams (1998). GTM : The generative topographic mapping. *Neural Comput 10*(1), 215–234.

Blansche, A., P. Gancarski, et J. Korczak (2006). Maclaw : A modular approach for clustering with local attribute weighting. *Pattern Recognition Letters 27(11)*, 1299–1306.

Grozavu, N., Y. Bennani, et M. Lebbah (2009). From variable weighting to cluster characterization in topographic unsupervised learning. In *IJCNN'09 : Proceedings of the 2009 international joint conference on Neural Networks*, pp. 609–614. Institute of Electrical and Electronics Engineers Inc., The.

Huang, J. Z., M. K. Ng, H. Rong, et Z. Li (2005). Automated variable weighting in k-means type clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence 27(5)*, 657–668.

Kohonen, T. (2001). *Self-organizing Maps*. Springer Berlin.

Lebbah, M., Y. Bennani, et N. Rogovschi (2008). A probabilistic self-organizing map for binary data topographic clustering. *International Journal of Computational Intelligence and Applications 7*(4), 363–383.

Lebbah, M., S. Thiria, et F. Badran (2000). Topological map for binary data. In *Proceedings European Symposium on Artificial Neural Networks-ESANN 2000, Bruges, April 26-27-28*, pp. 267–272.

Rogovschi, N., M. Lebbah, et Y. Bennani (2008). Probabilistic mixed topological map for categorical and continuous data. In *ICMLA*, pp. 224–231.

Summary

This paper introduces a weighted self-organizing map for clustering, analysis and visualization of mixed data (binary/continuous). The learning of weights and prototypes is done in a simultaneous manner assuring an optimised data classification. The learning oh these topological maps is combined with a weighting process of the different variables by computing weights wich influence the quality of clustering. We illustrate the power of this method with data sets taken from a public data set repository.