A weighted Self-Organizing Map for mixed continuous and categorical data

Nicoleta Rogovschi, Mustapha Lebbah, Younès Bennani

LIPN CNRS UMR 7030, Paris 13 University
99, J.-B. Clément, 93430 Villetaneuse, France
Name.Surname@lipn.univ-paris13.fr

Abstract

This paper introduces a weighted self-organizing map for clustering, analysis and visualization of mixed data (binary/continuous). We propose a formalism dedicated to mixed data in which cells are represented by a Bernoulli and Gaussian distribution. Each cell is characterized by a prototype with the same coding as used in the data space. The learning of weights and prototypes is done in a simultaneous manner assuring an optimized data classification. More a variable has a high weight, more the clustering algorithm will take into account the information transmitted by this variable. The learning oh these topological maps is combined with a weighting process of the different variables by computing weights which influence the quality of clustering.

We illustrate the power of this method with data sets taken from a public data set repository: a handwritten digit data set and other three data sets. The results show a good quality of the topological ordering and homogenous clustering.

1 Introduction

The topological map proposed by Kohonen (2001) uses a self-organization algorithm (SOM) which provides quantification and clustering of the observation space. More recently, new models of topological maps dedicated to specific data were proposed in (Bishop et al., 1998, Kaban et al., 2001, Lebbah et al., 2000). Some of these models are based on a probabilistic formalism and a learning procedure to maximize the likelihood function of the data set, the others are quantization methods.

In the literature there are approaches based on weighting as (Huang et al., 2005; Blansche et al., 2006; Grozavu et al., 2008) and approaches based on feature selection like (Basak et al., 1998; Bassiouny et al., 2004; Liu et al., 2005; Questier et al., 2005; Li et al., 2006; Wiratunga et al., 2006; Strickert et al., 2006; Li et al., 2007; Guerif and Bennani, 2007). For the continuous data, a model for local variables weighting using SOM was proposed, called lw-SOM (Grozavu et al., 2009). This algorithm is an adaptation to SOM of the weighting approach proposed for K-means by Huang et al. (2005). The model lw-SOM is dedicated to continuous variables and is not directly applicable to categorical data. Among the probabilistical method of variable selection we find the works of Kim et al. (2003) and Cord et al. (2006) where methods of variables selection are used with the EM algorithm. The main idea is that a variable wich wasn't selected don't have a big influence in the computation of the data likelihood.

In this paper we propose a topological self organizing algorithm for analyzing mixed (continuous and binary) data. It is a quantization model which provides a consistent set of prototypes whose particularity is to be interpreted (prototypes and data belong to the same space and have a meaningful interpretation). The variable weights provide to a user the relevance of each variable for the clustering. They correspond to the degree of use of variable in the clustering process.

In section 2, we present the model and the iterative algorithm. In the section 3, we present some applications of proposed method. The experiments involve handwritten numerals ($0-9$), and three other data sets available in Asuncion and Newman (2007). These data sets allow us to prove the importance of the weighting for the clustering process. Our conclusions are reported in section 4.

2 Local weighted Mixed Topological Map

As with a traditional self-organizing map, we assume that the lattice $\mathcal{C}$ has a discrete topology (discrete output space) defined by an undirect graph. Usually, this graph is a regular grid in one or two dimensions. We denote the number of cells in $\mathcal{C}$ as N_{cell}. For each pair of cells (i,j) on the map, the distance $\delta(i,j)$ is defined as the length of the shortest chain linking cells i and j. The lw-MTM (Local Weighted Mixed Topological Map) model is based on the quantization formalism of topological maps.

Let A be the learning data set $\mathbf{x}$ where each observation $\mathbf{x} = (x^1, x^2, ..., x^k, ..., x^d)$ is made of two parts: continuous part $\mathbf{x}^{r[.]} = (x^{r[1]}, x^{r[2]}, ..., x^{r[n]})$ ($\mathbf{x}^{r[.]} \in \mathcal{R}^n$) and categorical part $\mathbf{x}^{c[.]} = (x^{c[1]}, x^{c[2]}, ..., x^{c[j]}, ..., x^{c[k]})$ where the l^{th} component $x^{c[l]}$ have M_l modalities. Each categorical variable can be coded with a binary variable, thus, each categorical variable $x^{c[l]}$, is coded with the vector $x^{b[.]} = (x^{b[1]}, ..., x^{b[Ml]})$ where $x^{b[l]} \in \{0,1\}$). The categorical part can be represented by a binary part $\mathbf{x}^{b[.]} = (x^{b[1]}, x^{b[2]}, ..., x^{b[l]}, ..., x^{b[m]})$ such as each observation $\mathbf{x}$ is thus, a realization of a random variable which belongs to $\mathcal{R}^n \times \{0,1\}^m$. Using these notations a particular observation $\mathbf{x} = (\mathbf{x}^{r[.]}, \mathbf{x}^{b[.]})$ is a mixed vector (continuous and binary variables) of dimension $d = n + m$. In our model, we assume that a given data set has been drawn from N_{cell} clusters.

For each cell c of the grid, we associate a referent vector $\mathbf{w_c} = (\mathbf{w}_c^{r[.]}, \mathbf{w}_c^{b[.]})$ of dimension d, where $\mathbf{w}_c^r \in R^n$ and $\mathbf{w}_c^{b[.]} \in \beta^m$ which is a binary coding of multidimensional categorical variable $\mathbf{w}_c^{c[.]}$. We denote by $\mathcal{W}$ the set of the referents vectors, by $\mathcal{W}^r$ the set of the numerical part and by $\mathcal{W}^b$ the binary part of the referent vectors.

In the following section we present a new model of topological map dedicated to mixed data. The associated learning algorithm is derived from the batch version of the Kohonen algorithm dedicated to numerical data (Kohonen, 2001) and the BinBatch algorithm which is dedicated to binary data (Lebbah et al., 2000). These models are improved to take into account the variable weights. In this algorithm, the similarity measure and the estimation of the referent vectors are specific for each type of data : it is the Euclidian distance with the mean vector in the continuous case and the Hamming distance with the median center in the binary case.

2.1 Minimization of the cost function

As the classical topological maps we propose to minimize the following cost function.

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

Where τ is a fitting parameter necessary for the estimation of the set of the weight vectors $\mathcal{Y}$, and ϕ assigns each observation $\mathbf{x}$ to a single cell in $\mathcal{C}$. $\mathcal{K}^T$ is a neighborhood function depending on the parameter T (called temperature): $\mathcal{K}^T(\delta) = \mathcal{K}(\delta/T)$, where $\mathcal{K}$ is a particular kernel function which is positive and symmetric ($\lim\limits_{|x|\to\infty} \mathcal{K}(x) = 0$). Thus $\mathcal{K}$ defines for each cell j a neighborhood region in $\mathcal{C}$. The parameter T allows to control the size of the neighborhood influencing a given cell on the map. As with the Kohonen algorithm, we decrease the value of T between two values T_{max} and T_{min}. The vector $\mathbf{y_j} = (\mathbf{y}_j^{r[.]}, \mathbf{y}_j^{c[.]})$ is the weighted vector, where $\mathbf{y}_j^{r[.]}$ is the continuous weight part and $\mathbf{y}_j^{c[.]}$ is a categorical weight variable (not binary variable).

In this expression $||\mathbf{x} - \mathbf{w}_j||^2$ is the square of the Euclidian distance. Since for binary vectors the Euclidian distance is no more than the Hamming distance $\mathcal{H}$, then the Euclidian distance can be rewritten by:

$$||\mathbf{x} - \mathbf{w}_c||^2 = ||\mathbf{x}^{r[.]} - \mathbf{w}_c^{r[.]}||^2 + \mathcal{H}(\mathbf{x}^{b[.]}, \mathbf{w}_c^{b[.]})$$

As for the mixed topological map (MTM) algorithm, we use this expression to rewrite the cost function as:

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

Where

$$\mathcal{G}_{som}(\phi, \mathcal{W}, \mathcal{Y}) = \sum_{\mathbf{x}\in A}\sum_{j\in C} \mathcal{K}^T(\delta(\phi(\mathbf{x}), j)) \mathbf{y}_j^{r[.]} ||\mathbf{x}^{r[.]} - \mathbf{w}_j^{r[.]}||^2 \tag{3}$$

is the classical cost function used by the weighted Kohonen Batch algorithm (Grozavu et al., 2009), and

$$\mathcal{G}_{bin}(\phi, \mathcal{W}, \mathcal{Y}) = \sum_{\mathbf{x}\in A}\sum_{j\in C} \mathcal{K}^T(\delta(\phi(\mathbf{x}), j)) \mathbf{y}_j^{b[.]} \mathcal{H}(\mathbf{x}^{b[.]}, \mathbf{w}_j^{b[.]}) \tag{4}$$

is the modified cost function used in BinBatch algorithm (Lebbah et al. 2000). The old cost function proposed is :

$$\mathcal{G}_{bin}(\phi, \mathcal{W}) = \sum_{\mathbf{x}\in A}\sum_{j\in C} \mathcal{K}^T(\delta(\phi(\mathbf{x}), j)) \mathcal{H}(\mathbf{x}^{b[.]}, \mathbf{w}_j^{b[.]}) \tag{5}$$

Thus in this paper we propose a new cost function to deal with mixed data, in the same way we define a new function to binary data.

The minimization of the cost function (2), is made using an iterative process with two steps:

- **Assignment step**: assuming that $\mathcal{W}$ and $\mathcal{Y}$ are fixed, we have to minimize $\mathcal{G}(\phi, \mathcal{W}, \mathcal{Y})$ with respect to ϕ. This leads to use the following assignment function: $\phi(\mathbf{x}) = \arg\min_j((\mathbf{y}_j^{r[.]})^{\tau} ||\mathbf{x}^{r[.]} - \mathbf{w}_j^{r[.]}||^2 + (\mathbf{y}_j^{c[.]})^{\tau} \mathcal{H}(\mathbf{x}^{b[.]}, \mathbf{w}_j^{b[.]}))$

- **Quantization step**: assuming that ϕ and $\mathcal{Y}$ are fixed, this step minimizes $\mathcal{G}(\phi,\mathcal{W},\mathcal{Y})$ with respect to $\mathcal{W}$ in the space $R^n \times \beta^m$. The minimization of the cost function (2) leads to minimize the function $\mathcal{G}_{som}(\phi,\mathcal{W})$ (3) in R^n and $\mathcal{G}_{bin}(\phi,\mathcal{W})$ (4) in β^m. It is easy to see that these two minimizations allow to define:
  - the numerical part $\mathbf{w}_j^{r[.]}$ of the referent vector $\mathbf{w}_j$ as the mean vector as:

$$\mathbf{w}_j^{r[.]} = \frac{\displaystyle\sum_{i\in C} \mathcal{K}^T(\delta(i,j)) \sum_{\mathbf{x}\in\mathcal{A},\phi(\mathbf{x})=i} \mathbf{x}^{r[.]}}{\displaystyle\sum_{i\in C} \mathcal{K}^T(\delta(i,j)) n_i},$$

    where n_i represents the corresponding number of assigned observations.
  - the binary part $\mathbf{w}_j^{b[.]}$ of the referent vector $\mathbf{w}_j$ as the median center of the binary part of the observations $\mathbf{x} \in \mathcal{A}$ weihted by $\mathcal{K}^T(\delta(j,\phi(\mathbf{x})))$. Each component $\mathbf{w}_j^{b[.]} = (w_j^{b[1]},...,w_j^{b[l]},...,w_j^{b[m]})$ is then computed as follows:

$$w_j^{b[l]} = \begin{cases} 0 & \text{if } \left[\sum_{\mathbf{x}\in\mathcal{A}} \mathcal{K}^T(\delta(j,\phi(\mathbf{x})))(1-\mathbf{x}^{b[l]})\right] \geq \\ & \left[\sum_{\mathbf{x}\in\mathcal{A}} \mathcal{K}^T(\delta(j,\phi(\mathbf{x})))\mathbf{x}^{b[l]}\right] \\ 1 & \text{otherwise} \end{cases}$$

- **Quantization step**: assuming that ϕ and $\mathcal{W}$ are fixed, this step minimizes $\mathcal{G}(\phi,\mathcal{W},\mathcal{Y})$ with respect to $\mathcal{Y}$ in the space R^{n+m}. The weights are computed in the following way:

$$y_j^l = \begin{cases} 0, \text{ if } D_j^l = 0 \\ \frac{1}{\sum_t \left[\frac{D_j^l}{D_t^l}\right]^{\frac{1}{\tau-1}}}, \text{ otherwise} \end{cases}$$

  where

$$D_j^l = \sum_{\mathbf{x}\in\mathcal{A}} \sum_{i=1}^{C} K^T(\delta(i,j))(x_i^l - w_j^l)^2$$

The minimization of $\mathcal{G}(\phi,\mathcal{W},\mathcal{Y})$ is run by iteratively performing the two steps. At the end the vector $\mathbf{w}_c$, which shares the same code with the observations can be decoded in the same way, allowing a symbolic interpretation of binary part of referent vectors. The nature of the topological model reached at the end of the algorithm, the quality of the clustering and those of the topological order induced by the graph greatly depend on the neighborhood function $\mathcal{K}$. In practice, as for traditional topological map we use a smooth function to control the size of the neighborhood as $\mathcal{K}^T(\delta(c,r)) = \exp(\frac{-\delta(c,r)}{T})$. Using this kernel function, T becomes a parameter of the model. As in the Kohonen algorithm (Kohonen, 2001), we repeat the preceding iterations by decreasing T from an initial value T_{max} to a final value T_{min}.

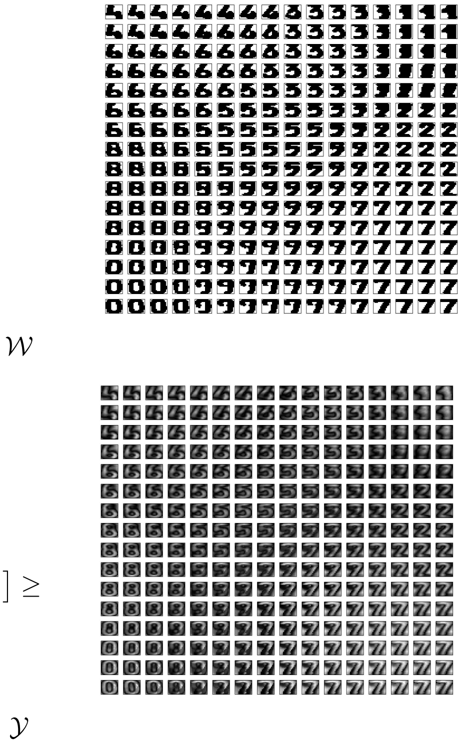

Figure 1: The map lw-MTM 16×16 representing the set of prototypes $\mathcal{W}$ and weights $\mathcal{Y}$.

3 Experimental validations

To evaluate the quality of clustering, we adopt the approach of comparing the results to a "ground truth". We use the clustering accuracy to measure the clustering results. This a common approach in the general area of data clustering. In general, the result of clustering is usually assessed on the basis of some external knowledge about how clusters should be structured. This may imply evaluating separation, density, connectedness, and so on. The only way to assess the usefulness of a clustering result is indirect validation, whereby clusters are applied to the solution of a problem and the correctness is evaluated against objective external knowledge. This procedure is defined by Jain and Dubes (1988) as "validating clustering by extrinsic classification", and has been followed in many other studies (Khan and Kant, 2007; Andreopoulos et al., 2006). We feel that this approach is reasonable one if we don't want to judge clustering results by some cluster validity index, which is nothing but a bias toward some preferred cluster property (e.g., compact, or well separated, or connected). Thus, to adopt this approach we need labeled data sets, where the external (extrinsic) knowledge is the class information provided by labels. Hence, if the lw-MTM finds significant clusters in the data, these will be reflected by the distribution of classes. Therefore we operate a vote step for clusters and compare them to the behavior methods from the literature. The so-called vote step consists in the following. For each cluster $c \in \mathcal{C}$:

- Count the number of observations of each class l (call it N_{cl}).
- Count the total number of observation assigned to the cell c (call it N_c).
- Compute the proportion of observations of each class (call it $S_{cl} = N_{cl}/N_c$).
- Assign to the cluster the label of the most represented class ($l = \arg\max_l(S_{cl})$.

A cluster c for which $S_{cl} = 1$ for some class labeled l is usually termed a "pure" cluster, and a purity measure can be expressed as the percentage of elements of the assigned class in a cluster. The experimental results are then expressed as the fraction of observations falling in clusters which are labeled with a different class from that of the observation. This quantity is expressed as a percentage and termed "error percentage" (indicated as $Err\%$ in the results).

3.1 Handwritten data

This experiment concerns a data set consisting of the handwritten numerals ("0"−"9") extracted from a collection of Dutch utility maps, Asuncion and Newman (2007). There are 200 samples of each digit such that there is a total of 2000 samples. Each sample is a 15×16 binary pixel image. The data set consisted of a 2000×240 binary data matrix. Each qualitative variable is a pixel with two possible values "On=1" and "Off=0". The figure 1 shows two maps obtained from the learning of lw-MTM map of 16×16 size with the fitting parameter $\tau = 2$. In the first figure we can visualize the prototypes of the map which are binary where each pixel “black/white” denotes the state of the binary variable (“Off/On“). In the second figure the grey shading shows the relevance of the variables. By analyzing these two figures we observe that the topological order is respected on the map ($\mathcal{W}$) and the contours of the numbers correspond to relevant variables which are detected by our proposed approach..

3.2 Other data sets

We use the following three categorical data sets obtained from UCI repository (Asuncion and Newman, 2007).

Heart disease
This is D. Detrano's heart disease data set that was generated by the Clevelande Clinic (Asuncion and Newman, 2007). The data set has 303 observations, each one is described by 6 continuous and 8 categorical variables. The observations are also classified into two classes, each class is either healthy (buff) or with heart-disease (sick). In both cases we use a binary coding to code a categorical variable. Hence, using a disjunctive coding we obtain $m = 17$ binary variables for *Heart disease* data set. The variable with two modalities is coded using only one binary variable indicating a presence or absence of modalities. The learning of a map with the dimensions 13×7 cells is made with all observations.

Credit Approval
This file concerns credit card applications. All attribute names and values have been changed to meaningless symbols to protect confidentiality of the data. This dataset is interesting because there is a good mix of attributes - continuous, nominal with small numbers of values, and nominal with larger numbers of values. There are also a few missing values. The data set has 666 observations, each one is described by 9 continuous and 6 categorical variables. Examples represent positive and negative instances of people who were and were not granted credit.

Thyroid disease
This dataset contains thyroid disease records supplied by the Garavan Institute and J. Ross ; Quinlan, New South Wales Institute, Syndney, Australia in 1987. The data set has 3163 observations, each one is described by 7 continuous and 12 categorical variables. Five laboratories tests are used to try to predict whether a patient's thyroid to the class hypothyroidism or hyperthyroidism. The diagnosis (the class label) was based on a complete medical record, including anamnesis, scan etc. Table 1, provides a short description of used data sets.

Table 1: Data sets used in the experimentation. #obs: data set size; #cl: number of classes; dim.Cat: categorical dimension; dim.Re: continuous variable dimension.

Data sets	dim.Cat	dim.Re	#obs	#cl
Heart disease	8	6	303	2
Credit	6	9	666	2
Thyroid	12	7	3163	2

We use the clustering accuracy for measuring the clustering results. This index is a purity measure which can be expressed as the percentage of elements of the assigned class in a cluster. This is a common approach in the general area of data clustering. We compared the proposed lw-MTM model with the classical determinist algorithm MTM and the probabilistic algorithm PrMTM. We computed the purity index on 50 experiences. The table 2 shows the performances obtained by the proposed model lw-MTM and the other models MTM and PrMTM. We observe an improvement of map purity on all datasets.

Table 2: Comparison of lw-MTM, MTM and PrMTM using the purity index on 50 experimentations. MTM: Classical topological map dedicated to mixed data. PrMTM: Probabilistic mixed topological map using Gaussian and Bernoulli distributions.

$\%$ Purity	MTM	PrMTM	lw-MTM
Heart disease (13 $\times$ 7)	83.39	84.45	85.76
Credit (13 $\times$ 10)	82.66	84.57	86.44
Thyroid (21 $\times$ 14)	95.38	97.41	97.53

For example, analyzing the table 2 we observe for the *Heart disease* data set, an improvement of the purity index from 83.39% to 85.76% using the same map size. For $Credit$ data set, we observe also improvement of the purity index from 82.66% to 86.44%. Finally with *Thyroid* data set, we improve the performance from 95.38% to 97.53% . thank to the introduced weights during the learning process, we may observe a clear improvement in the purity rate with lw-MTM model .

4 Conclusion

In this paper, we proposed a weighted self-organizing map for clustering categorical and mixed data. The weighting of the distance during the learning phase allows to detect the degrees of participation of each variable during the clustering process. More a variable has a high weight, more the clustering algorithm will take into account the informations transmited by this variable. The distance weighting has the goal to adapt the (dis)similarity measure between the observations and to improve the clustering results by mainly strengthening the most relevant variables. The distance weighting is very useful in the case of mixed data, because if for the learning dataset the categorical part is much larger than the continous part (and vice versa), the weighting process allows us to regularize the adaptations during the learning phase and to take into account the relevance of each variable. A perspective of this work can be the use of the computed weights to select the most relevant variables and thus to reduce the dimensionality of the data.

Reference

Andreopoulos B., A. An and X. Wang (2006). *Bi-level clustering of mixed categorical and numerical biomedical data.* International Journal of Data Mining and Bioinformatics, v. 1, number 1, pages 19-56.

Asuncion A. and D.J. Newman (2007). UCI Machine Learning Repository. http://www.ics.uci.edu/∼mlearn/MLRepository.html, University of California, Irvine, School of Information and Computer Sciences.

Basak J., Rajat K. De and Sankar K. Pal (1998). *Unsupervised feature selection using a neuro-fuzzy approach;*, In Pattern Recogn. Lett., v. 19, number 11, pages 997-1006, Elsevier Science Inc., New York, NY, USA.

Bassiouny S., M. Nagi and M. F. Hussein (2004). *Feature Subset Selection in SOM Based Text Categorization.*, IC-AI, pages 860-866.

Bishop, C. M., Svensén M. and Williams C. K. I. (1998). *GTM: The generative topographic mapping.*Neural Comput journal, volume 10, pages 215-234

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

Cord A., Ambroise C. and J.-P. Cocquerez (2006). *Feature selection in robust clustering based on Laplace mixture.* In Pattern Recognition Letters, v. 27, number 6, pages 627-635.

Grozavu N., Y. Bennani and M. Lebbah (2008). *Pondération locale des variables en apprentissage numérique non-supervisé*, Extraction et Gestion des Connaissances (EGC 08), pages 45-54, Sophia-Antipolis, France.

Grozavu N., Bennani Y. and M. Lebbah (2009). *From variable weighting to cluster characterization in topographic unsupervised learning.* IJCNN'09: Proceedings of the 2009 international joint conference on Neural Networks, isbn 978-1-4244-3549-4, pages 609–614, Atlanta, Georgia, USA.

Guérif S. and Y. Bennani (2007). *Dimensionality reduction trough unsupervised features selection.*

International Conference on Engineering Applications of Neural Networks, Hellas.

Huang J. Z., Michael K. Ng, H. Rong and Z. Li (2005). *Automated Variable Weighting in k-Means Type Clustering.* IEEE Transactions on Pattern Analysis and Machine Intelligence, v. 27(5), pages 657-668.

Jain K. and C. Dubes (1988). *Algorithms for clustering data.* isbn 0-13-022278-X, Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Kaban A. and M. Girolami (2001). *A combined latent class and trait model for the analysis and visualization of discrete data.* IEEE Trans. Pattern Anal. Mach. Intell, V. 23, pages 859-872.

Khan S. and Kant S. (2007). *Computation of Initial Modes for K-modes Clustering Algorithm Using Evidence Accumulation.* IJCAI, pages 2784-2789.

Kim Y., Street W.N. and F. Menczer (2002). *Evolutionary model selection in unsupervised learning.* In Intelligent Data Analysis Journal, v. 6, pages 531-536

Kohonen, T. (2001). *Self-organizing Maps.* Springer Berlin, Vol. 30, 501 pages, ISBN=3-540-67921-9.

Lebbah M., S. Thiria and F. Badran (2000). *Topological Map for Binary Data.* In Proceedings of European Symposium on Artificial Neural Networks-ESANN 2000, Bruges, pages 267-272.

Li Y., Lu B.-L. and Z.-F. Wu (2006). *A Hybrid Method of Unsupervised Feature Selection Based on Ranking.* ICPR '06: Proceedings of the 18th International Conference on Pattern Recognition, isbn 0-7695-2521-0, pages 687-690, IEEE Computer Society, Washington, DC, USA.

Li Y., Lu B.-L. and Z.-F. Wu (2007). *Hierarchical fuzzy filter method for unsupervised feature selection.* Journal of Intelligent and Fuzzy Systems, v. 18, number 2, pages 157-169.

Liu L., Kang J., Yu J. and Z. Wang (2005), *A comparative study on unsupervised feature selection methods for text clustering.* pages 597-601.

Strickert M., Sreenivasulu N., Peterek S., Weschke W., Mock H.-P. and U. Seiffert (2006). *Unsupervised Feature Selection for Biomarker Identification in Chromatography and Gene Expression Data.* In ANNPR, pages 274-285.

Questier F., R. Put, D. Coomans, B. Walczak and Y. Heyden (2005). *The use of CART and multivariate regression trees for supervised and unsupervised feature selection.* pages 45-54.

Wiratunga N., Lothian R. and S. Massie (2006). *Unsupervised Feature Selection for Text Data.* In ECCBR, Lecture Notes in Computer Science, v. 4106, pages 340-354.