SELECTION OF CLUSTERS NUMBER AND FEATURES SUBSET DURING A TWO-LEVELS CLUSTERING TASK

Sébastien Guérif and Younès Bennani
Université Paris 13, LIPN - CNRS UMR 7030
F-93430 Villetaneuse, France
{sebastien.guerif,younes.bennani}@lipn.univ-paris13.fr

ABSTRACT
Simultaneous selection of the number of clusters and of a relevant subset of features is part of data mining challenges. A new approach is proposed to address this difficult issue. It takes benefits of both two-levels clustering approaches and wrapper features selection algorithms. On the one hands, the former enhances the robustness to outliers and to reduce the running time of the algorithm. On the other hands, wrapper features selection (FS) approaches are known to given better results than filter FS methods because the algorithm that uses the data is taken into account. First, a Self-Organizing Maps (SOM), trained using the original data sets, is clustered using k-means and the Davies-Bouldin index to determinate the best number of a clusters. Then, an individual pertinence measure guides the backward elimination procedure and the feature mutual pertinence is measure using a collective pertinence based on the quality of the clustering.

KEY WORDS
Clustering, feature selection, self-organizing maps, model selection

1 Introduction

During the last decade, it became obvious that adapted tools are needed to exploit more and more huge companies databases. Actually, databases contain important hidden knowledge and the matter of data mining is to emphasize it. The curse of dimensionality problem states that the number of needed examples for training grows exponentionnaly with the dimensionality of the data. That way, whereas Knowledge Discovery from Database (KDD) is only possible because of the data redundancy, too many redundant features stand in the way of the nuggets discovery. This issue can be addressed by one of the two main approaches, namely, features extraction or feature selection.
The former presents a major drawback, actually, an important effort from the user is required to interpret and understand the new representation his data. Among the techniques of this category, the most widely used are probably Principal Component Analysis (PCA) [1, 2] which suffers from numerical instabilities whenever the correlation of the data is ill-conditionned. Moreover, this methods assume that the most relevant dimensions are those with the largest variance which not always the case as it is showed by the figure 1. Other approaches that does not suffer from the same numerical instabilities has been proposed [3] although the features extracted are not as intuitive as the original features. Whereas, the problem of feature selection

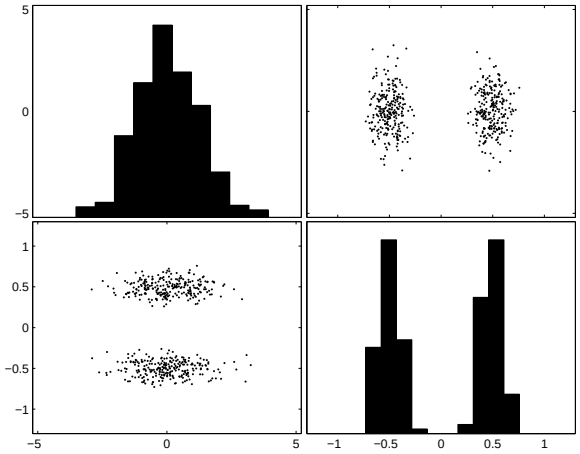

Figure 1. The feature variance is not always a relevant pertinence measure; actually, in this example, whereas $\sigma^2(X) = 1.03$ and $\sigma^2(Y) = 0.25$, the best separation is provided by the Y axis.

had been widely studied in the context of supervised learning, it gains researchers interest more recently in the context of unsupervised learning. In the context of supervised learning, feature selection is driven by the main purpose : achieve better accuracy on unseen data. Nevertheless, in the unsupervised learning framework, the issue is very different because neither the data labels nor their number are available. Therefore, the notion of feature relevance is not as obvious the latter context as in the former context. Anyway, selection of a relevant features subset remains a crucial stake for the data-mining techniques. In this paper, we propose an original method to find both the right number of clusters and the respective subset of features. Our approach is based on both the Davies-Bouldin index [4, 5] and the Test Values [6]. It is assumed that features that does not participate in the structure identified are irrelevant and should be thrusted away from the subset of features selected.

The rest of this paper is organized as follows. The two-levels clustering approache used is presented in section 2. Then, the feature selection method proposed is presented in section 3. Finally, some experimental results are given befor to conclude.

2 Method

2.1 Self-Organizing Maps

SOM was introduced by Pr. Teuvo Kohonen in the early 80's as a convenient clustering and visualization tool. High-dimensional data are projected on a low dimension discrete space, called the topological map, preserving the local topology of the initial space; thus, the observations which are close to each other are projected on a localized area. A map should be viewed as a set of neurons (or units), organized according to a grid that defines their neighbourhood relationships. Each neuron is associated to one point of the observations' space: its prototype.

Self-Organizing Maps (SOM) implement a particular form of competitive artificial neural networks; when an observation is recognized, activation of an output cell competition layer leads to inhibit activation of other neurons and reinforce itself. It is said that it follows the so called *Winner Takes All* rule. Actually, neurons are specialized in the recognition of one kind of observations. The learning is unsupervised because neither the classes nor their number is fixed a priori. A SOM consists in a two dimensional layer of neurons which are connected to the inputs with exciting connections and to their neighbors with inhibiting links.

The training set is used to organize these maps under topological constraints of the input space. Thus, a mapping between the input space and the network space is constructed; closed observations in the input space would activate two closed units of the SOM. An optimal spatial organization is determined by the SOM from the received information, and when the dimension of the input space is lower than three, both position of weights vectors and direct neighbourhood relations between cells can be represented visually.

2.2 Learning algorithms

For convenience, let us mention some notations : let N be the number of sample points in the data set Ω, n be the number of features in the original feature set F, r be the number of features in the reduced feature set F_R, M be the size of the map units set U and ω_j be the prototype of the j^{th} unit.

Connectionist learning is often presented as a minimization of a risk function (cost function). In our case, it will be carried out by the minimization of the distance between the input samples and the map prototypes (referents), weighted by a neighbourhood function h_{ij}. The criterion to be minimized is defined by:

$$R_{SOM} = \frac{1}{N} \sum_{x_i \in \Omega} \sum_{j \in U} h_{b_i j}. \left\| \omega_j - x_i \right\|^2 \quad (1)$$

where b_i is the *Best Matching Unit* (BMU) of the sample point $x_i \in \Omega$ and is defined as the unit with the closest prototype:

$$b_i = \arg\min_{j \in U} \{ \left\| \omega_j - x_i \right\|^2 \}$$

In our experiments, we use the gaussian neighborhood function h defined

$$h_{ij} = \exp\left(-\frac{d^2(i,j)}{2.\sigma^2(t)} \right)$$

where $d(i,j)$ is the distance between units i and j on the map and $\sigma(t)$ is a decreasing function that defines the size of the neighborhood considered at step t.

Two main approaches can be used to optimize the criterion mentionned above, namely the *on-line algorithm* and the *batch algorithm*. Whereas the latter suffers from several drawbacks [7], it provides faster convergence. So we choose the batch Kohonen's algorithm [8] because our approach necessitates several running of the learning of the learning algorithm. The weights of all the neurons are updated until stabilization according to the following adaptation rules:

$$\omega_j(t+1) = \frac{\sum_{i \in \Omega} h_{b_i j} x_i}{\sum_{i \in \Omega} h_{b_i j}} \quad (2)$$

2.3 SOM segmentation

Whereas both agglomerative and partitive clustering algorithm have been successfully applied to the segmentation of SOM [9], several specific approaches have been proposed to take into account the topological ordering of the unit maps. They rely on either the contiguity study [10] or the U-matrix (the matrix of distances between adjacent map units) [11, 12, 13]. We adopted the *kmeans* based approach proposed by J. Vesanto [9]. Although the number of clusters is needed to run the *kmeans* algorithm, it is not known in the unsupervised learning framework. So several values should be tried and the best one according to the Davies-Bouldin index [4] is selected. Assuming that C, $S_c(k)$ and $d_{ce}(k,l)$ respectively refers to the number of clusters, the mean quantization error in cluster k and the distance between the centers of clusters k and l, the Davies-Bouldin index is defined by

$$I_{DB} = \frac{1}{C} \sum_{k=1}^{C} \max_{l \neq k} \left\{ \frac{(S_c(k) + S_c(l))}{d_{ce}(k,l)} \right\}$$

It should be noticed that the kmeans algorithm is a special case of the SOM training algorithm when no neighborhood constraints are imposed to the center. In other words, the neighborhood function $h_{b_i j}$ is replaced by the chronecker symbol $\delta_{b_i j}$.

3 Feature Selection

Feature Selection necessitates three essential elements [14]:

- A pertinence measure
- A search procedure
- A stop criterion

3.1 Pertinence measure

Whereas in the supervised learning case, a pertinence measure can be easily defines using the performance of the model in the task it has been designed to, in the unsupervised learning framework, it is not possible anymore.
So we have to define new criteria. We propose to use two different feature evaluation criteria : an individual criteria, $R_{individual}(j)$, to guide the search procedure and a collective criteria, $R_{collective}(j)$, to take the mutual relevance of features.
We propose to select features that involve a good clustering; thus, the SOM is segmented using the method presented above and the test-values [6] of each feature according each cluster are computed. Therefore, the maximum of absolute test values along the the different clusters is used as an individual relevance measure. The first individual relevance criteria is defined by

$$R_{individual}(j) \quad = \quad \max_{k=1,\ldots,C} \left\{ \left| \frac{\mu_{kj} - \mu_j}{\sigma_{kj}} \right| \right\} \qquad (3)$$

where C, μ_j, μ_{kj} and σ_{kj} are respectively the number of clusters, the mean values of the feature j in the whole data set and in the cluster k, and the standard deviation of feature j in the cluster k.
Then, whenever the removing of a feature involves an increasing of the I_{DB}, we consider that it is relevant according the current clustering. Thus, we define the collective relevance of a feature as the increasing of the I_{DB} involved by its removing :

$$R_{collective}(j) \quad = \quad I_{DB} - I_{DB}|_{F_R \setminus \{j\}} \qquad (4)$$

where $I_{DB}|_{F_R \setminus \{j\}}$ is the Davies-Bouldin index evaluated without taking in account the feature j.
Whereas these criteria have been successfully apply to several data set from UCI [15], they present some drawbacks. On the one hand, they rely on the kmeans algorithm which is well known for its strong dependance with the initial centers. So, to insure the reliability of the result several running of the algorithm have to be done at each step of the feature selection procedure and for each possible number of clusters. On the other hand, when many features are noisy or irrelevant, they may prevent kmeans algorithm and Davies-Bouldin to identified the right clusters; therefore the feature selection procedure might fail. Two other criteria which avoid the additional computational cost due to the map segmentation and the possible weak of robustness of the above criteria are presented in the next paragraph.

3.2 Search procedure

To find an optimal solution requires either an exhaustive search or the monotonicity of the pertinence measure. On the ones hand, the former involves the pertinence evaluation of 2^n subsets where n is the number of features and it becomes infeasible since n is large. On the other hand, the latter is difficult to insure. We propose a Backward Elimination procedure that takes into account both the individual and the collective pertinence measures defined in the previous section. It begins with the whole features set and progressively eliminates the less interesting features. The individual measure guides the selection and the collective pertinence insures that the removing of the feature candidate do not alter the quality of the model. The threshold θ in the algorithm 1 is used to balance the relative importance of the two pertinence measures.

Algorithm 1 Feature Selection Procedure

```
F_R ← F
while (¬stopping criterion) do
    Build a model.
    Evaluate individual relevance R_individual(j)
    Sort features according ascending individual relevance ordering
    found ← false
    while (¬found) do
        Evaluate the collective criterion R_collective(j) of the less relevant feature according individual criterion
        if (R_collective(j) ≤ θ) then
            found ← true
            R ← R \ {j}
        end if
    end while
    if (¬found) then
        j ← arg min_{k∈R} {R_collective(k)}
        R ← R \ {j}
    end if
end while
```

3.3 Stop criterion

We use the statistic criterion proposed by T. Cibas [16] to evaluate whether a feature subset gives any additionnal information according another one. Therefore, the backward elimination procedure is stopped since the removing of the feature selected involves a loss of information.
Assuming that F, the set of features, and $F \setminus F_R$, the removed features subset, are distributed according a gaussian law

$$N(\mu^{(k)}, \Sigma) \ : \ k = 1, \ldots, C$$

where $\mu^{(k)}$, the mean of the features from F in the cluster

k, and Σ, the covariance matrices, are defined as follows

$$\mu^{(k)} = \left(\mu_1^{(k)}, \mu_2^{(k)}\right), \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

where 1 and 2 as index respectively stand for F_R and $F \setminus F_R$. Then, the null hypothesis which says that $F \setminus F_R$ does not give any additionnal information than F_R is expressed as follows :

$$H_0 : \mu_2^{(k)} - \mu_2^{(h)} - \Sigma_{21}\Sigma_{11}^{-1}\left(\mu_1^{(k)} - \mu_1^{(h)}\right) = 0 \qquad (5)$$

with $k \neq h = 1, \ldots, C$.
A test of this hypothesis is based on Wilks statistics. Let B and W be respectively the between and the within covariance matrices :

$$\begin{aligned} B &= \sum_{k=1}^{C} N^{(k)} \left(\mu^{(k)} - \overline{\mu}\right)\left(\mu^{(k)} - \overline{\mu}\right)^T \\ W &= \sum_{k=1}^{C}\sum_{i=1}^{N^{(k)}} \left(x_i^{(k)} - \mu^{(k)}\right)\left(x_i^{(k)} - \mu^{(k)}\right)^T \end{aligned}$$

where $N^{(k)}$ is the number of elements in the cluster k and $\overline{\mu}$ is the mean of the features from F for the whole sample. Then, the same block decomposition as for Σ can be applied to the matrices B, W and their sum T :

$$\begin{aligned} B &= \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} \\ W &= \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix} \\ T &= B + W = \begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix} \end{aligned}$$

Therefore, the determinants of the matrices W and T can be written

$$\begin{aligned} |W| &= |W_{11}|\left|W_{22} - W_{21}W_{11}^{-1}W_{12}\right| \\ |T| &= |T_{11}|\left|T_{22} - T_{21}T_{11}^{-1}T_{12}\right| \end{aligned}$$

Thus, we denote

$$K = \frac{\left|W_{22} - W_{21}W_{11}^{-1}W_{12}\right|}{\left|T_{22} - T_{21}T_{11}^{-1}T_{12}\right|}$$

which has $\frac{(N-C-r)}{(C-1)}$ degrees of freedom. With the above notations, the Wilks statistics for n variables are :

$$\begin{aligned} \Lambda_F &= \frac{|W|}{|T|} \\ &= K.\frac{|W_{11}|}{|T_{11}|} \\ &= K.\Lambda_{F_R} \end{aligned}$$

which shows that, with a small value of K, the clusters separability is larger with n than r features. Therefore, the null hypothesis (5) is true if and if only features from F_R involve the same separability as the whole features set F. Then, the Wilks statistic Λ is equivalent to the Fisher-Snedecor one :

$$F_s = \frac{(N-C-r)}{(C-1)}\frac{1-K}{K}$$

which is distributed according $F\left(C-1, N-C-r\right)$

4 Experiments and results

The method presented above has been apply to several commonly used UCI machine learning data sets [15]. Whereas the data labels havent't been used during the learning stage, they can be used for evaluation purpose; actually, the ability of our approach to identified the true clusters can be measured using the following criterion :

- the number of identified clusters refered by C_T
- the couple error which measures how far the discovered partition is from the *true* classes and is defined by $E_C = \frac{2}{N(N-1)}\sum_{(i,j)\in\{1,\ldots,N\}^2,\ i<j}\epsilon_{ij}$ where ϵ_{ij} is null when samples points i and j are either grouped or separated in both true and discovered partitions.
- the Purity of clusters in term of known classes $P_R = \frac{1}{N}\sum_{k=1}^{C_T} maxM_k$ where M is the confusion matrix.

In our experiments, we used the *batch* Kohonen's algorithm and the *fast global k-means* algorithm [17] which are both deterministic. For each of the data sets considered, we run five 10-folds validation and we summarized the results obtained in Table 1. Then, the figure 4 shows the evolution of the Davie-Bouldin index during the feature selection process. The last model index value can be considered as an outlier, therefore, the best model according to the Davies-Bouldin index is obtained when five features have been removed. Nevertheless, our stop criterion indicates that the model with eleven removed features should be retained.

5 Conclusion

A new approach to select both the number of clusters and the related features subset has been proposed in an unsupervised learning framework. Whereas the preliminary results are encourageous, the stop criterion proposed can not always be uses. For instance, it requires that $N - c \geq p$, where N, c and p are respectively the number of map units, the number of identified clusters and the total number of features, to insure that the within covariance matrix W is not singular. Research work are on the way to enhance the proposed method to data sets with more features than observations.

				Training set		Testing set	
		C_T [σ_{C_T}]	n_{FS} [$\sigma_{n_{FS}}$]	E_C [σ_{E_C}]	P_R [σ_{P_R}]	E_C [σ_{E_C}]	P_R [σ_{P_R}]
Glass	F	7.04 [0.73]	9.0 [–]	0.301 [0.012]	56.25 [2.56]	0.295 [0.068]	67.52 [9.01]
189 - 21	F_R	5.10 [1.83]	2.84 [1.46]	0.376 [0.082]	50.83 [6.54]	0.382 [0.121]	58.38 [10.40]
Wine	F	6.86 [0.81]	13.0 [–]	0.171 [0.022]	93.59 [1.97]	0.165 [0.064]	95.28 [5.11]
189 - 21	F_R	5.70 [2.34]	6.3 [2.1]	0.247 [0.060]	80.32 [12.02]	0.239 [0.096]	83.44 [13.78]
Cancer	F	9.72 [0.67]	30.0 [–]	0.414 [0.014]	93.83 [1.56]	0.417 [0.026]	94.16 [3.03]
242 - 27	F_R	2.72 [1.96]	12.4 [3.3]	0.182 [0.077]	91.53 [1.04]	0.184 [0.091]	91.60 [3.49]
Wave	F	6.18 [2.56]	40.0 [–]	0.304 [0.016]	68.64 [8.48]	0.309 [0.014]	66.17 [7.82]
500 - 4500	F_R	4.82 [1.55]	28.2 [9.56]	0.304 [0.020]	66.93 [6.62]	0.306 [0.018]	65.97 [6.68]

Table 1. The two numbers under the data set name indicates the size of the training and testing sets respectively. Then F and F_R stands for the whole features set and the reduced subset selected.

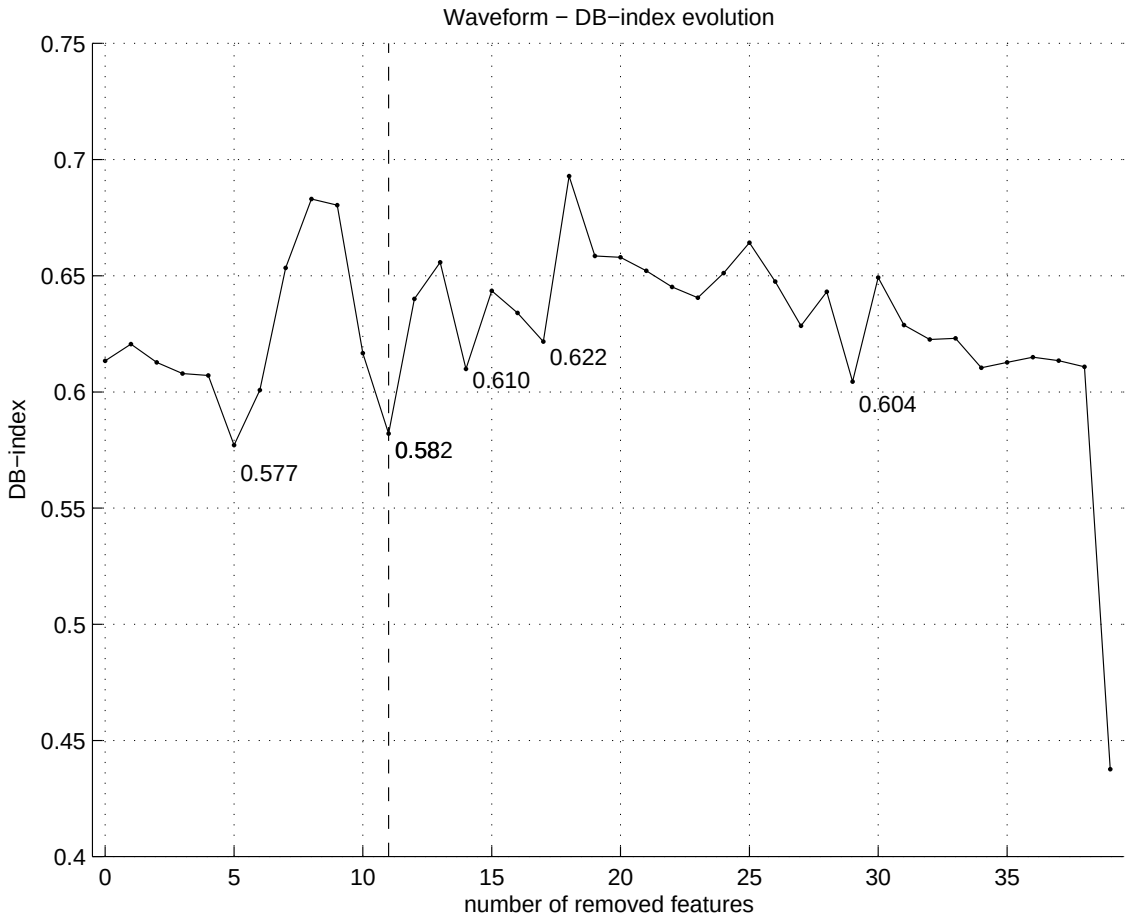

Figure 2. Evolution of the Davies-Bouldin index during the backward features elimination procedure : the vertical dash line indicates the model retained by our stop criterion and some of the best index values are indicated too.

References

[1] L. Lebart, A. Morineau, and M. Piron. *Statistique exploratoire multidimensionnelle*. Éditions Dunod, 1995.

[2] G. Saporta. *Probabilités, analyse des données et statistique*. Editions Technip, Paris, France, 1990.

[3] S. K. Pal, R. K. De, and J. Basak. Unsupervised feature evaluation: A neuro-fuzzy approach. *IEEE Transactions on Neural Networks*, 11(2):366–376, 2000.

[4] D. Davies and D. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 1(2):224–227, 1979.

[5] J. Vesanto and E. Alhoniemi. Clustering of the self-organizing map. *IEEE-NN*, 11(3):586–600, May 2000.

[6] A. Morineau. Note sur la caractérisation statistique d'une classe et les valeurs-tests, 1984.

[7] J-C. Fort, P. Letrémy, and M. Cottrell. Advantages and drawbacks of the batch kohonen algorithm. In M.Verleysen Ed., editor, *ESANN'2002 Proceedings, European Symposium on Artificial Neural Networks, Bruges (Belgium)*, pages 223–230, Bruxelles, Belgium, 2002. Editions D Facto.

[8] T. Kohonen. *Self-Organizing Maps*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2001.

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

[10] F. Murtagh. Interpreting the Kohonen self-organizing feature map using contiguity-constrained clustering. *Pattern Recognition Letters*, 16(4):399–408, April 1995.

[11] F. Moutarde and A. Ultsch. U*F clustering: a new performant cluster-mining method based on segmentation of Self-Organizing Maps. In *Proceedings of the 5th Workshop On Self-Organizing Maps (WSOM'05)*, pages 25–32, Paris 1 Panthéon-Sorbonne University, France, September 2005.

[12] D. Opolon and F. Moutarde. Fast semi-automatic segmentation algorithm for Self-Organizing Maps. In *Proceedings of ESANN'2004 , European Symposium on Artificial Neural Networks, Bruges (Belgium)*, pages 507–512, 2004.

[13] A. Ultsch. Clustering with SOM: U*C. In *Proceedings of the 5th Workshop On Self-Organizing Maps (WSOM'05)*, pages 75–82, Paris 1 Panthéon-Sorbonne University, France, September 2005.

[14] D. Cakmakov and Y. Bennani. *Feature Selection for Pattern Recognition*. Informa, Skopje, Macedonia, 2002.

[15] C.L. Blake D.J. Newman, S. Hettich and C.J. Merz. UCI repository of machine learning databases, 1998.

[16] T. Cibas. *Contrôle de la complexité dans les réseaux de neurones : régularisation et sélection de caractéristiques*. PhD thesis, University of Paris XI Orsay, Paris, France, December 1996.

[17] J. J. Verbeek. *Mixture Models for Clustering and Dimension Reduction*. PhD thesis, University of Amsterdam, Amsterdam, The Netherlands, December 2004.