

Unsupervised Variable Selection: when random rankings sound as irrelevancy

Sébastien Guérif

Université Paris 13 - LIPN / CNRS UMR 7030, F-93430 Villetaneuse, France,
sebastien.guerif@lipn.univ-paris13.fr,
WWW home page: <http://www-lipn.univ-paris13.fr/~guerif>

Abstract. Whereas the variable selection has been extensively studied in the context of supervised learning, the unsupervised variable selection has attracted attention of researchers more recently as the available amount of unlabeled data has exploded. Many unsupervised variable ranking criteria were proposed and their relevance is usually demonstrated using either external cluster validity indexes or the accuracy of a classifier which are both supervised criterion. Actually, the major issue of the variable subset selection according to a ranking measure has been addressed only by few authors in the unsupervised learning context. In this paper, we propose to combine multiple ranking to go ahead toward a stable consensus variable subset in a totally unsupervised fashion.

1 Introduction

The amount of available data has exploded as the storage technologies were improved. The companies databases gathered a lot of information which are not all relevant for their exploitation by computational methods. It has been proved that dimension reduction of the data generally leads to performance improvement: this process can be applied either to the sample size, to their description or even to both dimension. In this paper, we are interested in the reduction of the description space dimension. Two kind of approaches can be considered to reduce the dimension of the sample points description, namely, feature extraction and feature selection. The latter consists in combining the original dimension to construct new features while the last reduces the dimension of the input space by dropping some irrelevant or weakly relevant features.

The most widely used feature extraction techniques is probably the Principal Component Analysis (PCA) [1] which builds new uncorrelated dimensions, called factors, by maximising the variance. Another linear technique is the Multidimensional Scaling (MDS) [1] which starts with a dissimilarity matrix and builds a representation in an Euclidian space by maximising the preservation of the distances between sample points. Some non linear methods have also been proposed such as the Isometric Mapping (Isomap) [2] or the Locally Linear Embedded (LLE) [3]. The efficiency of these methods have been demonstrated on a large range of application domain, but the interpretation of the new features is not obvious and an important effort is required from the user. This paper deals

with feature selection which has the advantage of keeping the sample points in a subspace of the original space. Hence, the reduced description is directly understandable and does not require any additional interpretation work.

Although feature selection has been extensively studied in the context of supervised learning, this field is relatively new in the unsupervised learning. Several evaluation measures have been proposed but the authors rarely present a full selection procedure; actually, the definition of a stopping criterion is not often considered. In this paper, a general methodology is proposed for variables ranking based feature selection. The approach proposed eliminates the irrelevant features but does not remove the redundant features whereas they are only weakly relevant.

The rest of the paper is organised as follows. Section 2 gives a brief overview of related works. Section 3 presents the approach proposed. Section 4 reports our experimental results. Finally, section 5 concludes.

2 Related work

Feature selection consists in choosing one feature subset among all the possible one and a feature selection procedure is composed from essentially in three elements: an evaluation measure, a subset generator and a stopping criterion [4–6]. Three categories of methods are generally distinguished: the filter, the wrapper and the embedded approaches [7]. The filter approaches only take into account the properties of the data-set independently of the future use of the data. On the contrary, in the wrapper approaches, the performance of the algorithm which uses the data are considered either to define the evaluation measure, to guide the search procedure or to define a stopping criterion. In the last kind of techniques, the procedure is embedded in the algorithm used to build the model. A deeper introduction to the feature and variable selection can be found in [4], [5] or [6]. The feature selection problem has been extensively studied in the context of supervised learning but relatively few works have been published in the unsupervised context. A brief overview of the domain is given by classifying the approaches according the following taxonomy: redundancy based, entropy based, auto-associative model based and clustering based. The reader interested in unsupervised feature selection for text data should refer to [8] and [9].

Redundancy based approaches

A broad part of the methods which proposed to achieve feature selection in the unsupervised context try to eliminate the redundancy among a feature subset and thus they rely either on correlation or an estimation mutual information [10–12]. Actually, P. Mitra et al. used a similarity measure that corresponds to the lowest eigenvalue of correlation matrix between two features [11], J. Vesanto et al. proposed to visually detect correlation using a SOM-based approach [12] and S. Guérif and al. used a similar idea and integrated a weighting mechanism in the SOM training algorithm to reduce the redundancy side effects [10].

Auto-associative model based approaches

Some authors reduce the unsupervised case issue to the supervised one by using auto-associative models: supervised selection approaches can be applied with regression techniques and furnish us with an unsupervised method. For instance, the auto-associative regression trees (ART) have been used in [13].

Entropy based approaches

Assuming that the relevant features are those which lead to clusters in a data-set, an entropy like criteria can be used as evaluation measure: a uniformly distributed feature does not provide any useful information for clustering and on the contrary, feature that gathers the sample points in small group are relevant. Actually, assuming that $s(x, y)$ is the similarity between two sample points x and y , an entropy measure can be defined as [14, 15]:

$$H = \sum_{(x,y) \in X^2} (s(x, y) \log s(x, y)) + (1 - s(x, y)) \log(1 - s(x, y)) \quad (1)$$

where X is the data-set. This measure is used in combination with a sequential backward elimination procedure in [14] or with a sequential forward selection procedure in [15]. It has also been exploited by the neuro-fuzzy approach proposed in [16].

Clustering based approaches

Clustering quality assessment can be used either as a subset evaluation measure [17] or as a stopping criterion [15, 18, 19]. Actually, a Davies-Bouldin index based evaluation measure was proposed in [17] and assuming that the features are normally distributed, the authors used the Λ Wilks statistic to stop their backward elimination procedure since the separability of the clusters decreases significantly. In [15, 18, 19], the scatter matrices and separability criteria used in multiple discriminant analysis were used by the authors to stop their feature selection process. Model-based clustering usually estimates the probability distribution of the data; this gives an additional insight of the data. In other respects, the maximum likelihood criterion measures how likely a model fits the data. Hence, select a feature subset amounts to select a model [18–21]. The authors of [18, 19] pointed out that both the scatter separability and the maximum likelihood criteria are biased with the respect of the number of features and they provided a normalisation term to correct this bias. The subspace clustering can be viewed as a related issue and [22] provides an overview of this topic.

3 Proposed approach

The approach proposed relies on two reasonable assumptions. On the one hand, since a scoring function is consistent with the addressed problem, the irrelevant

features are ranked at the end. On the other hand, we assume the existence of a partition such that the features are gathered in equivalence classes in respect with an evaluation measure; actually, the ordering defined on each of the equivalence classes by the scoring function considered is nothing more than an artefact. Hence, we are interested in detecting only the equivalence class which gathers the irrelevant features. This can be achieved by identifying the rank from which the variables are randomly ranked. According to the definition of the equivalence classes, while a set of features gathers variables from more than one class, it can be (at least partially) ordered. Now, rank the features according their decreasing relevance, as soon as the set of the last features can not be properly ordered, it gathers only features from the same equivalence class: the irrelevant features one.

We propose to compute multiple ranking of a variable set using resampling technique and to identify a subset of the best features with respect to the different rankings; a user defined parameter controls the minimal agreement. Actually, we assume that the irrelevant features equivalence class can not be properly ordered. Thus, when the irrelevant feature equivalence class begins at rank k , there is no reason for one irrelevant feature to be preferably ranked at any of the last ranks. This leads to assume that the distribution of the features at each of the remaining the k^{th} rank is uniform. Our null hypothesis H_0 can be formulated as follows: "at the k^{th} rank the variables are uniformly distributed". Thus, a χ^2 test can be used to test the agreement between the empirical distribution of the variables at the k^{th} rank and the theoretical distribution under a uniform random distribution.

3.1 Notations

Let $F = \{1, \dots, n\}$ be a set of variables and $S = \{s_i : i = 1, \dots, p\}$ be a set of p scoring functions $s_i : F \rightarrow \mathbb{R}$. The scoring function s_i on F defines a total order on F and the corresponding permutation is noted by σ_i while the set of the permutations induced by S is $R = \{\sigma_i : i = 1, \dots, p\}$. Let $\sigma_i|_{1:k}$ and $R|_{1:k} = \{\sigma_i|_{1:k}\}$ be respectively the k first values of the permutation σ_i and the set of the restricted permutations. Let $\hat{c}_j|_k$ denotes the number of permutations $\sigma_i \in R$ such that the variable j appears at the k^{th} rank and let $c_j|_k$ be its expectation when only the $(k-1)$ first values of each permutation are known.

3.2 Theoretical distribution

Assuming that only the $(k-1)$ first values of each permutation are known, and that the $(n-k+1)$ last values are obtained independently at random, $c_j|_k$ follows a Bernoulli distribution and can be easily computed as:

$$c_j|_k = \sum_{l=1}^{p_{kj}} \binom{p_{kj}}{l} \times l \times (q_k)^l \times (1 - q_k)^{p_{kj}-l} \quad (2)$$

where p_k and q_k are defined as:

$$p_{kj} = p - \sum_{i=1}^p \sum_{l=1}^{k-1} \hat{c}_j|l \quad (3)$$

$$q_k = \frac{1}{n - k + 1} \quad (4)$$

and refer to the number of permutations for which j does not appear before the k^{th} rank and the probability such j appears at the k^{th} rank (if has not appeared before) respectively.

3.3 χ^2 -test

The deviation of the empirical values from the theoretical distribution is measured using the χ^2 statistic defined as:

$$\chi_{n-1}^2 = \sum_{j=1}^n \frac{(\hat{c}_j|k - c_j|k)^2}{c_j|k} \quad (5)$$

Under the null hypothesis H_0 , the following expression follows a Laplace-Gauss distribution $LG(0, 1)$:

$$\left[\left(\frac{\chi_{n-1}^2}{n-1} \right)^{\frac{1}{3}} + \frac{2}{9(n-1)} - 1 \right] \sqrt{\frac{9(n-1)}{2}} \quad (6)$$

since $(n-1) > 100$. Thus, the χ^2 test amounts to test whether the expression (6) exceeds the critical value of the unilateral test for a Laplace-Gauss distribution. Assuming that a false negative rate of 5% (probability of a Type II error) is acceptable, the threshold $\theta = 1.65$ would be used for the expression (6).

3.4 Subset selection

Let \tilde{k} be the last rank for which the expression (6) is lower than the threshold θ . The subset selected by the method proposed is defined as follows:

$$\tilde{F}_\alpha = \left\{ j \in F : \alpha \cdot p \leq \sum_{k=1}^{\tilde{k}} c_j|k \right\} \quad (7)$$

where the $\alpha \in [0, 1]$ parameter controls the required degree of agreement between the different rankings.

3.5 Algorithm and complexity

The following high level algorithm summarised the different steps of the method proposed:

1. Compute the set of scoring functions S using p independent sub-samples of the data-set: $\theta(p.m)$
2. Build the set of rankings R induced by S : $\theta(p.n.\log_2 n)$
3. For each variable j and each rank k , compute $\hat{c}_j|_k$ the number of permutations such that j appears at the k^{th} rank: $\theta(p.n)$
4. Compute the binomial coefficients: $\theta(p^2)$
5. Compute $c_j|_k$ the expected values of $\hat{c}_j|_k$ under the null hypothesis: $\theta(p.n^2)$
6. Compute the χ^2 statistics: $\theta(n^2)$

The time complexity of each step is indicated above and the overall time complexity of the method proposed is $\theta(p(n^2 + m) + p^2)$ where m , n and p are respectively the complexity of the scoring function evaluation, the number of variables and the number of scoring function evaluations.

4 Experimental results

4.1 Data-sets

Here, we demonstrate the effectiveness of the stop criterion presented above on four data-sets, namely, Arcene, Gisette, Madelon and Waveform. They are all available at the UCI Machine Learning Repository [23]. The Arcene, Gisette and Madelon data-sets were originally proposed during the Feature Selection Challenge organised during NIPS 2003 [24] and they point out different difficulties. Actually, Arcene data-set illustrates the case where the number of available sample points is small with respect to the data dimension; this problem arises frequently in real application such text mining, mass spectrometry or bioinformatic. Madelon data-set was artificially constructed to emphasised the difficulty of the feature selection when no feature is informative by itself. Waveform data-set illustrates the case where the different classes largely overlap. Table 1 gathers some the information about these data-sets. The *probes* are artificial features which were added to facilitate the assessment of the feature selection methods; they were drawn at random from a distribution resembling that of the real features except for the waveform data-set were they are normally distributed.

Table 1. Data-sets used in our experiments: N is the sample size, n is the number of features, n_{probe} (resp. $\%_{probe}$) is the number (resp. rate) of probes and C is the number of classes.

Data-set	Domain	N	n	n_{probe}	$\%_{probe}$	C
Arcene	Mass spectrometry	200	10000	3000	30.0 %	2
Gisette	Digit recognition	7000	5000	2500	50.0 %	2
Madelon	Artificial	2600	500	480	96.0 %	2
Waveform	Artificial	5000	40	19	47.5 %	3

4.2 Laplacian Score

In our experiments, the Laplacian Score [25] was retained as pertinence criteria; it measures the locality preserving power of the features. The steps of the algorithm given by [25] are the following:

1. Construct the k nearest neighbours graph G : $\theta(n.N^2)$.
2. If nodes i and j are connected in G , put $S_{ij} = e^{-(d_{ij}/t)^2}$ where d_{ij} is the distance between sample points i and j , and t is a suitable constant. Otherwise, put $S_{ij} = 0$: $\theta(k.N)$.
3. Let f_i be the column vector of the i^{th} feature values and $L = D - S$ be the Laplacian of the graph G where $D = \text{diag}(S \cdot \mathbf{1})$. Compute the Laplacian Score L_i of the i^{th} feature as

$$L_i = \frac{\tilde{f}_i^T L \tilde{f}_i}{\tilde{f}_i^T D \tilde{f}_i} \quad (8)$$

$$\text{where } \tilde{f}_i = f_i - \frac{f_i^T D \mathbf{1}}{\mathbf{1}^T D \mathbf{1}} \mathbf{1} \quad (9)$$

This step requires $\theta(n.N^2)$ operations.

The time complexity of each step is indicated above and the overall time complexity of the method proposed is $\theta(n.N^2)$ where n and N are respectively the number of variables and the number of sample points. In our experiments, the number of neighbours k was set to 5 and the constant t was chosen as the maximum distance between to connected sample points.

4.3 Evaluation methodology

The labels of the sample points from the data-sets used in our experiments were available and the scoring functions chosen aims to preserve the local topology of the data space; thus the accuracy of the k -nearest neighbours classifiers was used as performance measure of the method proposed. The stability of the subset selected is measured using the Jaccard and the Rand statistics averaged on the subset selected pairs set. A 10-fold cross-validation approach was used to evaluate the accuracy of a k -nearest neighbours classifier. The original sample was randomly partitioned into 10 sub-samples. Of the 10 sub-samples, a single sub-sample is retained as the validation data for evaluating the performance of the system, and the remaining 9 sub-samples are used as training data. The cross-validation process is then repeated 10 times, with each of the 10 sub-samples used exactly once as the validation data. The 10 results from the folds then are averaged.

In our experiments, we estimated the scoring function on $p = 100$ independent random sub-samples of the training subset with replacement. The α control parameter was set to 0.5. The scoring function was computed using only sub-samples of the training set and r denotes the ratio between the number of

Table 2. Stability of the subset selected: r is the ratio between the number of sample points and the number of features, the values indicated are averaged over the 10 folds and the standard deviation are given between brackets.

Data-set	Method	Proposed	1-NN	3-NN	5-NN
Arcene ($r = 0.018$)	Jaccard	0.957 [0.008]	0.743 [0.180]	0.684 [0.192]	0.671 [0.181]
	Rand	0.958 [0.005]	0.758 [0.176]	0.708 [0.180]	0.696 [0.174]
Arcene ($r = 0.01$)	Jaccard	0.894 [0.012]	0.645 [0.167]	0.638 [0.160]	0.659 [0.190]
	Rand	0.936 [0.008]	0.687 [0.155]	0.987 [0.148]	0.684 [0.181]
Gisette ($r = 1$)	Jaccard	0.911 [0.018]	0.787 [0.115]	0.806 [0.124]	0.782 [0.137]
	Rand	0.949 [0.011]	0.853 [0.101]	0.878 [0.087]	0.858 [0.100]
Gisette ($r = 0.1$)	Jaccard	0.976 [0.014]	0.709 [0.124]	0.650 [0.144]	0.609 [0.133]
	Rand	0.977 [0.013]	0.789 [0.096]	0.705 [0.135]	0.697 [0.137]
Madelon ($r = 1$)	Jaccard	0.999 [0.002]	0.991 [0.010]	0.986 [0.015]	0.996 [0.003]
	Rand	0.999 [0.002]	0.991 [0.010]	0.987 [0.014]	0.996 [0.003]
Madelon ($r = 0.1$)	Jaccard	0.986 [0.007]	0.969 [0.011]	0.967 [0.011]	0.966 [0.013]
	Rand	0.986 [0.007]	0.970 [0.010]	0.968 [0.011]	0.967 [0.013]
Waveform ($r = 10$)	Jaccard	0.617 [0.104]	0.779 [0.125]	0.784 [0.123]	0.811 [0.113]
	Rand	0.699 [0.105]	0.861 [0.086]	0.870 [0.080]	0.887 [0.073]
Waveform ($r = 1$)	Jaccard	0.638 [0.142]	0.728 [0.110]	0.681 [0.152]	0.752 [0.117]
	Rand	0.765 [0.109]	0.834 [0.082]	0.798 [0.112]	0.855 [0.077]

features over the sub-sample size in the following. To evaluate the robustness of our method to the small sample size issue, we use two different values of the r parameter. For a given data-set, the lower is r , the more difficult is the relevance estimation.

The results achieved by the unsupervised method proposed are then compared with those of the best k -nearest neighbours classifiers that can be obtained given the set of rankings R and the value of the α parameter (see equation 7); they are referred as "supervised (k -NN)".

4.4 Results and discussion

Figure 1 compares the size of the original feature subset (before adding probes), the size of the feature subset selected by the method presented above and the size of the best subsets according the performance of the 1-NN, 3-NN and 5-NN classifiers. In addition, table 2 indicates the averaged agreement between the subset selected and the collection of the top- k feature subsets. Figure 2 summarises the error rate of the 1-NN classifiers using the whole set of features, the subset selected by the method presented above and the subset that leads to the best classifier; the performance of the 3-NN and 5-NN classifiers are very similar and have been disregarded due to space limitation. As shown by the figure 2, a small increasing of the 1-NN classifier error rate is observed with the

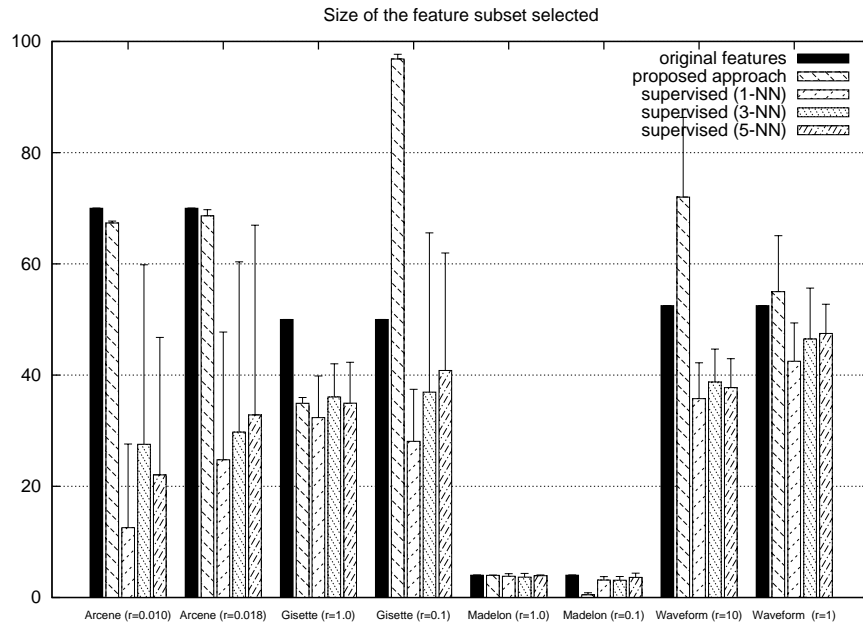


Fig. 1. Size of the feature subset selected: index 100 corresponds to the whole set of features (original features + probes), r is the ratio between the number of sample points and the number of features, the values indicated are averaged over the 10 folds and the standard deviation are given between brackets. The supervised method refers to the choice of the classifier with the lowest error rate.

Madelon data-set when there are too few sample points to estimate the scoring function. Actually, few highly redundant features were identified as relevant by the Laplacian Scores; hence, an important part of the relevant information is lost (see fig. 1). On the contrary, the selection does not operate on the Gisette data-set when too few sample points are available to compute the scoring function.

The proposed method does not perform as well as the supervised criterion with the Arcene and the Waveform data-sets whereas most of the noisy dimensions are removed. Although no significant improvements in respect with the baseline (no feature selection) are observed with the Arcene data-set, all the probes are eliminated. With the Gisette and the Madelon data-sets, since the sample size is sufficient to estimate the scoring function, the accuracy of the 1-NN using the unsupervised method proposed are comparable with those obtained by the supervised criteria.

Anyway, the stability of the subset selected by the method proposed is emphasised by the table 2: the subset selected by the supervised method vary a lot among the different folds. The resampling causes fluctuations of the laplacian scores which might account for the greater stability of the method proposed

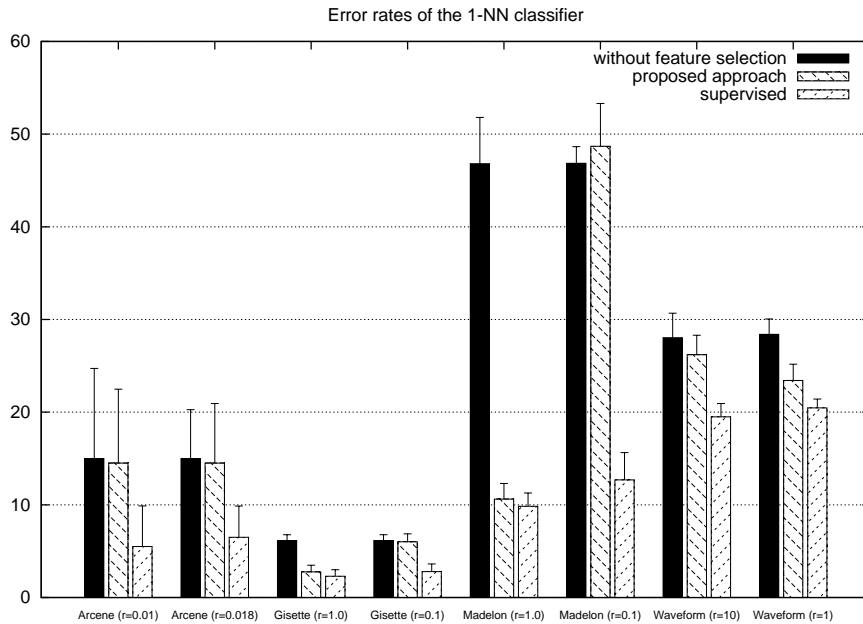


Fig. 2. Error rates of the 1-NN classifiers: r refer to the ratio between the number of sample points and the number of features, the values indicated are averaged over the 10 folds. The supervised method refers to the choice of the classifier with the lowest error rate.

against the supervised approach. On the one hand, the supervised method operates separately for each scoring function evaluation. On the other hand, the approach proposed operates globally which improves its robustness. Then, the lowest values of the Jaccard and Rand indexes are observed with the Waveform data-sets where the results can be much improved; this leads us to assume that the stability of the subset selected can be thought as a potential indication to predict whether the feature selection correctly operates or not. Obviously, this point should be further investigated before to conclude.

The approach proposed relies on resampling and the detection of randomness. It should be noticed that the use of subsampling methods involves a high computational overload. Nevertheless, the computation needed by our approach can be easily distributed to overcome this problem: (i) the different scoring function evaluations are independent from each other, (ii) the expected number of permutations such the variable j appears at the k^{th} does not depends on any other variable and (iii) the χ^2 statistics can be computed separately for each rank k .

5 Conclusion and perspectives

In this paper, we propose a general method to address the subset selection issue when only a scoring function ranks the features according their relevancy is available. The approach proposed relies on resampling and the detection of randomness. It should be noticed that the use of subsampling methods involves a high computational overload, but the computation needed by our method can be easily distributed to overcome this problem.

The effectiveness of the method and the stability of the subset selected have been demonstrated on 4 data-sets which span at least three major issues of feature selection: small sample size with respect to the data dimension, lack of feature informative by itself and classes overlapping. The experiments presented above were done in the unsupervised learning context but obviously it applies in the semi-supervised or supervised context as soon as the additional information available is included in the scoring function; we suppose that this can lead to robustness and stability improvement. The results presented point out that the method proposed does not perform well when the sub-sample size used to compute the scores is too small, namely lower than 100. We feel that this problem might be corrected by increasing the number of scores computed.

Future work includes a rigorous study of the relationships between the data dimension n , the number of sub-samples used p and the size of the sample used to estimate the relevance of the features; anyway, we feel that these relationships depend on the scoring function selected and we plan to repeat the experiments presented using the weighting coefficients computed by the AVW-k-means algorithm [26] and by the ω -SOM algorithms [27].

References

1. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification (2nd Edition). Wiley-Interscience (2001)
2. Tanenbaum, J.B., de Silva, V., Langford, J.C.: A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science* **290** (Dec 2000) 2319–2323
3. Roweis, S.T., Saul, L.K.: Nonlinear Dimensionality Reduction by Local Linear Embedding. *Science* **290** (Dec 2000) 2323–2326
4. Blum, A., Langley, P.: Selection of relevant features and examples in machine learning. *Artificial Intelligence* **97**(1-2) (1997) 245–271
5. Dash, M., Liu, H.: Feature selection for classification. *Intelligent Data Analysis* **1** (1997) 131–156
6. Guyon, I., Elisseeff, A.: An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research* **3** (2003) 1157–1182
7. Kohavi, R., John, G.H.: Wrappers for feature subset selection. *Artificial Intelligence* **97**(1-2) (1997) 273–324
8. Liu, L., Kang, J., Yu, J., Wang, Z.: A comparative study on unsupervised feature selection methods for text clustering. In: Proceedings of 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering. (2005) 597–601
9. Wiratunga, N., Lothian, R., Massie, S.: Unsupervised feature selection for text data. In Roth-Berghofer, T., Göker, M.H., Güvenir, H.A., eds.: ECCBR. Volume 4106 of Lecture Notes in Computer Science., Springer (2006) 340–354

10. Guérif, S., Bennani, Y., Janvier, E.: μ -som : Weighting features during clustering. In: Proceedings of the 5th Workshop On Self-Organizing Maps (WSOM'05). (Sep 2005) 397–404
11. Mitra, P., Murthy, C., Pal, S.: Unsupervised Feature Selection Using Feature Similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**(4) (2002)
12. Vesanto, J., Ahola, J.: Hunting for Correlations in Data Using the Self-Organizing Map. In Bothe, H., Oja, E., Massad, E., Haefke, C., eds.: *Proceeding of the International ICSC Congress on Computational Intelligence Methods and Applications (CIMA '99)*, ICSC Academic Press (1999) 279–285
13. Questier, F., Put, R., Coomans, D., Walczak, B., Heyden, Y.V.: The use of CART and multivariate regression trees for supervised and unsupervised feature selection. *Chemometrics and Intelligent Laboratory Systems* **76**(1) (2005) 45–54
14. Dash, M., Liu, H., Yao, J.: Dimensionality reduction of unsupervised data. In: *ICTAI*. (1997) 532–539
15. Dash, M., Liu, H.: Feature selection for clustering. In: *PADKK '00: Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Current Issues and New Applications*, London, UK, Springer-Verlag (2000) 110–121
16. Basak, J., De, R.K., Pal, S.K.: Unsupervised feature selection using a neuro-fuzzy approach. *Pattern Recognition Letter* **19**(11) (1998) 997–1006
17. Guérif, S., Bennani, Y.: Selection of clusters number and features subset during a two-levels clustering task. In: *Proceedings of the 10th IASTED International Conference Artificial intelligence and Soft Computing 2006*. (Aug 2006) 28–33
18. Dy, J.G., Brodley, C.E.: Feature subset selection and order identification for unsupervised learning. In: *Proceedings 17th International Conference on Machine Learning*, Morgan Kaufmann, San Francisco, CA (2000) 247–254
19. Dy, J.G., Brodley, C.E.: Feature Selection for Unsupervised Learning. *Journal of Machine Learning Research* **5** (2004) 845–889
20. Law, M.H.C., Figueiredo, M.A.T., Jain, A.K.: Simultaneous feature selection and clustering using mixture models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **26**(9) (2004) 1154–1166
21. Raftery, A., Dean, N.: Variable selection for model-based clustering. *Journal of the American Statistical Association* **101** (2006) 168–178
22. Parsons, L., Haque, E., Liu, H.: Subspace clustering for high dimensional data: a review. *SIGKDD Explor. Newsl.* **6**(1) (June 2004) 90–105
23. D.J. Newman, S. Hettich, C.B., Merz, C.: *UCI repository of machine learning databases* (1998)
24. Guyon, I., Gunn, S.R., Ben-Hur, A., Dror, G.: Result analysis of the nips 2003 feature selection challenge. In: *NIPS*. (2004)
25. He, X., Cai, D., Niyogi, P.: Laplacian score for feature selection. In Weiss, Y., Schölkopf, B., Platt, J., eds.: *Advances in Neural Information Processing Systems 18*. MIT Press, Cambridge, MA (2006) 507–514
26. Huang, J.Z., Ng, M.K., Rong, H., Li, Z.: Automated variable weighting in k-means type clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27**(5) (2005) 657–668
27. Guérif, S., Bennani, Y.: Dimensionality Reduction Through Unsupervised Features Selection. In: *Proceedings of the 10th International Conference on Engineering Applications of Neural Networks (EANN2007)*, Publishing Centre Alexander T.E.I. of Thessaloniki (August 2007) 98–106