Unsupervised Variable Selection: when random rankings sound as irrelevancy

Sébastien Guérif SEBASTIEN.GUERIF@LIPN.UNIV-PARIS13.FR
LIPN / CNRS UMR 7030
Université Paris 13
F-93430 Villetaneuse, France

Editor: Saeys et al.

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 criteria. Actually, the major issue of the variable subset selection according to a ranking measure has been adressed 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 dimensions. In this paper, we are interested in the reduction of the description space dimension. Two kinds of approaches can be considered to reduce the dimension of the sample points description, namely, feature extraction and feature selection. The former consists in combining the original variables to construct new features while the latter reduces the dimension of the input space by dropping some irrelevant or weakly relevant variables.

The most widely used feature extraction technique is probably the Principal Component Analysis (PCA) (Duda et al., 2001) which builds new uncorrelated dimensions, called factors, by maximising the variance. Another linear technique is the Multidimensional Scaling (MDS) (Duda et al., 2001) 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) (Tanenbaum et al., 2000) or the Locally Linear Embedded (LLE) (Roweis and Saul, 2000). The efficiency of these methods have been demonstrated on a large range of application domains, 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 ones and a feature selection procedure is composed from essentially three elements: an evaluation measure, a subset generator and a stopping criterion (Blum and Langley, 1997; Dash and Liu, 1997; Guyon and Eliseeff, 2003). Three categories of methods are generally distinguished: the filter, the wrapper and the embedded approaches (Kohavi and John, 1997). 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 Blum and Langley (1997), Dash and Liu (1997) or Guyon and Eliseeff (2003). 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 to 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 Liu et al. (2005) and Wiratunga et al. (2006).

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 (Guérif et al., 2005; Mitra et al., 2002; Vesanto and Ahola, 1999). Actually, P. Mitra et al. (2002) used a similarity measure that corresponds to the lowest eigenvalue of correlation matrix between two features, Vesanto and Ahola (1999) proposed to visually detect correlation using a SOM-based approach and Guérif et al. (2005) used a similar idea and integrated a weighting mechanism in the SOM training algorithm to reduce the redundancy side effects.

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 Questier et al. (2005).

Entropy based approaches

Assuming that the relevant features are those which lead to clusters in a data-set, an entropy like criterion can be used as evaluation measure: a uniformly distributed feature does not provide any useful information for clustering and on the contrary, features that gather the sample points in small groups 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 (Dash et al., 1997; Dash and Liu, 2000):

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

where X is the data-set. This measure is used in combination with a sequential backward elimination procedure in Dash et al. (1997) or with a sequential forward selection procedure in Dash and Liu (2000). It has also been exploited by the neuro-fuzzy approach proposed in Basak et al. (1998).

Clustering based approaches

Clustering quality assessment can be used either as a subset evaluation measure (Guérif and Bennani, 2006) or as a stopping criterion (Dash and Liu, 2000; Dy and Brodley, 2000, 2004). Actually, a Davies-Bouldin index based evaluation measure was proposed in Guérif and Bennani (2006) 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 Dash and Liu (2000), Dy and Brodley (2000) and Dy and Brodley (2004), 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 (Dy and Brodley, 2000, 2004; Law et al., 2004; Raftery and Dean, 2006). Dy and Brodley (2000, 2004) 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 Parsons et al. (2004) 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 to 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. Hence, 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 $F = \{1, \ldots, n\}$ be a set of variables and $S = \{s_i \ : \ i = 1, \ldots, p\}$ 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. The set of the permutations induced by S is denoted by $R = \{\sigma_i \ : \ i = 1, \ldots, p\}$. Let $\sigma_i|_{1:k}$ and $R|_{1:k} = \{\sigma_i|_{1:k} \ : \ i = 1, \ldots, p\}$ 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:

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

where p_k and q_k are defined as:

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

$$q_k = \frac{1}{n-k+1} \tag{4}$$

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:

$$\chi^2_{n-1} = \sum_{j=1}^{n} \frac{(\hat{c}_j|_k - c_j|_k)^2}{c_j|_k} \tag{5}$$

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

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

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:

$$\tilde{F}_\alpha = \left\{ j \in F \;:\; \alpha.p \leq \sum_{k=1}^{\tilde{k}} c_j|_k \right\} \tag{7}$$

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 dataset: $\theta(p.m)$
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: $\theta(n^2)$

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

The time complexity of each step is indicated above and the overall time complexity of the method proposed is $\theta\left(p(n^2+m)+p^2\right)$ 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

Two assumptions are made by the method proposed: (i) the irrelevant features are ranked at the end by the scoring function, and (ii) the features can be gathered in equivalence classes such that no ordering can be properly defined on each of the equivalence classes by the scoring function. Here, we show the validity of these assumptions using two toys data-sets: Art-1 and Art-2. The Art-1 data-set is composed of 500 sample points gathered in 5 clusters which are normally distributed. They have the same size, the same covariance matrix, and their respective means are the following: $(0, 0)$, $(1,\ 1)$, $(1, -1)$, $(-1, -1)$ and $(-1, 1)$. Then, 8 normally distributed noisy dimensions have been added. The Art-2 data-set is build by replacing 2 of the noisy dimensions by noisy copies of the original variables.

Then, we demonstrate the effectiveness of the stopping criterion presented above on four data-sets, namely, Arcene, Gisette, Madelon and Waveform. They are all available at the UCI Machine Learning Repository (D.J. Newman and Merz, 1998). The Arcene, Gisette and Madelon data-sets were originally proposed during the Feature Selection Challenge organised during NIPS 2003 (Guyon et al., 2004) 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 as text mining, mass spectrometry or bioinformatic. Madelon data-set was artificially constructed to emphasise 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.

4.2 Laplacian Score

In our experiments, the Laplacian Score (He et al., 2006) was retained as pertinence criterion; it measures the locality preserving power of the features. The steps of the algorithm given by He et al. (2006) 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 $D = diag(S.\overrightarrow{1})$. Compute the Laplacian Score L_i of the i^{th} feature as

$$L_i = \frac{\tilde{f}_i^T L \tilde{f}_i}{\tilde{f}_i^T D \tilde{f}_i} \tag{8}$$

$$\text{where } \tilde{f}_i = f_i - \frac{f_i^T D \overrightarrow{1}}{\overrightarrow{1}^T D \overrightarrow{1}} \overrightarrow{1} \tag{9}$$

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 two connected sample points.

4.3 Validity of the underlying assumptions

Here, we validate the two underlying assumptions of the method proposed: (i) the scoring function ranks the irrelevant features at the end, and (ii) the scoring function does not distinguish between irrelevant features. Figures 1 and 2 shows the repartition of the variables at each rank for the Art-1 and Art-2 data-sets, respectively. $Variable_i$ refers to the original variables, $Noise_i$ denotes the noisy dimensions and $Copy_i$ stands for the noisy copy of the $Variable_i$. In both figures (fig. 1 and 2), the irrelevant variables appears significantly at the end of the ranking; this accounts for the first assumption validity. Then, it seems that variables with the same level of relevancy are not properly distinguished; this is more clear in figure 2 where the noisy dimensions appear globally with the same frequencies at ranks from 5 to 10. This accounts for our second assumption.

4.4 Evaluation methodology

The labels of the sample points from the data-sets used in our experiments were available and the scoring function 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

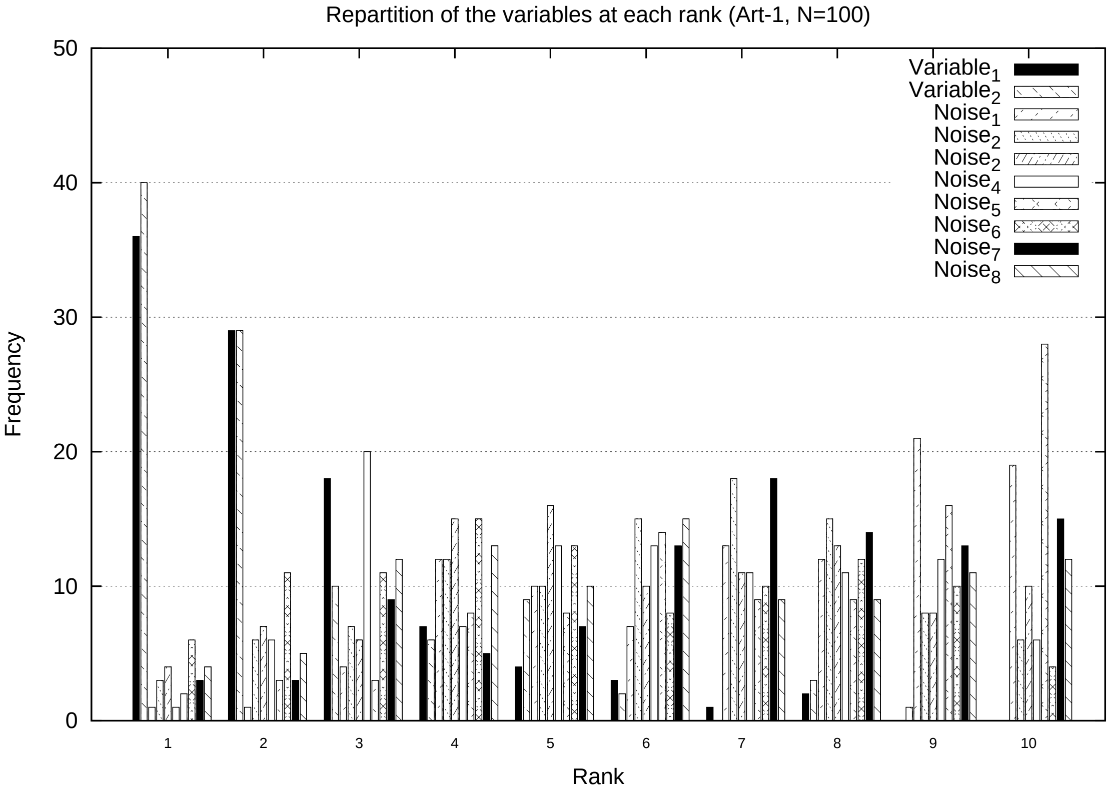

Figure 1: Repartition of the variables at each rank for the Art-1 data-set: each of the 100 scoring function evaluations was performed using sub-samples of $N = 100$ sample points from the Art-1 data-set. $Variable_i$ refers to the original variables and $Noise_i$ denotes the noisy dimensions. It clearly appears that the most represented variables at the first rank are the relevant one. Then, the repartition of the noisy dimensions is more or less uniform from the third or the fourth rank to the end.

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 features over the sub-sample size in the following. To evaluate the robustness of our method to the small sample size issue, we

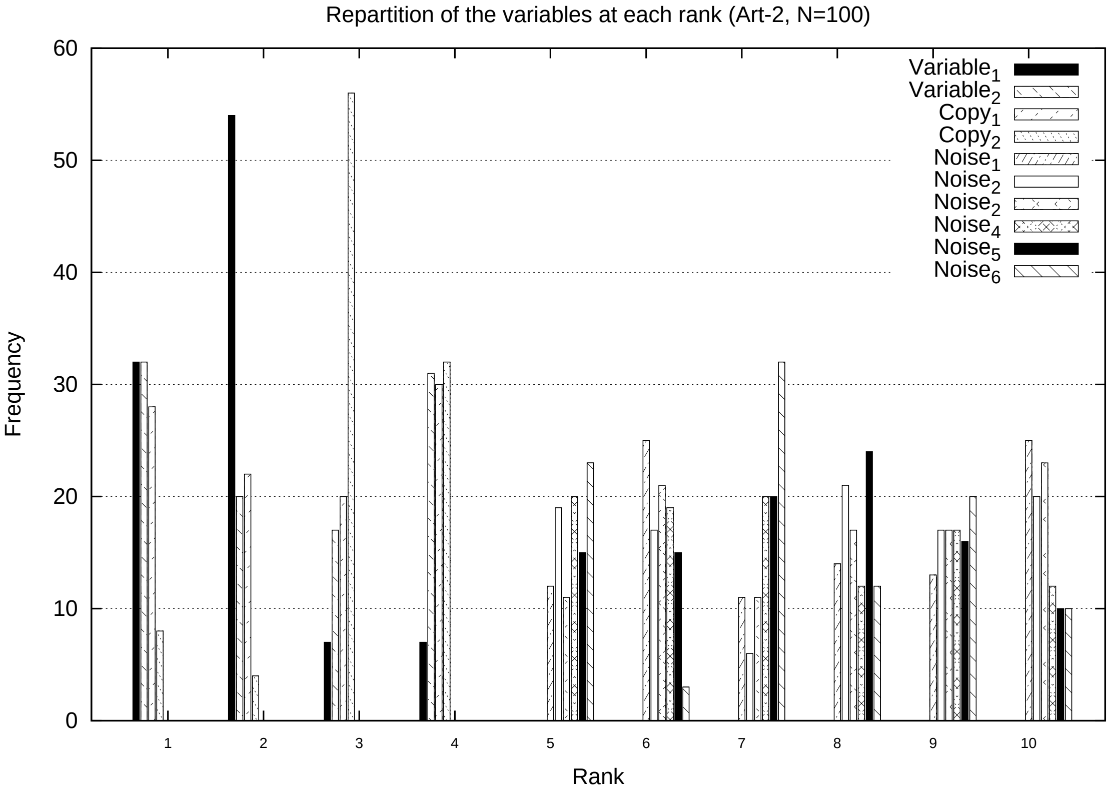

Figure 2: Repartition of the variables at each rank for the Art-2 data-set: each of the 100 scoring function evaluations was performed using sub-samples of $N = 100$ sample points from the Art-2 data-set. $Variable_i$ refers to the original variables, $Noise_i$ denotes the noisy dimensions and $Copy_i$ stands for the noisy copy of the $Variable_i$. The detection of the irrelevant variables is quite perfect here; the original variables and their noisy copies are always ranked between the first and the fourth positions. The irrelevant variables come next without any preference between them.

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)*".

4.5 Results and discussion

Figure 3 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 to 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

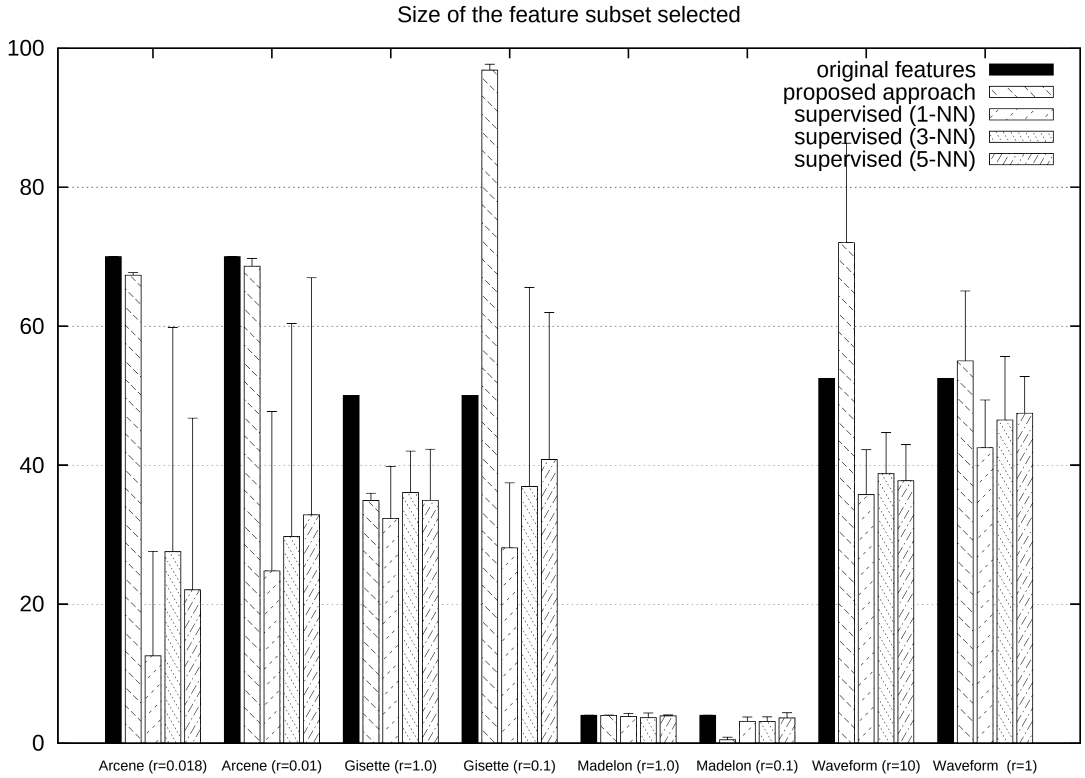

Figure 3: Size of the feature subset selected: index 100 corresponds to the whole set of features (original features + probes), r is the ratio between the number of sample points and the number of features, the values indicated are averaged over the 10 folds and the standard deviation are given between brackets. The supervised method refers to the choice of the classifier with the lowest error rate.

top-k feature subsets. Figure 4 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 disguarded due to space limitation. As shown by the figure 4, a small increasing of the 1-NN classifier error rate is observed with the 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. 3). 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 criteria 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

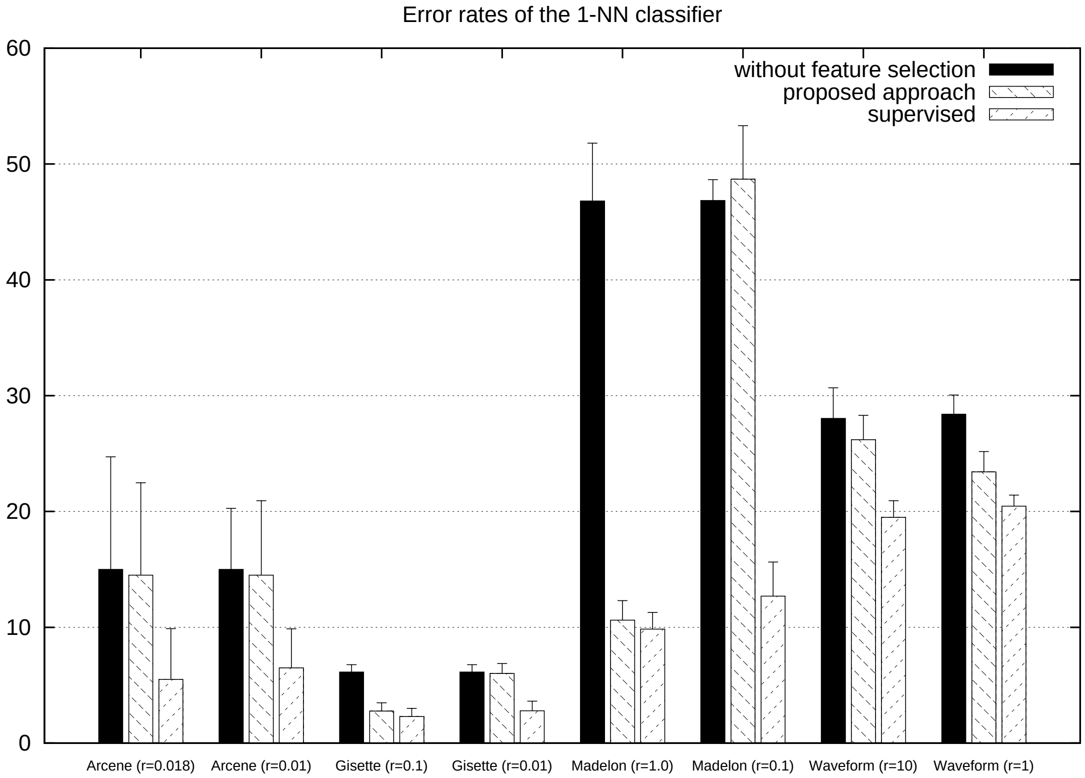

Figure 4: Error rates of the 1-NN classifiers: r refer to the ratio between the number of sample points and the number of features, the values indicated are averaged over the 10 folds. The supervised method refers to the choice of the classifier with the lowest error rate.

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 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.

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

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. The running times observed during our experiments are gathered in table 3. It appears that most of the running time is spent to evaluate the scoring function over each sub-samples since their size is large enough. The computation time of the stopping criterion proposed does not depend the sub-samples size but only on the number of variables (and the number of sub-samples). Anyway, the running times observed seem to us rather reasonable in the context of exploratory analysis. Moreover, the computation needed by our approach can be easily distributed to improve its scalability: (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 to their relevancy is available. The approach proposed relies on resampling and the detection of randomness. It should be

Table 3: Running times: r is the ratio between the number of sample points N and the number of features n. The running times shown respectively refer to the Laplacian Scores computation, the time required to compute our stopping criterion and the sum of both. The values indicated are averaged over the 10 folds. These running times have been measured using Matlab 2006a running on a Dell Precision 670 Workstation (single Intel Xeon 2.8 GHz processor) under the GNU Linux 32 bits operating system.

Data-set		N	n	LS	Criterion	Overall
Arcene	$(r = 0.018)$	180	10000	650.0 s	983.7 s	1633.7 s
	$(r = 0.01)$	100	10000	527.2 s	981.9 s	1508.7 s
Gisette	$(r = 0.1)$	500	5000	1213.9 s	400.5 s	1614.4 s
	$(r = 0.01)$	50	5000	230.5 s	381.3 s	611.8 s
Madelon	$(r = 1)$	500	500	146.0 s	21.7 s	167.7 s
	$(r = 0.1)$	50	500	24.4 s	22.5 s	46.9 s
Waveform	$(r = 10)$	400	40	15.8 s	0.5 s	16.4 s
	$(r = 1)$	40	40	2.7 s	0.8 s	3.5 s

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 they apply 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 improvements. 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 rigourous 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 (Huang et al., 2005) and by the ω-SOM algorithms (Guérif and Bennani, 2007).

References

Jayanta Basak, Rajat K. De, and Sankar K. Pal. Unsupervised feature selection using a neuro-fuzzy approach. *Pattern Recognition Letter*, 19(11):997–1006, 1998.

A. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(1-2):245–271, 1997.

M. Dash and H. Liu. Feature selection for classification. *Intelligent Data Analysis*, 1: 131–156, 1997.

Manoranjan Dash and Huan Liu. 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*, pages 110–121, London, UK, 2000. Springer-Verlag.

Manoranjan Dash, Hua Liu, and J. Yao. Dimensionality reduction of unsupervised data. In *ICTAI*, pages 532–539, 1997.

C.L. Blake D.J. Newman, S. Hettich and C.J. Merz. UCI repository of machine learning databases, 1998. URL `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2001. ISBN 0-471-05669-3.

Jennifer G. Dy and Carla E. Brodley. Feature subset selection and order identification for unsupervised learning. In *Proceedings 17th International Conference on Machine Learning*, pages 247–254. Morgan Kaufmann, San Francisco, CA, 2000.

Jennifer G. Dy and Carla E. Brodley. Feature Selection for Unsupervised Learning. *Journal of Machine Learning Research*, 5:845–889, 2004.

S. Guérif and Y. Bennani. 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*, pages 28–33, Aug 2006.

S. Guérif, Y. Bennani, and E. Janvier. μ-som : Weighting features during clustering. In *Proceedings of the 5th Workshop On Self-Organizing Maps (WSOM'05)*, pages 397–404, Sep 2005.

Sébastien Guérif and Younès Bennani. Dimensionality Reduction Through Unsupervised Features Selection. In *Proceedings of the 10th International Conference on Engineering Applications of Neural Networks (EANN2007)*, pages 98–106. Publishing Centre Alexander T.E.I. of Thessaloniki, August 2007.

Isabelle Guyon and André Eliseeff. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

Isabelle Guyon, Steve R. Gunn, Asa Ben-Hur, and Gideon Dror. Result analysis of the nips 2003 feature selection challenge. In *NIPS*, 2004.

Xiaofei He, Deng Cai, and Partha Niyogi. Laplacian score for feature selection. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 507–514. MIT Press, Cambridge, MA, 2006.

Joshua Zhexue Huang, Michael K. Ng, Hongqiang Rong, and Zichen Li. Automated variable weighting in k-means type clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):657–668, 2005.

Ron Kohavi and George H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324, 1997.

M. H. C. Law, M. A. T. Figueiredo, and A. K. Jain. Simultaneous feature selection and clustering using mixture models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(9):1154–1166, 2004.

Luying Liu, Jianchu Kang, Jing Yu, and Zhongliang Wang. 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*, pages 597– 601, 2005.

P. Mitra, C.A. Murthy, and S.K. Pal. Unsupervised Feature Selection Using Feature Similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4), 2002.

Lance Parsons, Ehtesham Haque, and Huan Liu. Subspace clustering for high dimensional data: a review. *SIGKDD Explor. Newsl.*, 6(1):90–105, June 2004.

F. Questier, R. Put, D. Coomans, B. Walczak, and Y. Vander Heyden. The use of CART and multivariate regression trees for supervised and unsupervised feature selection. *Chemometrics and Intelligent Laboratory Systems*, 76(1):45–54, 2005.

A.E. Raftery and N. Dean. Variable selection for model-based clustering. *Journal of the American Statistical Association*, 101:168–178, 2006.

Sam T. Roweis and Lawrence K. Saul. Nonlinear Dimensionality Reduction by Local Linear Embedding. *Science*, 290:2323–2326, Dec 2000.

Joshua B. Tanenbaum, Vin de Silva, and John C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290:2319–2323, Dec 2000.

Juha Vesanto and Jussi Ahola. Hunting for Correlations in Data Using the Self-Organizing Map. In H. Bothe, E. Oja, E. Massad, and C. Haefke, editors, *Proceeding of the International ICSC Congress on Computational Intelligence Methods and Applications (CIMA '99)*, pages 279–285. ICSC Academic Press, 1999.

Nirmalie Wiratunga, Robert Lothian, and Stewart Massie. Unsupervised feature selection for text data. In Thomas Roth-Berghofer, Mehmet H. Göker, and H. Altay Güvenir, editors, *ECCBR*, volume 4106 of *Lecture Notes in Computer Science*, pages 340–354. Springer, 2006.