Available online at www.sciencedirect.com
SCIENCE DIRECT®
Pattern Recognition Letters 27 (2006) 627–635
Pattern Recognition Letters
www.elsevier.com/locate/patrec

Feature selection in robust clustering based on Laplace mixture

Aurélien Cord *, Christophe Ambroise, Jean-Pierre Cocquerez

Department of Pattern Recognition (ASTRID), Université de Technologie de Compiègne, CNRS HEUDIASYC/UMR 6599, Centre de Recherche de Royallieu BP 20529, 60205 COMPIEGNE Cedex, France

Received 17 September 2004; received in revised form 12 September 2005
Available online 29 November 2005

Communicated by M.A.T. Figueiredo

Abstract

A wrapped feature selection process is proposed in the context of robust clustering based on Laplace mixture models. The clustering approach we consider is a generalization of the *K*-median algorithm. The selection process makes use of the statistical model and recursively deletes features using hypothesis tests. We report simulations and applications to real data sets which illustrate the relevance of the proposed approach. We propose a strategy to select a reasonable number of remaining features. It uses the test statistic to choose the most relevant features, then an evaluation of the clustering error to discard the redundant ones from among them. This strategy appears to produce a good compromise between the selection of features and the performance of the clustering.

Keywords: Clustering; Feature selection; Laplace distribution; Kruskal–Wallis statistical test; EM algorithm

1. Introduction

Clustering algorithms are developed and used in many different fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics. Clustering involves finding subsets (clusters) of ''similar'' observations—similarity often being defined by a distance measure. In real life we generally have to deal with untypical values, called ''outliers'', occurring in data sets. These may impact considerably the statistical tools that can be used for clustering. Centroid-based methods, for example, group observations around a representative sample, which can often be defined as a weighted mean of the cluster. One solution involves using an alternative centroid based on a robust statistic such as the median.

More and more often, data sets in numerous research areas contain a very large number (from hundreds to tens of thousands) of features, providing a powerful contextual description of the observed phenomena (Turney, 1993). Some features, however, are not relevant: their presence can obscure important structures and generally confuse the description process. To tackle this problem there exist methods for selecting only those features which are relevant. These methods combine a variety of aims, such as decreasing computation time and storage requirements, increasing data interpretability, facilitating data visualization, and eliminating noisy features that can adversely affect the performance of most learning algorithms.

This paper proposes an original method for feature selection when using clustering based on Laplace mixture models. This method is a generalization of the *K*-median algorithm. The selection process recursively deletes features using statistical hypothesis tests.

Both clustering and feature selection rely on statistical models. This approach offers several advantages: (1) It allows a well-justified criterion to be proposed. (2) The assumptions in the model, even though they might be simple, are well known and can allow the user to determine if they are adapted to a particular problem. (3) It can be

* Corresponding author. Fax: +33 3 44 23 44 77.
E-mail addresses: aurelien.cord@hds.utc.fr, aurelien.cord@rssd.esa.int (A. Cord).

0167-8655/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2005.09.028

adjusted to the problem: it is possible to leave only a small number of free parameters and thus to propose parsimonious models. For instance, it is possible to consider equal mixing proportion, if one suspects that subpopulations have comparable sizes.

We first describe the Laplace distribution mixture. We then outline the feature-selection method relying on a Kruskal–Wallis statistical test. Finally, we illustrate the performance of our algorithm both on synthetic and real data sets.

2. Mixture of Laplace distributions and median clustering

Dealing with outliers is a recurrent problem in data analysis. Outliers can occur for a variety of reasons, including noise and errors (during the data acquisition or transmission, for instance), or they can simply be correct but untypical values existing in the data set. Outliers have a large influence on normal statistics and can dramatically affect perceived distributions. Although the sample mean is by far the most commonly quoted measure of location, it is strongly affected by outliers and, moreover, it often does not depict the typical outcome.

An alternative measure of the distribution center can be based on order statistics. In particular, for the majority of skewed data sets, the sample median is likely to be a more realistic measure of center than the mean.

The use of Gaussian distributions for clustering relies on the quadratic distance between outcomes and their mean, and it is therefore clear that this kind of process will be sensitive to the presence of outliers (Banfield and Raftery, 1993).

A typical solution is to use an algorithm such as k-median clustering (Bradley et al., 1997), which is based on L1-norm distance. This generates a partition of the data set into K groups such that the sum of the distances from the observations to their cluster's centroid is minimized. It is an alternative optimization algorithm, similar to the very classical k-mean clustering algorithm (Selim and Ismail, 1984), except that k-median uses the L1-norm distance as its distance measure, whereas k-mean uses L2-norm distance. The criterion optimized by k-median is the sum of the absolute distances between the observations and their class's centroid:

$$J(\boldsymbol{x},\boldsymbol{\mu}) = \sum_{k=1}^{K} \sum_{i/x_i \in \mathrm{Class}_k} |\boldsymbol{x}_i - \boldsymbol{\mu}_k|. \tag{1}$$

2.1. Mixture analysis

Basing cluster analysis on a mixture model has become a classical and powerful approach (MacLachlan and Peel, 2000; Figueiredo and Jain, 2002). Each observation is assumed to have been drawn from a mixture of parametric distributions. Data $\boldsymbol{x} = (\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N)$ are assumed to arise from a random vector with density

$$f(\boldsymbol{x};\boldsymbol{\Phi}) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x};\boldsymbol{\theta}_k), \tag{2}$$

where $\boldsymbol{\Phi} = (\pi_1,\ldots,\pi_K,\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_K)$, where $\pi_1,\ldots,\pi_K$ denote the proportions of the mixture and $\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_K$ the parameters of each component density. Obviously, being probabilities, the π_k must satisfy $\forall k \in [1,\ldots,K]$, $\pi_k \geqslant 0$ and $\sum_{k=1}^{K} \pi_k = 1$.

In this context two commonly used maximum likelihood (ML) approaches have been proposed: the mixture approach and the classification approach. Loosely speaking, the mixture approach aims to maximize the likelihood over the mixture parameters, whereas the classification approach aims to maximize the likelihood over the mixture parameters and over the identifying labels of the mixture component origin for each point.

In the mixture approach $\boldsymbol{\Phi}$ is chosen to maximize the *log-likelihood*

$$C_{\mathrm{mel}}(\boldsymbol{x},\boldsymbol{\Phi}) = \sum_{i=1}^{N} \log\left(\sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x};\boldsymbol{\theta}_k)\right).$$

A partition of the data can be directly derived from the ML estimates of the mixture parameters by assigning each $\boldsymbol{x}_i$ to the component which provides the greatest conditional probability.

In the classification approach the likelihood is maximized over the mixture parameters and over the identifying labels of the mixture component origin for each point:

$$C_{\mathrm{cla}}(\boldsymbol{x},\boldsymbol{\Phi}) = \sum_{k=1}^{K} \sum_{i/x_i \in \mathrm{Class}_k} \log \pi_k f_k(\boldsymbol{x};\boldsymbol{\theta}_k).$$

A partition of the data is given directly by the identifying labels.

2.2. A robust clustering algorithm

As a generalization of the k-median algorithm, we propose using Laplace distributions, that is to say a mixture-based clustering approach with distributions relying on the median (Ernst, 1998). This algorithm was proposed in (Dang, 1998).

We assume that all the components of the mixture are D-variate Laplace distributions. Moreover, we suppose that the features are independent within each class. Thus each D-dimensional Laplace law $\mathscr{L}(\mu_k,\lambda_k)$ is a product of D mono-dimensional Laplace distributions $\mathscr{L}(\mu_{kd},\lambda_{kd})$.

The mono-dimensional Laplace distribution $\mathscr{L}(\mu,\lambda)$ is defined as

$$f_{\mathscr{L}}(x|\mu,\lambda) = \frac{1}{2\lambda}\exp\left(-\frac{|x-\mu|}{\lambda}\right), \tag{3}$$

where $x \in \mathbb{R}$, $\lambda > 0$ and μ is the median.

The density function of the D-dimensional Laplace law $\mathscr{L}(\boldsymbol{\mu}_k,\boldsymbol{\lambda}_k)$ is therefore

$$f_{\mathscr{L}}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\lambda}_k) = \prod_{d=1}^{D} \frac{1}{2\lambda_{kd}} \exp\left(-\frac{|x_d - \mu_{kd}|}{\lambda_{kd}}\right). \quad (4)$$

Relying on a mixture approach, we propose using the Expectation Maximization (EM) algorithm (Dempster et al., 1977) for the maximum likelihood estimate of the mixture parameters.

After convergence, clustering is performed by assigning each observation to the class having the highest posterior probability (Maximum A Posteriori strategy, MAP).

It should be noted that in using the mixture of Laplace distributions with the assumption of equal mixing proportion and equal covariance structure, the classification approach is equivalent to finding the partition maximizing the k-median criterion:

$$\begin{aligned} C_{\text{cla}}(\boldsymbol{x}, \boldsymbol{\Phi}) &= \sum_{k=1}^{K} \sum_{i/x_i \in \text{Class}_k} \log \pi_k f_k(\boldsymbol{x}; \boldsymbol{\theta}_k) \\ &= \sum_{k=1}^{K} \sum_{i/x_i \in \text{Class}_k} \log \pi_k \left(\prod_{d=1}^{D} \frac{1}{2\lambda_{kd}} \exp\left(-\frac{|x_{id} - \mu_{kd}|}{\lambda_{kd}}\right)\right) \\ &= \sum_{k=1}^{K} \sum_{i/x_i \in \text{Class}_k} \sum_{d}^{D} |\boldsymbol{x}_{id} - \boldsymbol{\mu}_{kd}| + C \end{aligned}$$

where C denotes a constant.

In this context the CEM algorithm (Celeux and Govaert, 1992) is equivalent to the k-median algorithm, and the approach we propose can therefore be considered as an EM-type generalization of the k-median algorithm.

The mixture of Laplace distributions has been used in different contexts including texture analysis (Amin and Guan, 2004), source separation (Mitianoudis and Stathaki, 2005) and speech recognition (Ortmanns et al., 1997).

3. Feature selection

Over the past few years a lot of research areas within the field of machine learning have had to explore domains containing hundreds to tens of thousands of variables or features (e.g., clustering in molecular biology, web search engines, image retrieval). It might be supposed that greater volumes of data imply better descriptions, but in practice some features are not significant and can get in the way of the description process. Deciding which features are relevant becomes fundamental.

The case of supervised learning scenarios has been largely studied in the literature (Guyon and Elisseeff, 2003). The problem is well defined and consists of removing features to improve the classification rate on new data. Different methods are described in the literature (Kohavi and John, 1997): wrapped methods treat the learning machine as a black box in order to rank subsets of features. Filter methods employ a pre-processing step in which subsets of features are selected, independently of the chosen predictor. Embedded methods perform feature selection as part of the training process.

Feature selection for unsupervised learning is a difficult task because there are no class labels for the data and so no obvious criteria to guide the search (Dy and Brodley, 2000; Roth and Lange, 2003; Law et al., 2004).

3.1. Statistical tests for feature ranking

The feature selection method we propose is based on a wrapped method. Its crucial component is the determination and sorting of all potentially relevant features, leading to the exclusion of features that have only a small influence on the data.

It relies on statistical models and tests. The idea is simple: for each feature ($d = 1, \ldots, D$), we test

H0: All the medians of the different classes are similar ($\mu_{1d} = \mu_{2d} = \cdots = \mu_{Kd}$). This means that the feature in question d does not separate the classes and is therefore not relevant.

H1: At least one of the medians is different from the others.

This produces a test statistic whose value is used to rank the features in terms of their relevance.

The Kruskal–Wallis statistical test is a non-parametric test that makes no assumptions about the distribution of the data (e.g., normality) (Hollander and Wolfe, 1973; Gibbons and Chakraborty, 1992). This test is an alternative to the independent group ANOVA, when the assumption of normality or equality of variance may not apply. Like many non-parametric tests it uses data rank rather than raw values to calculate the statistic.

Let $n_1, n_2, \ldots, n_K$ represent the sample sizes for each of the K classes. The total sample size is $N = \sum_{k=1}^{K} n_k$. We rank the combined sample and compute the sum of the ranks for the class k: $R_k = \sum_{i|x_i \in \text{Class}_k} \text{rank}(x_i)$. The Kruskal–Wallis test statistic is then $H = \frac{12}{N(N+1)} \sum_{k=1}^{K} \frac{R_k^2}{n_k} - 3(N+1)$.

If the null hypothesis of equal median holds, this test statistic corresponds approximately to a chi-square distribution with $K - 1$ degrees of freedom. The larger the test statistic H, the weaker the null hypothesis becomes, since a strong separation of the medians indicates that the feature under consideration has a high clustering power. In the following test statistics are used to classify features according to their level of relevance.

3.2. Recursive feature selection for Laplace mixture clustering

For the feature selection we used a kind of wrapped technique called the Recursive Feature Elimination Algorithm. Often known as sequential backward search, it is one of the most popular heuristics for feature selection (Pudil et al., 1994). It was adapted by Guyon et al. (2002) for selecting relevant features in two-class problems using linear SVM. It is a recursive process: starting by considering all available

features, each step consists of ranking the features according the level of relevance and discarding some of the less discriminant features. The clustering algorithm is then re-applied in the reduced feature space.

In summary, we rely on statistical models to propose both a generalization of k-median and a wrapper feature-selection procedure. The algorithm we propose is the following:

(1) Clustering:
  (a) Use the EM algorithm to estimate the Laplace mixture parameters.
  (b) Assign each observation to the component with the highest posterior probability (Maximum A Posteriori strategy).
(2) For each feature $d = 1, \ldots, D$ compute the Kruskal–Wallis statistics.
(3) Use the test statistics to sort features and delete a fixed number or fixed percentage of the less relevant ones.
(4) Stop if there are no remaining features, else reiterate from the first step in the reduced space.

An analysis of computational requirements is carried out for the first real data set (Section 4.2).

4. Results

To illustrate the performance of our algorithm we test it on two synthetic data sets and three publicly-available data sets from the UCI Machine Learning Repository (Murphy and Aha, 1992), as well as on one from Alon et al. (1999). The entire data set is used both for the clustering and evaluation of error rates. For each test we run our algorithm 20 times with different initializations. The features are discarded one by one: only the least relevant feature is eliminated during each iteration. To avoid local convergence, at each iteration of our algorithm the EM-algorithm is repeated at least five times with a random initialization, and the solution corresponding to the best log-likelihood is chosen.

For the sake of clarity three curves are shown: the test statistic, the classification error and the clustering error in relation to the number of remaining features. The test statistic curve traces the smallest values of the Kruskal–Wallis test as a function of the number of remaining features. It corresponds to the test value of the feature that is eliminated during the current iteration. The classification error is the percentage of incorrectly classified examples, and is computed by finding the permutation of the actual labels associated with each cluster that minimizes the difference between estimated and true classes. The true class labels are used only in generating the classification error and not in obtaining the clusters. The clustering error is computed in exactly the same way, except that instead of the true class label we use the class label of the clustering computed when all the features are retained. It will be noticed that the clustering error on the data set containing all the features is zero by definition. In all the curves error bars show one sample standard deviation above and below each point.

In the case of unlabelled data clustering, only two of the three curves are available: the test statistic and the clustering error. In the following we will show how the two curves are complementary and how they can be combined in order to choose a reasonable number of features to retain.

4.1. Synthetic data sets

We first consider a "toy" synthetic data set, easy to deal with: it consists of 800 observations from a mixture of four Gaussian distributions $\mathcal{N}(\boldsymbol{m}_i, \boldsymbol{I}), i = \{1, 2, 3, 4\}$, where $\boldsymbol{m}_1 = (0, 3)$, $\boldsymbol{m}_2 = (1, 9)$, $\boldsymbol{m}_3 = (6, 4)$ and $\boldsymbol{m}_4 = (7, 10)$ mixed in the same proportions. Eight non-relevant features, sampled from a $\mathcal{N}(0, 1)$, are added to this data. It gives a 10-dimensional pattern.

The classification error plotted in Fig. 1a shows that in all 20 executions the four components are always correctly identified as long as there are at least two features. The class label of the clustering computed when all the features are retained is very similar to the actual class label (three

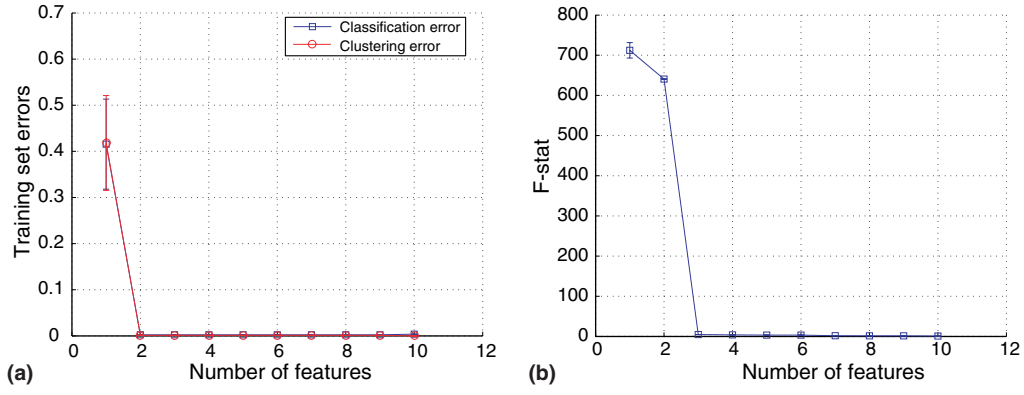

Fig. 1. Results of the 10-class synthetic data set. (a) Error curves. (b) Test-statistic curve. Note that the error bars are generally not visible because of the low variance between runs on this data set.

classification errors for 800 points) and the classification and clustering errors are identical.

The values of the statistical test (Fig. 1b) are drastically different between the two relevant features and the others. This example gives us the extreme values that the test statistic can reach for a 10-dimensional pattern: less than 5 for totally irrelevant, and greater than 500 for the totally relevant features.

Here separation of the relevant from the non-relevant features is very easy and can be done with the clustering error or the test statistic curves indifferently.

The second synthetic data set was first proposed by Trunk (1979). It consists of 1000 points from a mixture of two 20-dimensional Gaussian $\mathcal{N}(\boldsymbol{m}_1, \boldsymbol{I})$ and $\mathcal{N}(\boldsymbol{m}_2, \boldsymbol{I})$ where $\boldsymbol{m}_1 = \left(1, \frac{1}{\sqrt{2}}, \ldots, \frac{1}{\sqrt{20}}\right)$, $\boldsymbol{m}_2 = -\boldsymbol{m}_1$. Note that the features are sorted in order of decreasing relevance. The lower the rank, the more discriminant the feature. To produce the results curves (Fig. 2) with meaningful error bars, the algorithm runs 20 times simulating a new set of synthetic data each time. Both the errors and the values of the test statistic decrease as the number of features increases, in accordance with the true characteristics of the data. In particular, it confirms the link between test statistic values and the level of relevance of the features.

We should note that it is not possible to use classical p-values to select the features, because almost all the features have a p-value very close to 0. It therefore becomes almost impossible to distinguish the different levels of relevance of the data. This is a consequence of the fact that the p-value in clustering will be exaggerated. In particular, a data set without any cluster structure can have a very small p-value. A simple example is to consider k-means clustering (with $k = 2$) on a data set consisting of points uniformly distributed in the interval $[0, 1]$. Cluster A will be centered at 0.25, whereas cluster B will be centered at 0.75. If one does a hypothesis testing on

H0: Cluster A and cluster B are the same (has the same mean).
H1: Cluster A and cluster B are different (has different means).

One will get a very small p-value and reject H0. The cause for this is that the clustering process "inflates" the separation of the classes. Notice that it is only a warning and that our technique is not affected by this remark because no hypothesis testing is involved for the feature sorting.

4.2. Real data sets

In order to compare our approach with exiting methods we chose publicly available data sets that were used by Mangasarian and Wild (2004) and Law et al. (2004). This comparison is described in Section 4.3.

We tested our method on the wine-recognition data set. These data are the results of a chemical analysis of 13 constituents found in 178 wines produced in the same region in Italy but by three different growers.

Fig. 3 shows the results of our analysis. The total time required to generate the resulting curves (which entailed running the EM-algorithm 1300 times, selecting the features 260 times and plotting the graphs, all within MATLAB 7) was 78 s on a 2 GHz 512 MB RAM desktop machine running Windows XP.

In Fig. 3b a gap appears between the last six relevant features (from 8 to 13) and the first seven (from 1 to 7). Classification and clustering errors are lowest when exactly seven features are retained. These features all have a more or less equivalent capacity to separate the different clusters, when considering the value of the Kruskal–Wallis test statistic, but there are probably some redundant features that can be eliminated at only a small cost in terms of errors.

In Fig. 3a both the classification and the clustering error curves rise gently as the number of features decreases from 13 to 4, but then rise steeply as the number of features decreases from 4 to 1. In this example these two curves have the same behavior. For the treatment of unlabelled data the classification error is not available. However, the similarity of the error curves would seem to indicate that the clustering error alone is sufficient to determine a compromise between a tolerable error magnitude and the number of features to be retained.

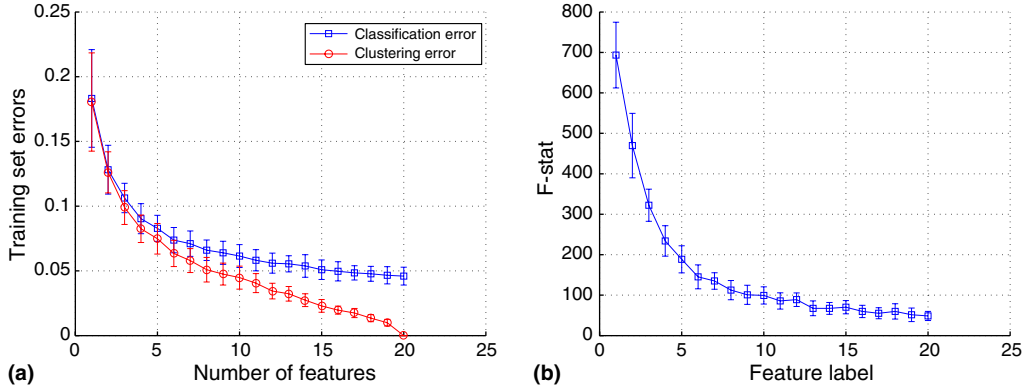

Fig. 2. Results of the 20-class synthetic data set from Trunk (1979). (a) Error curves. (b) Test-statistic in decreasing order of relevance of features.

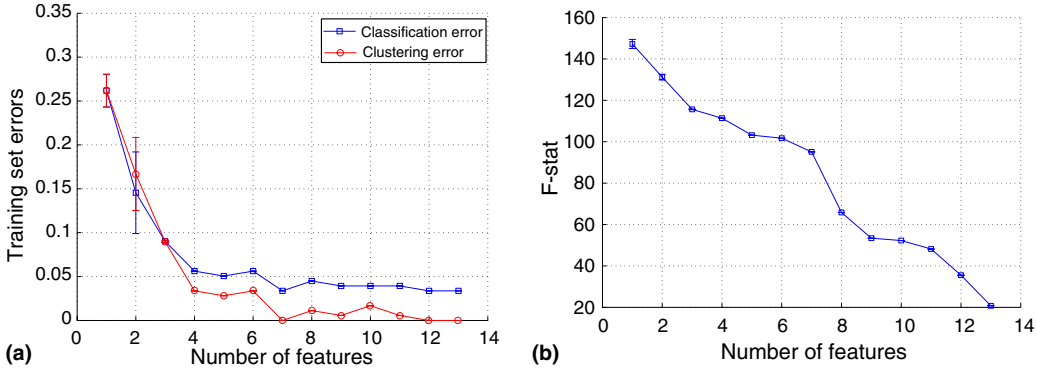

Fig. 3. Results of the wine data set from Murphy and Aha (1992). (a) Error curves. (b) Test-statistic curve.

It is interesting to note that the test statistic curve does not have the same behavior as the error curves. The two curves need to be combined to provide a better insight into the clustering. There are two possible strategies for this data set: we can either retain seven features, thus minimizing the errors; alternatively we can retain only four features, which represents a good compromise between the errors and the number of features.

The Cleveland Heart data set consists of a database of heart disease diagnoses collected from the Long Beach and Cleveland Clinic Foundation (Robert Detrano, M.D., Ph.D.). This database contains 76 attributes, but all published experiments have used a subset of just 14 of them, and we shall do likewise. The "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. Experiments with the Cleveland database have so far attempted only to distinguish presence (values 1, 2, 3, 4) from absence (value 0).

The classification error curve is surprising insofar as it falls almost continuously as features are eliminated (Fig. 4a). The clustering error curve has very wide error bars, and in this example the parallel between the two curves is not clear. We notice that the partition obtained with all the features is dramatically different from the partition obtained with a reduced number of features, since the clustering error jumps from 0% to 20%. This could be explained by the fact that some features with low clustering power have influenced the resulting partition during the first steps of the procedure.

However, the test statistic in Fig. 4b is quite simple: it increases slightly as the number of features decreases from 13 to 2, staying below 50, but then leaps to almost 300 when the number of features is reduced from 2 to 1. A decision about which feature are relevant is possible considering the test statistic curve related to the Cleveland data set.

We go on to examine further how these curves can be combined.

Fig. 5 gives the results from the colon data analyzed initially in (Alon et al., 1999). It consists of 62 tissue samples described by 2000 human gene expressions (40 tumors and 22 normal tissues). The actual implementation of the algorithm has difficulties dealing with so many features (precision problems) and we thus consider a subset of the available features extracted by a simple filtering procedure: the 210 features with the highest t-statistics (computed from the two real groups) were extracted from the 2000 initial features. Notice that the purpose of processing this data set is to test the method on a real data set having a large amount of features (more than 200), but in context of unsupervised clustering it may not be the best possible experimental setting.

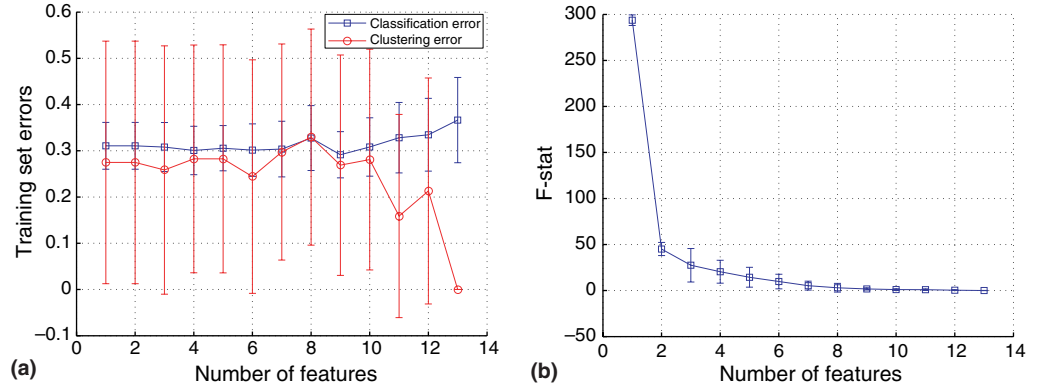

Fig. 4. Results of the Cleveland Heart data set from Murphy and Aha (1992). (a) Error curves. (b) Test-statistic curve.

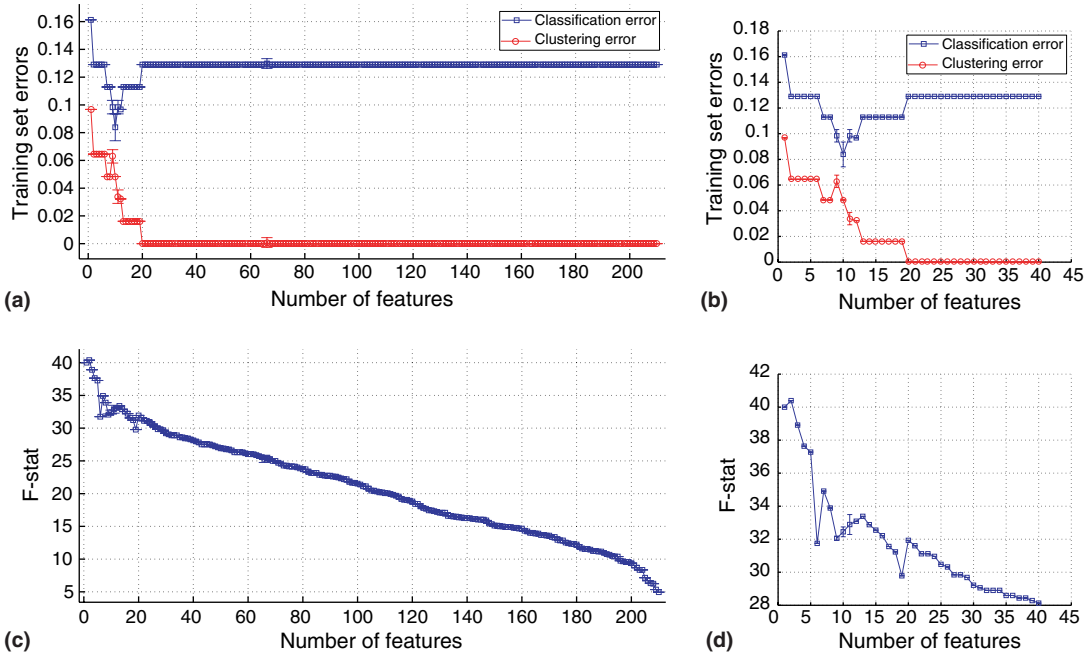

Fig. 5. Results of the colon data set from Alon et al. (1999). (a) Error curve. (b) Zoom on the error curve. (c) Test statistic curve. (d) Zoom on the test statistic curve.

From Fig. 5a it is clear that there is no variation in clustering results when we eliminate the features from 210 to 20. A first approximation might be to stop the selection here. However, to pursue the analysis further we shall examine Fig. 5b and d, where we have zoomed in on the last 40 features. In the test-statistic curve a small gap appears between the sixth and the fifth features (from 32 to 37), showing that the first five features are really more relevant than the others. The clustering error curve then remains horizontal from the fifth to the second feature, showing that these features are probably redundant, and suggesting that only the first 2 among the 210 initial features be retained. The classification error, in this case, is the same with 2 as with 210 features, confirming the validity of our method. Notice that a feature subset of 10 leads to the smallest classification error. However, this result is not detectable in the unsupervised case, with our curves, which shows that our method is not perfect in all cases, although it leads to a good compromise for all the examples we have considered.

We propose the following strategy: use the test-statistic curve to choose the most relevant features, and then from among these eliminate the redundant features indicated by the minimum of the clustering error curve. This strategy appears to converge to a good compromise between the number of features and the performance of the clustering, in particular if the parallelism between classification and clustering error curves is not well established.

Our final example is based on the Wisconsin Diagnostic Breast Cancer (WDBC) data set. This contains 576 data points having 30 features. They are computed from characteristics of cell nuclei present in digitized images produced via a fine needle. The goal is to predict the diagnosis (benign, malignant).

From Fig. 6 it would not appear easy at first sight to choose an optimal number of features from the clustering error and test-statistic curves. However, using our method, we shall propose different strategies as in the case of the wine recognition data set: from the test-statistic curve (Fig. 6b) the most relevant features can be separated into a handful of groups of decreasing relevance: from 1 to 6, from 7 to 12, from 13 to 18, from 19 to 26. After the 27th feature the test-statistic values decline significantly, indicating that any further features should not be taken into consideration. For each group we can determine the local minimum for the clustering error: it can be seen from Fig. 6a that the local minima are to be found at features 3, 10, 16 and 22. The choice between the different numbers of features should be made in accordance with the degree of precision a particular problem requires. We should notice that those values correspond to local minima of the classification errors.

4.3. Comparison with other results

In the study by Law et al. (2004) the authors present the results of their clustering algorithm for the wine and WDBC data sets. Their classification errors, 6.61 and 9.35, respectively, are larger than ours, which are from 3.8 to 5.8 and from 6.2 to 6.4, respectively. Their algorithm, however, does not have precisely the same purpose as ours: it not only yields a feature selection, but also allows an evaluation of the best number of classes needed to model the data.

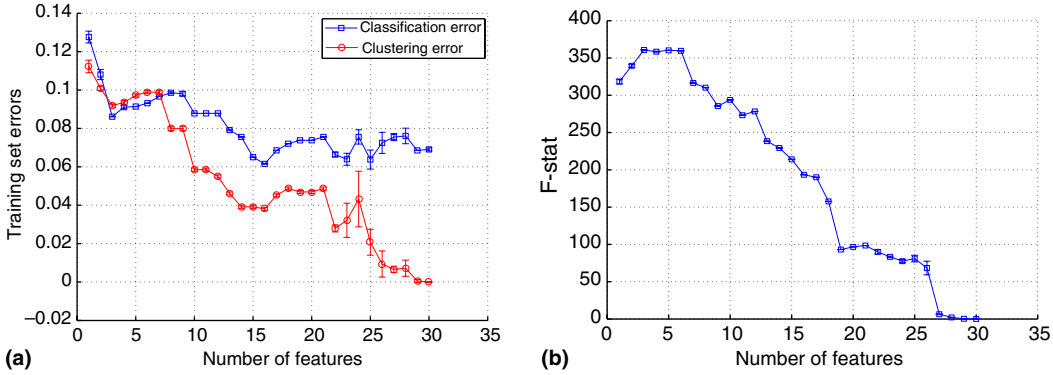

Fig. 6. Results of the WDBC data set from Murphy and Aha (1992). (a) Error curves. (b) Test-statistic curve.

Table 1
Comparison of results from different approaches

Data set	Error (Law et al., 2004)	Error (Mangasarian and Wild, 2004)	Nb features (Mangasarian and Wild, 2004)	Error (this work)	Nb features (this work)
Wine	6.61	4	4	5.8/3.8	4/7
WDBC	9.35	9	7	8.8/8.8/6.2/6.4	3/10/16/22
Cleveland Heart	N/A	28	8	30.5	2

The respective errors are expressed as percentages.

The algorithm of Mangasarian and Wild (2004) is closer to our method. Indeed, their algorithm can be interpreted as a CEM algorithm assuming a mixture of Laplace distributions with equal covariance structure. Our approach offers more flexibility for modeling when the cluster covariances are assumed to be different. It will be noticed from Table 1 that the classification error rates we obtain are similar or better.

Moreover, Mangsarian and Wild use only the clustering error curve to select the number of relevant features. This technique does not work with all the data sets (cf. Cleveland Heart or WDBC data sets). Indeed, the test statistics of Kruskal–Wallis produce precious complementary information for feature selection, as demonstrated in the results analysis section.

5. Conclusion

In this study a cluster analysis based on a mixture model of Laplace distributions is proposed. Generally speaking, the number of iterations needed by the EM algorithm to converge is smaller when using Laplace rather than Gaussian distributions. This compensates for the fact that one EM iteration for a Laplace distribution has a greater time-overhead than its equivalent Gaussian iteration. Moreover, it is clearly less sensitive to the presence of outliers because the measure of the distribution center is based on an order statistic (median). The feature selection is based on a wrapper method, it is then compu- tationally intensive and that can be a limitation to this approach.

In order to propose a practical solution for reducing the number of features we consider that a feature is not essential if the partition does not change greatly when the feature is ignored. This observation is not sufficient in itself to select a feature subset, since the reference partition obtained using all the features can sometimes be substantially influenced by features with low a clustering power. We therefore propose the use of an additional criterion to evaluate the clustering power of a given feature. The whole approach relies on statistical modeling and uses statistical hypothesis testing to measure clustering power.

Applying our method to real data sets illustrates the strategy we propose for selecting a reasonable number of remaining features. This strategy involves combining two steps. First, we extract a group of features that have a high clustering power, corresponding to the largest test statistics. Then, from among these, we discard the redundant features, i.e. those which do not modify the partition when they are ignored. This strategy is a practical tool for exploratory data analysis which simplifies the selection process, too often subjective and ad hoc, performed in a clustering context when using Laplacian mixture models.

Appendix A. EM algorithm for Laplacian mixture

The detailed EM algorithm for Laplacian mixtures necessarily includes a definition of weighted median. Given a set of N scalars $\{x_1,\ldots,x_N\}$ and N weights $\{w_1,\ldots,w_N\}$, positive or null scalar values with at least one weight different from 0, the weighted median $wmedian(\boldsymbol{x},\boldsymbol{w})$ is defined as a scalar that minimizes

$$J(a)=\sum_{i=1}^{N} w_i|x_i-a|, \tag{A.1}$$

where M is the rank and satisfies $\sum_{i=1}^{M-1} w_i < \frac{1}{2}\sum_{i=1}^{N} w_i \leqslant \sum_{i=1}^{M} w_i$.

Two situations arise:

$$wmedian(x,w) = x_M \quad \text{if} \quad \sum_{i=1}^{M} w_i > \frac{1}{2}\sum_{i=1}^{N} w_i > \sum_{i=1}^{M-1} w_i,$$

$$wmedian(x,w) \in]x_M, x_{M+1}[\quad \text{if} \quad \sum_{i=1}^{M} w_i = \frac{1}{2}\sum_{i=1}^{N} w_i.$$

In the second case, we choose $wmedian(x,w) = \frac{x_M + x_{M+1}}{2}$.

The two steps of the EM algorithm at the qth iteration are:

E-step: For $i = 1,\ldots,N$ calculate the probability that x_i comes from the kth component of the mixture:

$$c_{ik}^{(q+1)} = \frac{\pi_k \prod_{d=1}^{D} \frac{1}{2\lambda_{kd}^{(q)}} \exp\left(-\frac{|x_{id} - \mu_{kd}^{(q)}|}{\lambda_{kd}^{(q)}}\right)}{\sum_{k=1}^{K} \pi_k \prod_{d=1}^{D} \frac{1}{2\lambda_{kd}^{(q)}} \exp\left(-\frac{|x_{id} - \mu_{kd}^{(q)}|}{\lambda_{kd}^{(q)}}\right)}. \tag{A.2}$$

M-step: Evaluate the parameters μ^{q+1} and λ^{q+1} that maximize the *log-likelihood*. For $k = 1,\ldots,K$ and $d = 1,\ldots,D$ we have:

$$\mu_{kd}^{(q+1)} = wmedian\left(\left\{(x_{id}, c_{ik}^{(q+1)}), i = 1,\ldots,N\right\}\right), \tag{A.3}$$

$$\lambda_{kd}^{(q+1)} = \frac{1}{n_k^{(q+1)}} \sum_{i=1}^{N} c_{ik}^{(q+1)} |x_{id} - \mu_{kd}^{(q+1)}|, \tag{A.4}$$

$$\text{with } n_k^{(q+1)} = \sum_{i=1}^{N} c_{ik}^{(q+1)}, \tag{A.5}$$

$$\pi_k = \frac{n_k^{(q+1)}}{\sum_{k=1}^{K} n_k^{(q+1)}}. \tag{A.6}$$

References

Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., Levine, A.J., 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. In: Proc. National Academy of Sciences, vol. 96, pp. 6745–6750.

Amin, T., Guan, L., 2004. Interactive content-based image retrieval using Laplacian mixture model in wavelet domain. In: Proc. IEEE Internat. Symp. on Circuits and Systems. Vancouver, Canada.

Banfield, J., Raftery, A., 1993. Model-based Gaussian and non-Gaussian clustering. Biometrics 49, 803–821.

Bradley, P.S., Mangasarian, O.L., Street, W.N., 1997. Clustering via concave minimization. In: Advances in Neural Information Processing Systems, vol. 9. MIT Press, Cambridge, MA, pp. 368–374.

Celeux, G., Govaert, G., 1992. A classification EM algorithm for clustering and two stochastic versions. Comput. Statist. Data Anal. (14), 315–332.

Dang, V.M., 1998. Classification de donnees spatiales: modeles probabilistes et critere de partionnement. Ph.D. thesis, Universite de Technologie de Compiegne.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the em algorithm. J. Roy. Statist. Soc. Ser. B 39, 1–38.

Dy, J.G., Brodley, C.E., 2000. Feature subset selection and order identification for unsupervised learning. In: Proc. Seventeenth Internat. Conf. on Machine Learning, pp. 247–254.

Ernst, M.D., 1998. A multivariate generalized Laplace distribution. Comput. Statist. 13, 227–232.

Figueiredo, M.A.T., Jain, A.K., 2002. Unsupervised learning of finite mixture models. IEEE Trans. Pattern Anal. Machine Intell. 24 (3), 381–396.

Gibbons, J.D., Chakraborty, S., 1992. Nonparametric Statistical Inference, third ed. Marcel Dekker, New York.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182.

Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Mach. Learn. 46, 389–422.

Hollander, M., Wolfe, D.A., 1973. Nonparametric Statistical Methods. John Wiley and Sons, Inc., New York.

Kohavi, R., John, G., 1997. Wrappers for feature selection. Artificial Intell. 97 (1–2), 273–324.

Law, M.H., Figueiredo, M.A.T., Jain, A.K., 2004. Simultaneous feature selection and clustering using mixture models. IEEE Trans. Pattern Anal. Machine Intell. 26 (9), 1154–1166.

MacLachlan, G., Peel, D., 2000. Finite Mixture Models. Wiley.

Mangasarian, O.L., Wild, E.W., 2004. Feature selection in k-median clustering. In: SIAM International Conference on Data Mining, Workshop on Clustering High Dimensional Data and its Applications, La Buena Vista, FL, pp. 23–28.

Mitianoudis, N., Stathaki, T., 2005. Overcomplete source separation using Laplacian mixture models. IEEE Signal Process. Lett. 12 (4), 277–280.

Murphy, P.M., Aha, D.W., 1992. Uci machine learning repository. URL: www.ics.uci.edu/mlearn/MLRepository.html.

Ortmanns, S., Firzlaff, T., Ney, H., 1997. Fast likelihood computation methods for continuous mixture densities in large vocabulary speech recognition. In: Proc. EUROSPEECH-97: Eur. Conf. Speech Technol., pp. 139–142.

Pudil, P., Novovicovà, J., Kittler, J., 1994. Floating search methods in feature selection. Pattern Recog. Lett. 15, 1119–1125.

Roth, V., Lange, T., 2003. Feature selection in clustering problems. In: Advances in Neural Information Processing Systems, vol. 16. MIT Press, Cambridge, MA.

Selim, S., Ismail, M., 1984. *K*-means-type algorithms: A generalized convergence theorem and characterization of local optimality. IEEE Trans. Pattern Anal. Machine Intell. 6 (1), 81–87.

Trunk, G.V., 1979. A problem of dimensionality: A simple example. IEEE Trans. Pattern Anal. Machine Intell. 1 (3), 306–307.

Turney, P., 1993. Robust classification with context-sensitive features. In: Proceedings of the Sixth International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (IEA/AIE-93), Edinburgh, Scotland, pp. 268–276.