A NEW LEARNING ALGORITHM FOR CLASSIFICATION IN THE REDUCED SPACE

Luminita State
Department of Computer Science, University of Pitesti, Pitesti, Romania

Catalina Cocianu, Ion Rosca
Department of Computer Science, Academy of Economic Studies, Bucharest, Romania

Panayiotis Vlamos
Department of Computer Science, Ionian University, Corfu, Greece

Keywords: Feature extraction, informational skeleton, principal component analysis, unsupervised learning, cluster analysis.

Abstract: The aim of the research reported in the paper was twofold: to propose a new approach in cluster analysis and to investigate its performance, when it is combined with dimensionality reduction schemes. Our attempt is based on group skeletons defined by a set of orthogonal and unitary eigen vectors (principal directions) of the sample covariance matrix. Our developments impose a set of quite natural working assumptions on the true but unknown nature of the class system. The search process for the optimal clusters approximating the unknown classes towards getting homogenous groups, where the homogeneity is defined in terms of the "typicality" of components with respect to the current skeleton. Our method is described in the third section of the paper. The compression scheme was set in terms of the principal directions corresponding to the available cloud. The final section presents the results of the tests aiming the comparison between the performances of our method and the standard *k-means* clustering technique when they are applied to the initial space as well as to compressed data.

1 INTRODUCTION

Basically, a cluster analysis method can be viewed as an unsupervised learning technique and usually it is a pre-processing step in solving a pattern recognition problem. The objective of cluster analysis is simply to find a convenient and valid organization of the data, not to establish rules for separating future data into categories.

The most intuitive and frequently used criterion function in partitional clustering techniques is the squared error criterion, which tends to work well with isolated and compact clusters. The *k*-means is the simplest and most commonly used algorithm employing a squared error criterion (McQueen 1967).

The aim of the present paper is to propose a new kind of approach in cluster analysis. Our attempt is based on group skeletons defined by a set of orthogonal and unitary eigen vectors (principal directions) of the sample covariance matrix. According to the well known result established by Karhunen and Loeve, a set of principal directions corresponds to the maximum variability of the "cloud" from metric point of view, as well as from informational point of view. The performance of our algorithm is tested against the *k-means* method in the initial representation space as well as in the reduced space of features given by principal directions. In our approach the skeleton of a group is represented by the principal directions of this sample.

Since similarity is fundamental to the definition of a cluster, a measure of the similarity between two

State L., Cocianu C., Rosca I. and Vlamos P. (2008).
A NEW LEARNING ALGORITHM FOR CLASSIFICATION IN THE REDUCED SPACE.
In *Proceedings of the Tenth International Conference on Enterprise Information Systems - AIDSS*, pages 155-160
DOI: 10.5220/0001676501550160

patterns drawn from the same feature space is essential to most clustering procedures. It is most common to calculate the dissimilarity between two patterns using a distance measure defined on the feature space. The dissimilarity measure used in our method is defined in terms of the Euclidian distance between the group skeletons.

Our developments impose a set of quite natural working assumptions on the true but unknown nature of the class system. The search process for the optimal clusters approximating the unknown classes towards getting homogenous groups, where the homogeneity is defined in terms of the "typicality" of components with respect to the current skeleton. Our method is described in the third section of the paper. The final section presents the results of the tests aiming to derive comparative conclusions abut the performances of our method and the *k-means* in the initial representation space and the reduced spaces.

2 A SKELETON-BASED DISSIMILARITY MEASURE

Let us assume that the recognition task is formulated as a discrimination problem among *M* classes or hypothesis. We denote by *H* the set of hypothesis. The Bayesian point of view is usually expressed in terms of an a priori probability distribution ξ on *H*, where for each $h \in H$, $\xi(h)$ stands for the probability of getting an example coming from class *h*.

In the supervised framework, for each class h, a sample of examples coming from this class $\{X_1^{(h)}, X_2^{(h)},..., X_{N_h}^{(h)}\}$ is available. We denote by

$$\aleph = \bigcup_{h \in H} \{X_1^{(h)}, X_2^{(h)},..., X_{N_h}^{(h)}\},\ N = \sum_{h \in H} N_h \ .$$

Therefore, each element of $\aleph$ can be viewed as a tagged component, where the tag is the label of the provenience class. For each class, the sample mean $\mu_{N_h}^{(h)} = \frac{1}{N_h} \sum_{i=1}^{N_h} X_i^{(h)}$ can be viewed as a template or prototype for the class which typicality depends on the variability existing within the sample. The components of the sample covariance matrix

$$\Sigma_{N_h}^{(h)} = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} \left(X_i^{(h)} - \mu_{N_h}^{(h)}\right)\left(X_i^{(h)} - \mu_{N_h}^{(h)}\right)^T$$

express the global correlations between the attributes measured in the representation space with respect to the sample coming from class *h*. Therefore, the variability degree of each class h is usually expressed in terms of a real valued function *f* of $\mu_{N_h}^{(h)}$ and $\Sigma_{N_h}^{(h)}$.

The global prototype and overall sample covariance matrix are given by the mixture of $\{(\mu_{N_h}^{(h)}, \Sigma_{N_h}^{(h)}), h \in H\}$ with respect to ξ, that is

$$\mu_N = \sum_{h \in H} \xi(h) \mu_{N_h}^{(h)} \quad (1)$$

$$\Sigma_N = \sum_{h \in H} \xi(h) \Sigma_{N_h}^{(h)} \quad (2)$$

The value of $f(\mu_{N_h}^{(h)}, \Sigma_{N_h}^{(h)})$ represents a measure of the overall variability existing in the "cloud" $\aleph$. In cases the probability distribution ξ is unknown, it is usually estimated by the relative frequencies, that is, for each $h \in H$, $\xi(h) \approx \frac{N_h}{N}$.

In the unsupervised case, the available data is represented by $\aleph = \{X_1, X_2,..., X_N\}$, an untagged set of examples of a certain volume *N*, coming from the classes of *H*. The task is to develop suitable algorithm to identify the groups of examples coming from each class. Usually, these groups are referred to as clusters. The problem is usually solved using a conventional dissimilarity measure defined in terms of the measured attributes, whose value for each pair of examples expresses in which extent these examples "are different".

In our attempt we define a dissimilarity measure to express the fitness degree of an element with respect to a cluster by a function expressing a measure of disturbance of cluster structure induced by the decision of including this element into the given cluster. Our developments are based on the following set of working assumption.

1. Each data of $\aleph$ is the realization of a certain random vector corresponding to an unique but unknown class of the set *H*. Let $M = |H|$, wher $|H|$ stands for the number of elements of *H*. We assume that *M* is known.

2. The classes are well separated in the representation space $\mathbf{R}^n$.

3. For each class $k \in H$, it is available an example P_k coming from this class

The idea behind our approach is to use the skeletons as basis in developing the search for *M*-homogenous groups starting with $P_1, P_2,.., P_M$ as initial *seeds*. The closeness degree of a particular data *X* to a cluster *C* is measured by the distance between skeletons of C and $C \cup \{X\}$. From

intuitive point of view, in case C includes mostly elements coming from the same class k, C results homogenous, and for X coming from k, the distance between C and $C \cup \{X\}$ is negligible.

The search process allots/re-allots data to the current set of clusters aiming to produce M clusters as homogenous as possible. The computation of the distance between the skeletons of C and $C \cup \{X\}$ can be simplified using first order approximation as follows. If $C = \{X_1, X_2, ..., X_r\}$, the sample means and the sample covariance matrices of C and $C \cup \{X\}$ are given by,

$$\mu_r = \frac{1}{r}\sum_{i=1}^{r} X_i \quad (3)$$

$$\mu_{r+1} = \frac{r}{r+1}\mu_r + \frac{1}{r+1}X \quad (4)$$

$$\Sigma_r = \frac{1}{r-1}\sum_{i=1}^{r}(X_i - \mu_r)(X_i - \mu_r)^T \quad (5)$$

$$\Sigma_{r+1} = \Sigma_r + \frac{1}{r+1}(X - \mu_r)(X - \mu_r)^T - \frac{1}{r}\Sigma_r \quad (6)$$

Let $\lambda_1^r \geq \lambda_2^r \geq ... \geq \lambda_n^r$ be the eigen values and let $\psi_1^r, ..., \psi_n^r$ be the orthonormal eigen vectors of Σ_r. In case the eigen values of Σ_r are pairwise distinct, the following first order approximations of the eigen values and eigen vectors of Σ_{r+1} hold,

$$\lambda_i^{r+1} = \lambda_i^r + (\boldsymbol{\psi}_i^r)^T \Delta\boldsymbol{\Sigma}_r \boldsymbol{\psi}_i^r = (\boldsymbol{\psi}_i^r)^T \boldsymbol{\Sigma}_{r+1} \boldsymbol{\psi}_i^r \quad (7)$$

$$\boldsymbol{\psi}_i^{r+1} = \boldsymbol{\psi}_i^r + \sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{(\boldsymbol{\psi}_j^r)^T \Delta\boldsymbol{\Sigma}_r \boldsymbol{\psi}_i^r}{\lambda_i^r - \lambda_j^r} \boldsymbol{\psi}_j^r \quad (8)$$

where $\Delta\Sigma_r = \Sigma_{r+1} - \Sigma_r$.

The closeness degree of X to C is defined by

$$D(X, C) = D(X, \Psi^r) = \frac{1}{n}\sum_{j=1}^{n} \left\| \psi_j^{r+1} - \psi_j^r \right\|_2 , (9)$$

where $\| \ \|_2$ stands for the Euclidian norm in $\mathbf{R}^n$.

Obviously, the performance in time of any unsupervised classification method is strongly dependent on the dimension of the input data. Consequently, the decrease of the input data dimension by some sort of compression scheme could become worth from time efficiency point of view. However, any dimensionality reduction scheme implies missing information therefore the accuracy could become dramatically affected. Therefore, in real cluster analysis task, getting a tradeoff between accuracy and efficiency by selecting the most informational features becomes extremely important. In case of unsupervised cluster analysis, the features have to be extracted exclusively from the available data.

3 THE DESCRIPTION OF THE PROPOSED CLUSTER ANALYSIS SCHEME

The aim of this section is to present a new unsupervised classification scheme (SCS) based on cluster skeletons. The input is represented by:

- the data $\aleph = \{X_1, X_2, ..., X_N\}$ to be classified;
- M, the number of clusters;
- the set of initial seeds, $P_1, ..., P_M$.

Parameters:

- n, the dimension of input data;
- θ, the threshold value to control the cluster size; $\theta \in (0,1)$;
- nr, the threshold value to control the cluster homogeneity;
- *Cond*, the stopping condition, expressed in terms of the threshold value *NoRe*, for the number of re-allotted data;
- ρ, the control parameter, $\rho \in (0,1)$, to control the fraction of "disturbing" elements identified as outliers and removed from clusters.

P1. The Generation of the Initial Clusters,
$\mathbf{C}^0 = \{C_1^0, C_2^0, ..., C_M^0\}$, $C_k^0 = \{P_k\}, k = 1, ..., M$
The initial clusters are determined around the seeds using a minimum distance criterion.

P2. Compute the System of Cluster Skeletons,
$\mathbf{S}^t = \{S_1^t, ..., S_M^t\}$, where $S_k^t = \{\psi_{k,1}^t, \psi_{k,2}^t, ..., \psi_{k,n}^t\}$ is the skeleton of the cluster k at the moment t. We denote by $S_k^{t,i} = \{\psi_{k,1}^{t,i}, \psi_{k,2}^{t,i}, ..., \psi_{k,n}^{t,i}\}$ the skeleton of $C_k^t \cup \{X_i\}$, $1 \leq i \leq N$.

P3.

REPEAT
t=t+1; $\mathbf{S}^t = \mathbf{S}^{t-1}$; $\mathbf{C}^t = \mathbf{C}^{t-1}$;
{Compute the new cluster system
$\mathbf{C}^t = \{C_1^t, C_2^t, ..., C_M^t\}$}
for $k = 1, ..., M$
{compute the cluster C_k^t}
$C_k^t = \emptyset$;

P3.1.

for $i = \overline{1, N}$

for $cl = \overline{1, M}$ compute $D(X_i, S_{cl}^t)$;
endfor
compute $l = \arg\min_{1 \le cl \le M} D(X_i, S_{cl}^t)$;
if $k=l$ **then**
$C_k^t \leftarrow C_k^t \cup \{X_i\}; C_p^t \leftarrow C_p^t \setminus \{X_i\}$,
where p is such that $X_i \in C_p^t$
endif
endfor

P3.2. {test the homogeneity of C_k^t }

compute c_k^t the center of C_k^t; $c_k^t = \frac{1}{|C_k^t|} \sum_{X \in C_k^t} X$

re-compute S_k^t, the skeleton of C_k^t;

compute

$$F_1 = \left\{ X \in C_k^t \Big/ \left\| X - c_k^t \right\|_2 > \theta \max_{X \in C_k^t} \left\| X - c_k^t \right\|_2 \right\};$$

compute $F_2 = \left\{ X \in C_k^t \Big/ \exists j \neq k, D(X, S_k^t) > D(X, S_j^t) \right\}$;

if $|F_1 \cup F_2| > nr$ **then** C_k^t is not homogenous

else C_k^t is homogenous

endif

P3.3. {extend C_k^t in case it is homogenous by adding the closest elements }

if C_k^t is homogenous **then**
for each $X \in \aleph \setminus C_k^t$
for $cl = 1,..., M$ compute $D(X, S_{cl}^t)$
endfor
compute $l = \arg\min_{1 \le cl \le M} D(X, S_{cl}^t)$;
if $k=l$ **then**
$C_k^t \leftarrow C_k^t \cup \{X_i\}$, $C_p^t \leftarrow C_p^t \setminus \{X_i\}$,
where p is such that $X_i \in C_p^{t-1}$
endif
endfor
else { C_k^t is not homogenous }
$elim = \rho |F|$;
compute *SET*1 the set of the most "disturbing" *elim* elements from *F* (identified as outliers with respect to C_k^t)
{elements of maximum distance to S_k^t }
for each $X \in SET1$
for $cl = 1,..., M$ compute $D(X, S_{cl}^t)$;
endfor
compute $l = \arg\min_{1 \le cl \le M} D(X, S_{cl}^t)$;
if $l<>k$ **then**
$C_l^t \leftarrow C_l^t \cup \{X\}$; $C_k^t \leftarrow C_k^t \setminus \{X\}$;
endif
endfor
endif

P3.4.

re-compute S_k^t, the skeleton of the new C_k^t ;

P3.5. {re-allot the elements of $C_k^{t-1} \setminus C_k^t$ }

for each $X \in C_k^{t-1} \setminus C_k^t$
for $cl = 1,..., M$ compute $D(X, S_{cl}^t)$
endfor
compute $l = \arg\min_{1 \le cl \le M} D(X, S_{cl}^t)$;
$C_l^t \leftarrow C_l^t \cup \{X\}$;
endfor

P3.6.

Compute the new set of skeletons $\mathbf{S}^t$
{the computation of C_k^t is over}
endfor
UNTIL *Cond*

The use of the previously presented classification scheme combined with a compression applied to reduce data dimensionality can be developed either by compressing with respect to the overall principal directions (variant 1) or with respect to the principal directions of each initial cluster (variant 2).

Set the value of m, $1 < m < n$,

Variant 1. The overall compression

1.1. Determine the principal directions of the initial data $\aleph$ using μ_N and Σ_N given by (1) and (2).

1.2. Get the m-dimensional representation $\aleph^m$ of $\aleph$ by projecting the components of $\aleph$ on the m-dimensional subspace represented by the first m principal directions

1.3. Apply the classification scheme to $\aleph^m$.

Variant 2. Cluster compression

2.1. Apply P1 to get the initial system of clusters $\mathbf{C}^0 = \{C_1^0, C_2^0, ..., C_M^0\}$

2.2. Determine the principal directions for each cluster of $\mathbf{C}^0$.

2.3. Get the compressed m-dimensional versions of data by compressing each element with respect to systems of principal directions corresponding to the cluster it belongs to.

2.4. Get $\aleph^m$ as the union of the resulted *m*-dimensional versions.

2.5. Apply the classification scheme to $\aleph^m$

4 EXPERIMENTAL PERFORMANCE EVALUATION OF THE PROPOSED ALGORITHM

A series of tests were performed in order to derive conclusions about the performance of our method as well as to test its performance against the *k-means* algorithm. The stopping condition *Cond=True* holds if IN the current iteration resulted at most *NoRe* re-allots; in our tests *NoRe* was set to *NoRe*=10. The tests were performed for *M*=4, the data being randomly generated by sampling from normal repartitions. Some of the repartitions were selected to correspond to "well separated" classes some others being generated to correspond to "bad separated" subsets of classes, the working assumption 2 not being necessarily fulfilled.

In order to obtain conclusions concerning algorithm sensitivity to data dimensionality, several tests were performed for *n*=2, *n*=4, *n*=6, *n*=8, *n*=10. The tests on our algorithm and *k-means* pointed out the following conclusions.

1. In cases when there is a natural grouping tendency in data, the initial system of skeletons is pretty close to the true one. In these cases, our algorithm gets stabilized in a small number of iterations.

2. In case of data of relatively small size, the number of misclassified components by our algorithm is significantly less than the number of misclassified data using *k-means*.

3. In cases of data of relatively small size, the performance of *k-means* algorithm in identifying the cluster structures is significantly less than the performance of our method.

4. The *k-means* algorithm is significantly more sensitive to data dimensionality, its performance decreasing dramatically as the dimension *n* increases.

5. In case of large sample sizes, the performance of our method is comparable to the performance of *k-means*.

Several tests were performed for "well separated" classes, relatively "separated" and "bad separated" respectively. In all tests, the performance of *k-means* proved moderated, while our method managed to identify the class structures and to correctly classify most of data. The closeness degree between the classes is computed in terms of the Mahalanobis distance.

Some of the results are reported below.

A. *M*=4, *n*=4 and relative small size data. The classes are weakly separated , the values of the Mahalanobis distances are

$$\begin{pmatrix} 0 & 428.1542 & 214.5993 & 351.6289 \\ 428.1542 & 0 & 265.3931 & 846.1386 \\ 214.5993 & 265.3931 & 0 & 369.9349 \\ 351.6289 & 846.1386 & 369.9349 & 0 \end{pmatrix}$$

In this case, the classification scheme managed to discover the true structure of data in the initial space, but using the compression for $m = 3$ and $m = 2$ its performance degraded dramatically. The *k-means* algorithm did not manage to identify the existing structure in the initial space. Some of he results are summarized in the following table.

Note that for the samples S_1 , S_2 and S_4 the *k-means* failed to identify the cluster structures.

Table 1: The comparison of our method against *k-means*.

The sample	S_1	S_2	S_3	S_4
Number of misclassified examples by our method	0	2	0	0
Number of misclassified examples by k-means	276	253	19	311
Number of iterations	3	2	2	2

B. *M*=4, *n*=4 and relative small size data. In this case, the true classes are better separated. The values of the Mahalanobis distances are

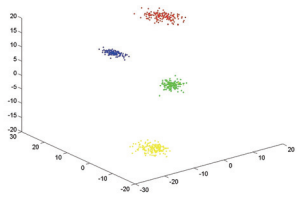
$$10^3 \begin{pmatrix} 0 & 0.4183 & 0.4139 & 0.9733 \\ 0.4183 & 0 & 0.2827 & 1.19 \\ 0.4139 & 0.2827 & 0 & 0.6171 \\ 0.9733 & 1.19 & 0.6171 & 0 \end{pmatrix}$$

In this case good results were obtained by applying the proposed classification scheme in the initial space as well as for m=3. All tests proved better performances of our method as compared to *k-means*. Some of the results are summarized in the following table.

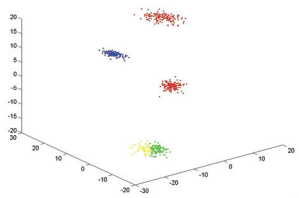
Table 2: The comparison of our method against *k-means*.

The sample	S_1	S_2	S_3	S_4	S_5
Number of misclassified examples by our method	0	0	0	0	0
Number of misclassified examples by k-means	315	0	325	318	0
Number of iterations	2	2	3	2	2

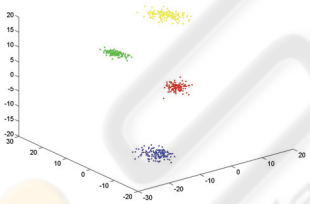
The 3-dimensional representations of data corresponding to S are depicted in figure 1.



a. The true system of classes



b. The clusters produced by *k-means* algorithm



c. The clusters computed by our method

Figure 1: The results on the sample S_1.

REFERENCES

Cocianu, C., State, L., Rosca, I., Vlamos, P, 2007. A New Adaptive Classification Scheme Based on Skeleton Information. In *Proceedings of ICETE-SIGMAP 2007*, Spain**.**

Diamantaras, K.I., Kung, S.Y., 1996. *Principal Component Neural Networks: theory and applications*, John Wiley &Sons

Everitt, B. S., 1978. *Graphical Techniques for Multivariate Data*, North Holland, NY

Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R., 1996. *Advances in Knowledge Discovery and Data Mining*, AAAI Press/MIT Press, Menlo Park, CA.

Goldberger, J., Roweis, S., Hinton, G., Salakhutdinov, R., 2004. Neighborhood Component Analysis. In *Proceedings of the Conference on Advances in Neural Information Processing Systems*

Gordon, A.D. 1999. *Classification*, Chapman&Hall/CRC, 2nd Edition

Hastie, T., Tibshirani, R., Friedman, J. 2001. *The Elements of Statistical Learning Data Mining, Inference, and Prediction.* Springer-Verlag

Hyvarinen, A., Karhunen, J., Oja, E., 2001. *Independent Component Analysis*, John Wiley &Sons

Jain,A.K., Dubes,R., 1988. *Algorithms for Clustering Data*, Prentice Hall,Englewood Cliffs, NJ.

Jain, A.K., Murty, M.N., Flynn, P.J. 1999. Data clustering: a review. *ACM Computing Surveys*, Vol. 31, No. 3, September 1999

Liu, J., and Chen, S. 2006. Discriminant common vectors versus neighborhood components analysis and Laplacianfaces: A comparative study in small sample size problem. *Image and Vision Computing* 24 (2006) 249-262

MCQueen, J. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics andProbability*, 281–297.

Panayirci,E., Dubes,R.C., 1983. A test for multidimensional clustering tendency. *Pattern Recognition*,16, 433-444

Smith,S.P., Jain,A.K., 1984. Testing for uniformity in multidimensional data, In *IEEE Trans.`Patt. Anal.` and Machine Intell.*, 6(1),73-81

State L., Cocianu C. 1997. The computation of the most informational linear features, *Informatica Economica*, Vol. 1, Nr. 4

Ripley, B.D. 1996. *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge.