International Journal of Medical, Medicine and Health Sciences
ISSN: 2517-9969
Vol:2, No:5, 2008

Observations about the Principal Components Analysis and Data Clustering Techniques in the Study of Medical Data

Cristina G. Dascălu, Corina Dima Cozma, and Elena Carmen Cotrutz

***Abstract*—**The medical data statistical analysis often requires the using of some special techniques, because of the particularities of these data. The principal components analysis and the data clustering are two statistical methods for data mining very useful in the medical field, the first one as a method to decrease the number of studied parameters, and the second one as a method to analyze the connections between diagnosis and the data about the patient's condition. In this paper we investigate the implications obtained from a specific data analysis technique: the data clustering preceded by a selection of the most relevant parameters, made using the principal components analysis. Our assumption was that, using the principal components analysis before data clustering - in order to select and to classify only the most relevant parameters – the accuracy of clustering is improved, but the practical results showed the opposite fact: the clustering accuracy decreases, with a percentage approximately equal with the percentage of information loss reported by the principal components analysis.

***Keywords*—**Data clustering, medical data, principal components analysis.

I. Introduction

THE most common difficulties met in the medical data statistical analysis come from the following facts: we have to deal, in the most cases, with a very large number of parameters, and the parameters are often very different in their nature – because, in order to establish a correct and accurate diagnosis, the physician needs to make many analysis, to observe many parameters that characterize the patient's condition and to get information using all the possible sources. Therefore, the physician has to manage a lot of data, very different in their nature: numerical values of different body markers, images, sounds (often also characterized by quantitative measures) and a lot of qualitative values – which are codes for the physician's observations, more or less subjective in their intimate structure, about the disease's anamnesis. In this context, a correct diagnosis come mostly from the physician's experience, often accumulated during many years of practice. The computer is not able yet to supply the knowledge, intuition and experience of a good physician, but the researchers work in order to create the so-called "expert systems" – computerized applications which are able to establish the diagnosis based on the values of a large amount of parameters collected and regarding the patient's condition.

A useful tool in this purpose is the data clustering, used in the following manner: we record first all the medical parameters that characterize a disease or a class of diseases, and we try to classify them in a number of clusters equal with the number of possible diagnosis, knowing also the right diagnosis for each record; in this way we find the clustering's accuracy, often expressed in percentages. Then, if the percentage is not good enough, we can change the clustering algorithm, or we can change the set of analyzed parameters by adding or removing them, until we obtain the best percentage of accuracy. After this stage we can use the obtained procedure in order to establish automatically the diagnosis for new patients, using only the values of the previous selected parameters and the clustering algorithm.

In order to obtain a good tool for automated diagnosis, the principle is to use the most significant medical parameters, without any concern about their amount (because, with the increasing of their number, theoretically the method's accuracy must improve). Even more, using this method we can find also new correlations between medical parameters specific for a certain disease, which can lead to new research directions in the medical field. The only problem with this method is generated by the long time necessary for the data processing, especially when we deal with many medical parameters and large databases of patients (with a few thousand of cases). Therefore, a possible question is: How to reduce the number of parameters selected for analysis, without the decreasing of the clustering's accuracy?

On the other hand, the principal components analysis is a statistical method well-known as very efficient for data reduction – being given a number of parameters, this technique helps us to select the most relevant parameters, with

Manuscript received October 24, 2006.

C.G. Dascălu, Ph.D., Lecturer, is with the University of Medicine and Pharmacy "Gr. T. Popa", Iaşi, Romania – The Medical Informatics and Biostatistics Department, Faculty of Dental Medicine (phone: 0040-232-206441, e-mail: cdascalu@umfiasi.ro).

C. Dima Cozma, MD, Assistant Professor, is with the University of Medicine and Pharmacy "Gr. T. Popa", Iaşi, Romania – The VI-th Medical Section, Faculty of Medicine (e-mail: cm72@email.ro).

E.C. Cotrutz, MD, Ph.D., Professor, is with the University of Medicine and Pharmacy "Gr. T. Popa", Iaşi, Romania – The Cell Biology Department, Faculty of Medicine (e-mail: cotrutz@yahoo.com).

a minimal loss of information.

Taking in consideration the specificity of these techniques, we assumed that, selecting the most relevant medical parameters by a principal component analysis and then clustering only those parameters, the final result will be an improving of the clustering accuracy, and we intended to check this supposition in a practical situation. The obtained results are presented in the following sections.

II. Material and Method

A. The Principal Components Analysis Method

This method of data analysis, described by Pearson (1901) and Hotelling (1933), concerns the finding of the best way to represent n samples by using vectors with p variables, in such a manner so the similar samples are represented by points as close as possible. In order to find the principal components from a set of variables, the method used is the analysis of eigenvalues and eigenvectors, which starts from a data representation using a symmetrical matrix and transforms it.

Let $X = \{x^1, x^2, ... x^p\}$ be a set of points that makes a cloud in the R^n space [1]. We try to find the directions u^1, u^2, ... where the cloud's dispersion is maximal. To do this, we find the cloud's centroid and the lines L_1, L_2, ... around which the cloud's points are closely grouped and pass through the centroid, so the directions u^1, u^2, ... will be those lines directions. Denoting by d_j the distance from the x^j point to a line L, the problem is to find the line where the quantity $J = \sum_{j=1}^{p} d_j^2$ is minimal. It is demonstrated that this line pass through the cloud's centroid [2]. Making a data normalization by the translation $x' = x - m$, $x \in X$ (m = the average value of the points $x \in X$), the cloud X becomes X', with average 0 and the centroid situated in the coordinate system's origin; denoting by u the direction vector of the searched line, $\| u \| = 1$, the criteria function becomes $J : R^n \to R$:

$$J(u) = \sum_{j=1}^{p} d^2(x^{j'}, u) = \sum_{j=1}^{p} \left(\|x^{j'}\|^2 - (x^{j'}, u)^2 \right)$$

$$\Rightarrow J(u) = \sum_{j=1}^{p} \|x^{j'}\|^2 - \sum_{j=1}^{p} (x^{j'}, u)^2$$

The minimization of the function J is then equivalent with the maximization of the function $I : R^n \to R$, defined by:

$$I(u) = \sum_{j=1}^{p} (x^{j'}, u)^2 = \sum_{j=1}^{p} (u^T M x^{j'})(x^{j'T} M u)$$

$$\Rightarrow I(u) = u^T \left[M \left(\sum_{j=1}^{p} x^{j'} x^{j'T} \right) M \right] u$$

Because $\| u \| = 1$ we determine for the I function the maximal value for the unit sphere vectors. The quadratic matrix with d order, $S = M \left(\sum_{j=1}^{p} x^{j'} x^{j'T} \right) M$, is the spreading matrix of the normalized points cloud. The function's $I(u) = u^T S u$ extreme values on the unit sphere vectors are identical with the proper vectors of the S matrix, and the proper vector corresponding to the biggest proper value of S is identical with the direction where I(u) is maximal.

The proper vectors of the S matrix, denoted by u^1, u^2, ... u^n, and taken in descendant order of their corresponding proper values, are the principal directions, or the principal components of the points cloud and shows the cloud's orthogonal directions. The cloud is maximally extended in the u^1 direction (main proper vector).

B. The Data Clustering: The Hierarchical Clustering and the K-Means Methods

The general algorithm of hierarchical clustering was proposed by S.C. Johnson and has the purpose to built a chain of partitions based on the set of input data and an ultra-metric distance, in such a way so, at each step, the diameters of the classes are growing. We denote the input data to classify by X:

$X = \{ x^1, x^2, x^3, ... x^p \}$.

The algorithm is [3, 4]:

Step 0:

Let P^0 be the discrete partition, $P^0 = \{ P_1^0, P_2^0, ...P_p^0 \}$, where $P_i^0 = \{ x^i \}$, i = 1, 2, ... p

Let denote:

$\delta (P_i^0, P_j^0) = \delta (x^i, x^j)$ – the ultrametric distance,

$\nu_0 = 0$

$L^0 = \{1, 2, ...n\}$ – the set of the indexes of found partitions;

Step t (t ≥ 1):

1. Find $\nu_t = \min \{\delta (P_i^{t-1}, P_j^{t-1}) \mid P_i^{t-1}, P_j^{t-1} \in P^{t-1}, i < j \}$
2. Define the coefficients sets of the identified pairs:
   $C^t = \{(i, j) \mid i, j \in L^{t-1}, i < j, \delta (P_i^{t-1}, P_j^{t-1}) = \nu_t \}$
   $I^t = \{i \in L^{t-1} \mid i < j, (i, j) \in C^t \}$
   $J^t = \{j \in L^{t-1} \mid i < j, (i, j) \in C^t \}$

(the elements identified by these coefficients will be further tested, because they will generate the searched classes).

3. Define the working variables:
   $I(1) = I^t$ - the set of coefficients to check
   $L(t) = \varnothing$ - the set of coefficients already checked
   r = 1
4. Find:
   $i_r = \min \{i \in I(r)\}$
   $J(r) = \{j \mid (i_r, j) \in C^t \}$
   $P_{ir}^t = P_{ir}^{t-1} \cup \{ P_j^{t-1} \mid j \in J(r) \}$
5. Update the working sets – by moving the i_r coefficient:
   $I(r+1) = I(r) - \{i_r\}$
   $L(t) = L(t) \cup \{i_r\}$
6. If $I(r+1) = \varnothing$, go to step 7 (all the coefficients are checked).

   Otherwise, r = r + 1; go back to step 4.
7. $L^t = L(t) \cup \{i \in L^{t-1} - (I^t \cup J^t)\}$ – update the coefficients set

   $\delta (P_i^t, P_j^t) = \delta (P_i^{t-1}, P_j^{t-1})$, i, j $\in L^t$, i < j – update the ultra-metric distance

   $P_i^t = P_i^{t-1}$, if $i \in L^{t-1} - (I^t \cup J^t)$

   $P^t = \{ P_i^t \mid i \in L^t \}$ – the partition found at step t
8. The algorithm's stop condition:

If $|L^t| = 2$, define $P^{t+1} = \{X\}$

$L = \{P^0, P^1, \ldots P^{t+1}\}$, STOP

If $|L^t| > 2$, define t = t +1 and repeat the step t.

This general algorithm can be used in different variants, according with the formula of the ultra-metric distance used; one of these variants is the so-called Average Linkage / Group Average Algorithm.

The Average Linkage algorithm uses the distance between classes:

$d_{med}(A_r, A_s) = \frac{1}{p_r p_s} \sum_{x \in A_r} \sum_{y \in A_s} d(x, y)$,

where p_i = number of elements in the A_i class.

The distance d_{med} can also be written as:

$d_{med}(A_r, A_s,) = \frac{1}{\sum_j A_{rj} \sum_k A_{sk}} \sum_j \sum_k A_{rj} A_{sk} d(x_j, y_j)$

where $A_{rj} = \begin{cases} 1, if\ x_j \in A_r \\ 0, otherwise \end{cases}$, and d(x, y) is a pseudodistance.

We notice that d minimize an average squared error. Denoting by:

$u_{jk} = \begin{cases} 1, if\ x_j \in A_r\ and\ x_k \in A_s \\ 0, otherwise \end{cases}$

$\Rightarrow u_{jk} = A_{rj} \cdot A_{sk}$,

we define the squared error $\sum u_{jk} (d(x_j, x_k) - d)^2$, which measures the deviation from the d distance of the distances between points. The condition of minimal squared error becomes:

$-2 \cdot \sum u_{jk} (d(x_j, x_k) - d) = 0$

$\Rightarrow d = \frac{\sum_{j,k} u_{jk} d(x_j, x_k)}{\sum_{j,k} u_{jk}} = d(A_r, A_s)$

This method has the following advantages:

1) If the data set contains a partition where the distances between the classes points are smaller than the distances between classes, then that partition is accurately detected.

2) If the data set has a tree-like structure, then the method detects this hierarchy.

3) Let $P = \{A_1, \ldots, A_m\}$ a partition of X set, obtained by this method. If we remove from X the elements of the A_i class, then the method applied to the rest of data leads to a partition with m – 1 classes, identical with the P partition without the A_i class.

The n-means algorithm is another well-known method used for data clustering, which has even better results than the general hierarchic clustering.

We denote, like in the previous case, by $X = \{x^1, x^2, \ldots x^p\}$ the set of the objects to cluster, which are represented as vectors in the n-dimensional Euclidean space. The used dissimilarity measure will be:

$D(x, y) = \| x - y \|^2$.

We assume that the X set is made by clusters approximately compact and well-differentiated, which can be represented by single points – the classes prototypes, denoted by $L_i \in R^n$, corresponding to the class C_i. We also know the number of clusters we need to obtain.

The dissimilarity between a point $x \in X$ and the prototype L_i is the error made when we approximate the point x by the prototype L_i, being expressed as:

$D(x, L_i) = \| x - L_i \|^2$.

Let also be I_C the characteristic function of the C set, and:

$C_{ij} = I_{Ci}(x^j) = \begin{cases} i, if\ x^j \in C_i \\ 0, otherwise \end{cases}$.

The criteria function will be:

$J(P, L) = \sum_{i=1}^{m} I(C_i, L_i) = \sum_{i=1}^{m} \sum_{x \in C_i} D(x, L_i)$

$\Rightarrow J(P, L) = \sum_{i=1}^{m} \sum_{x \in C_i} \|x - L_i\|^2$

$\Rightarrow J(P, L) = \sum_{i=1}^{m} \sum_{j=1}^{p} C_{ij} \cdot \|x^j - L_i\|^2$

Into a Euclidean space the scalar product has the form:

$(x, y) = x^T M y$

$\Rightarrow J(P, L) = \sum_{i=1}^{m} \sum_{j=1}^{p} C_{ij} \cdot (x^j - L_i)^T M (x^j - L_i)$

We must find the L representation which is a minimum for the care $J(P, \cdot)$ function:

$\frac{\partial J(P, L)}{\partial L_i} = 0$, $i \in \{1, 2, \ldots m\}$

$\Rightarrow -2 \sum_{j=1}^{p} C_{ij} M (x^j - L_i) = 0$ (the M matrix is not singular; we multiply the relation by M^{-1})

$\Rightarrow \sum_{j=1}^{p} C_{ij} x^j - \sum_{j=1}^{p} C_{ij} L_i = 0$

$\Rightarrow L_i = \frac{\sum_{j=1}^{p} C_{ij} x^j}{\sum_{j=1}^{p} C_{ij}}$, $i \in \{1, 2, \ldots m\}$

We denote the number of elements in the C_i class with p_i, $p_i = \sum_{j=1}^{p} C_{ij}$, therefore $L_i = \frac{1}{p_i} \sum_{x \in C_i} x$.

The L_i prototype is the centroid of the C_i class.

The representation $L = \{L_1, L_2, \ldots L_n\}$ induces a new partition obtained using the closest neighbor rule: a point x^j goes into the class with the closest centroid:

$x^j \in C_i$ if $\| x^j - L_i \| < \| x^j - L_k \|$, $k \in \{1, 2, \ldots m\}$, $k \neq i$.

In order to simplify the algorithm, this rule can also be expressed as:

$C_{ij} = \begin{cases} 1, if \|x^j - L_i\| < \|x^j - L_k\|, \forall k \neq i \\ 0, otherwise \end{cases}$.

The n-means algorithm is presented below [5]:

Step 1. Let $P^0 = \{C_1, C_2, \ldots C_m\}$ be an initial partition of X.

Step 2. Calculate the prototypes of this partition,

$$L_i = \frac{\sum_{j=1}^{p} C_{ij} x^j}{\sum_{j=1}^{p} C_{ij}} = \frac{1}{p_i} \sum_{x \in C_i} x .$$

Step 3.

3.1. Calculate a new partition using the rule:

For each input vector x^j, $j \in \{1, 2, ... p\}$, add this vector into the partition with the closest prototype:

$x^j \in C_i$ if $\| x^j - L_i \| < \| x^j - L_k \|$, $\forall\, k \in \{1, 2, ... m\}$, $k \neq i$, or

$$C_{ij} = \begin{cases} 1, if \left\| x^j - L_i \right\| < \left\| x^j - L_k \right\|, \forall k \neq i \\ 0, otherwise \end{cases}.$$

3.2. Update the prototypes of the new partition.

Step 4. Calculate the error function corresponding to the new partition,

$$E = \sum_{i=1}^{m} \sum_{x_j \in C_i} \left\| x_j - L_i \right\|^2$$

If the error function is not significantly different (the new partition is identically with the previous one), STOP.

Otherwise, go to Step 3.

The problem of choosing the initial partition can be easily solved by taking an arbitrary selection of m points from the data set.

The algorithm leads to good results when the data are separated in distinct and compact classes; its performance level highly depends on the initial partition and the number of classes to generate. The main elements which decrease its efficiency are [6]:

- The problem of clusters validity: it is necessary to establish by experiments the optimal number of classes into a given dataset, excepting the situations when this number is pre-defined.
- The pseudo-gravitational effect: when the classes have very different sizes, the criteria function tends to favor the partitions that broke the bigger classes against the partitions that keep that classes united; in this way the points situated in border positions are often misclassified.
- The noise presence: isolated points, whose integration in certain classes is difficult.

The data analysis was performed in SPSS 13.0, on a dataset made using Microsoft Visual FoxPro.

III. Results

In order to check the practical efficiency of this method, we took under analysis a set of 212 patients from three categories: healthy patients (65 cases), patients with liver cirrhosis (65 cases) and patients with liver hepatitis (82 cases). These patients were included into a study about the connections between the liver diseases and the heart's health - measured using detailed electrocardiograms. The study's purpose was to find if we can extract, using the heart's activity analysis, certain conclusions about the liver's state of health.

The heart's activity was recorded by ECG, and 38 parameters were analyzed, as it follows: the diastolic blood pressure; the systolic blood pressure; the cardiac frequency; the diameter of the aorta at the ring; the diameter of the ascendant aorta; the inter-ventricular septum width; the left ventricle posterior wall width; the left ventricle mass; the right atrium diameter; the right ventricle diameter; the left ventricle mass index; the diastolic left ventricle diameter; the systolic left ventricle diameter; the shortening ratio; the diastolic volume; the systolic volume; the ejection ratio; the cardiac flow; the left atrium diameter; the E wave velocity; the A wave velocity; the E / A ratio; the iso-volume relaxation time; the E wave deceleration time; the systolic pressure into the pulmonary artery; the pulmonary artery at the ring diameter; the average pulmonary arterial pressure; TAPSE; the inferior cave vein diameter; the average arterial blood pressure; the peripheral vascular resistance; the pre-ejection time; the ejection time; the pre-ejection time / ejection time ratio; the nervous conductivity speed; QTc and the O_2 arterial satiation (CLINO and ORTO).

All these parameters were numeric, so their clustering didn't require special data transformation. Using the hierarchical clustering – average linkage between groups method, based on the Euclidean distance and a predefined number of 3 clusters, the accuracy of fitting between the generated clusters and the right diagnosis was of 76.42 % (a common value for the hierarchical clustering method). Using the n-means clustering, also with a predefined number of 3 clusters, the accuracy of fitting was better (80.66 %), according with the Table I.

TABLE I
NUMBER OF CASES IN EACH CLUSTER (1ST METHOD)

Cluster	1	51.000
	2	53.000
	3	108.000
Valid		212.000
Missing		.000

After this step, we proceeded to a principal component analysis, in order to reduce the number of parameters used for the next clustering. From the variance analysis we found a number of 12 principal components (with initial eigenvalues > 1.00), according to the Scree Plot in Fig. 1. The cumulative initial eigenvalue of these components is 75.41 % - so these components correspond to an information (variability) loss of 24.59 %, compared with the whole set of parameters.

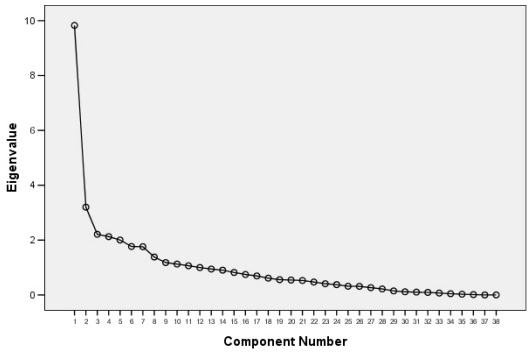

Fig. 1 The Scree Plot of the 38 parameters initial eigenvalues

By calculating the Component Scores Coefficient Matrix and selecting the parameters with the highest values, we found that the 12 principal components correspond to the following parameters: the left ventricle mass; the average pulmonary arterial pressure; the E wave deceleration time; the right ventricle diameter; the pre-ejection time / ejection time ratio; the systolic left ventricle diameter; the E wave velocity; the O_2 arterial satiation (CLINO); the systolic blood pressure; the ejection ratio; the E / A ratio and the diameter of the aorta at the ring, in this order.

Finally, we took these parameters and we proceeded to a new clustering, using only them. Using the same methods like in the previous step, we obtained the following results:
- using the hierarchical clustering – average linkage between groups method, based on the Euclidean distance and a predefined number of 3 clusters, the accuracy of fitting between the generated clusters and the right diagnosis is very weak – 47.17 %;
- using the n-means clustering, also with a predefined number of 3 clusters, the accuracy of fitting is also weak (56.13 %), according with the Table II.

TABLE II
NUMBER OF CASES IN EACH CLUSTER (2ND METHOD)

Cluster	1	51.000
	2	57.000
	3	104.000
Valid		212.000
Missing		.000

Comparing these results, it clearly follows that the number of parameters reduction by principal component analysis is not a good choice for the data clustering optimization.

IV. DISCUSSIONS AND CONCLUSION

In order to seek the reasons of this result, we turned back at the variability loss, reported by the principal components analysis; the corresponding value was 24.59 % - which is an acceptable value. Looking instead at the accuracy loss in clustering, we found the results reported in Table III.

TABLE III
THE CLUSTERING COMPARATIVE ACCURACIES

	Using all parameters (accuracy %)	*After the parameters selection (accuracy %)*	*The difference in accuracy*
The hierarchical clustering	76.42 %	47.17 %	*29.25 %*
The n-means clustering	80.66 %	56.13 %	*24.53 %*

Therefore we can conclude the following thing: the values of information loss by selecting the principal components from a set of parameters and of accuracy loss by reducing the number of parameters involved in clustering are almost equal – so, even if the information loss by selecting the principal components is acceptable – as all the statistics books suggest, this quantity is not absorbed when we proceed to any type of further data analysis. For example, in the case of data clustering, this information loss is entirely preserved and transferred negatively over the clustering's accuracy.

So, the clustering algorithms cannot be optimized by reducing the number of parameters taken into account, even if we use in this purpose classic methods (like the principal components analysis); the only possibilities in this purpose remain the structural changes in the algorithms – and, basically, the optimal choice of the dissimilarity measure.

REFERENCES

[1] Chernick, M.R., Friis, R.H., *Introductory Biostatistics for the Health Sciences*, John Wiley & Sons Publ., 2003.
[2] Zhou, X.H., Obuchowski, N.A., McClish, D.K., *Statistical Methods in Diagnostic Medicine*, John Wiley & Sons Publ., 2002.
[3] Saporta, G., Ştefănescu, M.V., *Analiza datelor şi informatică*, Ed. Economică, 1996 (in romanian).
[4] C. Dascălu, Boiculese, L., "The Usefulness of Algorithms Based on Clustering in the Diagnosis Finding in Medical Practice", in Lecture Notes of the ICB Seminars - Statistics and Clinical Practice, editors: L. Bobrowski, J. Doroszewski, E. Marubini, N. Victor, Warsaw, 2000, pg. 53 – 56.
[5] Alsabti, K., Ranka, S., Singh, V., "An Efficient K-Means Clustering Algorithm", in Proceedings of the 1st Workshop on High-Performance Data Mining, 1998.
[6] Dumitrescu, D., *Teoria clasificării*, Babeş – Bolyai University, Cluj – Napoca, 1991 (in romanian).