Diagnosing Patients Combining Principal Components Analysis and Case Based Reasoning

Carles Pous, Dani Caballero and Beatriz Lopez
University of Girona
Campus de Montilivi, Girona, Spain
carles@eia.udg.edu, u1048116@correu.udg.edu,beatriz.lopez@udg.edu

Abstract

This paper addresses the application of a PCA analysis on categorical data prior to diagnose a patients data set using a Case-Based Reasoning (CBR) system. The particularity is that the standard PCA techniques are designed to deal with numerical attributes, but our medical data set contains many categorical data and alternative methods as RS-PCA are required. Thus, we propose to hybridize RS-PCA (Regular Simplex PCA) and a simple CBR. Results show how the hybrid system produces similar results when diagnosing a medical data set, that the ones obtained when using the original attributes. These results are quite promising since they allow to diagnose with less computation effort and memory storage.

1. Introduction

Medical databases are usually constituted by a great amount of variables. So, it is tedious for the diagnostic system to cope with all these attributes. Hence, it is important to find methodologies to reduce the total amount of attributes without lose of performance. One of the well-known techniques that has been largely applied to reduce the dimensionality of numeric attributes is the Principal Component Analysis (PCA). The PCA analysis reduces data dimensionality by performing a covariance analysis between factors. The results of a PCA are usually discussed in terms of component scores and loadings [4].

Once the scores and loadings have been calculated, new patient's data have to be projected on the new $K-$dimensional space, and then apply the criteria to diagnose the patient, with any other decision support tool. In our case, we use a CBR (Case-Based Reasoning) system to diagnose the patient according to the scores and loadings provided by the PCA resulting in a hybrid PCA-CBR system. Particularly, we have used the RS-PCA (Regular Simplex PCA) method because the database where it has to be applied has categorical data. Results obtained after attribute reduction are similar to the ones obtained with the complete original data set. So, no information is lost, while the performance and storage requirements of the CBR are improved.

This paper is organized as follows. Next section explains some fundamentals of the RS-PCA and CBR methods. We continue in section 3 with the description of our hybrid PCA-CBR system. Then, section 4 give the experiments and results we have obtained when CBR uses the original data or the reduced set obtained by means of PCA. The paper ends with some conclusions and future work.

2. Background

Our research is concerned with the Regular Simplex PCA algorithm and Case-Based Reasoning systems. This section provides some fundamental issues in both fields.

2.1 The Regular Simplex PCA algorithm

The Regular-Simplex PCA (RS-PCA) method [7] is an unsupervised method that calculates covariances between categorical variables by means of the regular simplex expressions (RS). The Regular-Simplex is an extension of a regular triangle when the dimension of the space is greater than two. The definition of Regular-Simplex implies that the distance between vertexes is always equal. Each category of a variable is placed in a vertex. If the space dimension of our variable is greater than 3, the regular-simplex is a tetrahedron.

According to the definition of [7], the RS-Algorithm is based on the concept of regular simplex used in mathematics. A regular simplex is a geometrical structure analogue to a triangle in n-dimensions, with the same distance among its vertices. If we consider a categorical variable x_i, which have k_j categories, then following the definition of regular

978-0-7695-3326-1/08 $25.00 © 2008 IEEE
DOI 10.1109/HIS.2008.126

simplex, we can represent each category in a vertex of a regular simplex. To represent these vertices, we define $v^n(r_k)$ as the position of the $k-th$ vertex of a regular $(n-1)-$simplex.

Consider now that we have J variables $x_1, x_2...x_j$. For the $a-th$ instance, x_i takes the value x_{ia}. Therefore, we can represent x_{ia} as a vertex coordinates $v^{ki}(x_{ia})$ representing the value of instance a for the variable x_i. If this process is done for each variable, and the vertices coordinates are concatenated, it results in so called *List of Regular Simplex Vertices* (LRSV). It is noted as follows:

$$x(a) = (v^{k_1}(x_{1a}), v^{k_2}(x_{2a}), ..., v^{k_J}(x_{Ja})) \quad (1)$$

LRSV has a dimension $N \times M$, where N is the number of instances, and M the sum of all the categories of all the variables. With LRSV constructed, we can proceed to calculate the covariance matrix, A, which is defined as follows:

$$\mathcal{A} = \frac{1}{N}\sum_{a=1}^{N}(x(a) - \hat{x})^t(x(a) - \hat{x}) \quad (2)$$

where $\hat{x} = \frac{1}{N}\sum_{a=1}^{N} x(a)$ is an average of the LRSV. Next, the covariance matrix of the LRSV can be defined as follows:

$$\begin{bmatrix} A^{11} & A^{12} & \dots & A^{1J} \\ A^{21} & A^{22} & \dots & A^{2J} \\ \dots & \dots & \dots & \dots \\ A^{J1} & A^{J2} & \dots & A^{JJ} \end{bmatrix} \quad (3)$$

Now, with the covariance matrix calculated, we are in the same situation as the standard PCA analysis over numeric variables. PCA analysis projects n-dimensional data onto a lower-dimensional subspace in a way that is optimal in a sum-squared error sense. First, the mean vector and covariance matrix for the full data set are computed. Next, the eigenvectors and eigenvalues are computed and sorted according to decreasing eigenvalue. Finally, the k eigenvectors having the largest eigenvalues are chosen. The number of principal components to retain is normally selected by the sedimentation graph, either keeping the principal components that have an eigenvalue greater than 1, or when the slope of the sedimentation curve is almost constant.

Once the eigenvectors are chosen, the scores T for each original instance are calculated according to the following equation:

$$T = X \times P \quad (4)$$

where X is the set of LRSV, and P the eigenvectors.

2.2 Illustrative example

Let us suppose that the first attribute of our dataset has two categories, so it has each of its vertices located at [0,1] or [1,0]. Same for the second attribute. The third attribute has three categories, hence we have located the regular simplex vertices at [1,0,0], [0,1,0] and [0,0,1]. This process is done for all the variables and patients, and at the end a matrix of 1074 x 295 is obtained. A part of it is depicted in Figure 1.

	Atrib 1		Atrib 2		Atrib 3								
Patient 1	1	0	0	1	0	1	0	1	0	1 0	0	1
Patient 2	1	0	0	1	0	1	0	1	0	1 0	0	1
Patient 3	0	1	1	0	0	1	0	1	0	1 0	0	1
Patient 4	0	1	0	1	0	1	0	1	0	1 0	0	1
Patient 5	0	1	1	0	0	1	0	1	1	0 0	0	1
	.....									.....		
Patient 1072	0	1	1	0	1	0	1	0	1	0 0	0	1
Patient 1073	1	0	0	1	0	1	1	0	1	0 0	0	1
Patient 1074	1	0	0	1	0	1	1	0	1	0 0	0	1

Figure 1. LRSV vector for medical database

Then, using equation 2 and 3, the covariance matrix is computed. At this point, we are in conditions to apply the standard PCA analysis.

Once the RS-PCA is applied to our data set, we obtain the accumulated variance explained shown in table 1

Table 1. Accumulate variance explained

Component	Total variance explained
1	13.51 %
2	16.83 %
3	18.91 %
4	20.81 %
5	22.58 %
6	24.34 %
7	25.93 %

Drawing the sedimentation graph, Figure 2 is obtained. Observe that from component 5^{th} forward, the slope of the sedimentation graph is almost constant, so the number of principal components chosen is 5, which have a total explained variance of 22.58%, according to Table 1.

Once the number of principal components k is chosen, the matrix P is reduced to $k = 5$ columns, each one representing a principal component.

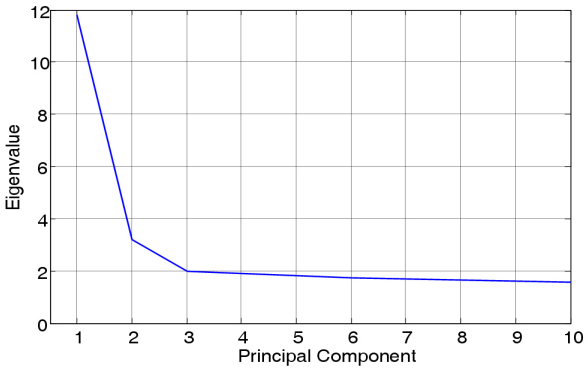

Figure 2. Screen plot of the 10 first components.

2.3 Case-based reasoning

Case Based Reasoning is an approach to problem solving that is able to use specific knowledge of previous experiences [5]. A new problem is solved by matching it with a similar past situation. If the problem is solved, this new situation will be retained in order to solve the new ones. In case of diagnosis, solving the problem means that the CBR-system proposes a solution satisfactory enough to identify the new fault. The CBR has been formalized as a four step process:

- Retrieve: Given a target problem, retrieve cases from memory that are relevant to solving it.
- Reuse: Use the solution of the retrieved case to solve the target problem.
- Revise: Having mapped the previous solution to the target situation, test the new solution in the real world and if necessary, revise.
- Retain: After the solution has been successfully adapted to the target problem, decide whether to store the resulting experience as a new case in memory.

3. Hybrid PCA-CBR Methodology

Our work concerns the hybridization of RS-PCA with CBR as shown in Figure 3, with the purpose of reducing the dimensionality of the attributes to be considered by the CBR.

Particularly, the methodology we propose is the following. First, we obtain a set of P eigenvectors for the given cases (patients)and a set of scores for each patient in the database. Therefore, we obtain a new database that is a matrix of scores. It has as many rows as patients and as many columns as principal components taken.

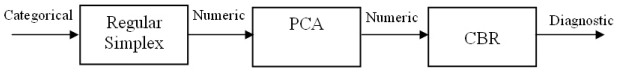

Figure 3. PCA-CBR schema.

When a new patient has to be diagnosed, the scores and the eigenvectors are calculated, and the new patient is projected on the new $K-$dimensional space. Then, the Euclidean distance between the patient's score and our matrix scores is calculated. That is, if $T_p = (T_p^1, T_p^2, ..., T_p^K)$ represents the new patient's scores and $T_i = (T_i^1, T_i^2, ..., T_i^K)$ the scores of a patient in the matrix scores, the distance D_E between them can be calculated as:

$$D_E = \sqrt{\sum_{j=1}^{K} (T_p^j - T_i^j)^2} \quad (5)$$

where K is the number of principal components chosen in the previous step.

When all the distances have been calculated, the nearest cases are taken to derive the diagnosis of the patient. This is what the reuse step does. In our case, as the paper focuses on testing the performance when using a reduced set of principal components, the reuse has been implemented simply as a voting procedure. It has been left for future works to improve this CBR step.

As a summary, the methodology that we propose is the following:

1. Calculate the covariance matrix for the actual training set.
2. Calculate the set of eigenvectors P according to the RS-PCA.
3. Project the cases (patients) to the new PCA K-dimensional space, obtaining a set of scores T for each patient in the data base.
4. Project all the instances on the test set.
5. Search the instances with minimum distance with the scores of instances on the test set.
6. Obtain the diagnostic of the instances on the test set from the selected cases.

Note, that steps 1 to 3 correspond to a classical training phase of a machine learning method, the 4 and 5 to the CBR retain phase, and the 6 to the reuse phase. We have not developed yet any further steps of CBR.

4. Experimentation

We have implemented the RS-PCA and CBR in Matlab^{TM}and tested on a medical data set. Next, we describe the preprocessing required in the data, the experimental scenario carried out, and the results obtained.

4.1 Data Preprocessing

Before starting with RS-PCA and CBR, the data set must be adapted to the experiment needs. As the motivation of the paper is to test how a CBR system performs when using projected data instead of the original attributes, non categorical variables have been not taken into account for the experiment. Also, there are some variables that contains no information related with medical data, as patient code, for example. Other variables have always the same value for all patients, giving no useful information, and hence not taken into account.

In particular, our medical database[1] contains 1074 patients, each one with 112 attributes. After removing the numerical variables and the ones with no variability or useless information, 104 attributes remain per each patient.

As many data sets, there are a lot of missing values, noise and inconsistent data. In our case, noisy and inconsistent data are replaced by missing values, and the missing values replaced by the mode of the variable. At the end, our data set contains only categorical variables, all of them containing values from 1 to 3, the categories defined per each categorical variable.

It is also important to mark which is the classification variable, the one that contains information on whether the patient is healthy or not.

4.2 Other CBR methods

In order to compare our results with the ones obtained when using the original data, as the original attributes are categorical, a second simple-CBR system has been defined with the Hamming distance for these kind of attributes. The Hamming distance is 1 if $x(a) \neq x(p)$ and 0 if $x(a) = x(p)$. The instances p with the minimum distance to the new case are taken, and used to diagnose the patient a.

The two different CBR scenarios compared are depicted in Figure 4

4.3 Evaluation methodology

In each system, the PCA-CBR and the simple-CBR, classification is performed by a $10-$fold cross-validation.

[1]The medical database used in this experiment, is not public domain.

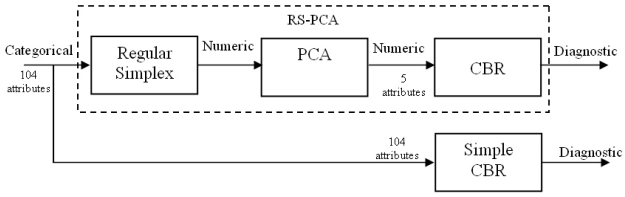

Figure 4. PCA-CBR versus simple-CBR schema.

From the 10 subsets, a single subset is retained as the validation data for testing the model, and the remaining 9 are used as training data. The cross-validation process is then repeated 10 times, for each of the 10 subsets used once as the validation data. Each $10-$fold cross-validation has been repeated 10 times, so each algorithm has been run 100 times.

Then, the 100 results obtained are averaged to produce a single estimation.

4.4 Results

Percentage of true positives obtained with PCA-CBR are shown in Figure 5. The average of successes can be seen that is approximately 89%.

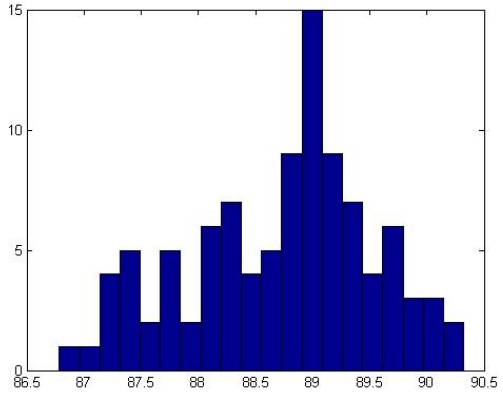

Figure 5. Results of 100 execution of RS-PCA and CBR.

Now, the results with a CBR on the original data will be compared. Making 100 times the execution of CBR, the histogram in Figure 6 have been obtained, where it is possible to be seen that the average of successes is 83.25%. So using PCA-CBR we got an increment of almost 3 points over

the simple-CBR. In addition to that, memory usage is saved since with PCA-CBR fewer values have been stored.

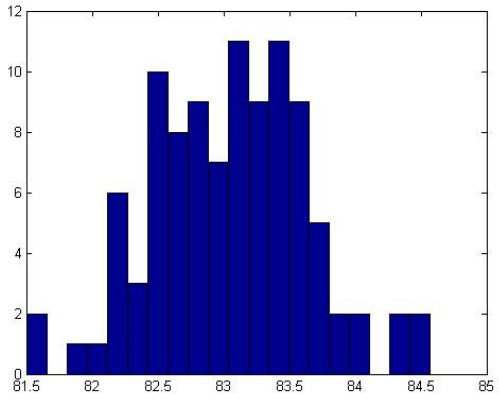

Figure 6. Results of 100 execution of CBR on original data.

5 Related work

In the literature, several works can be found that integrates PCA for attribute reduction and CBR [9] [3] [8]. But the standard PCA techniques can not deal with data with categorical variables, such as our medical data set. Hence, it is necessary to find another method that can handle categorical variables. Different PCA algorithms have been proposed in the literature, all of them designed to perform an optimal scaling of categorical data sets.

For example, the well-known statistical package SPSS uses the CATPCA (CATegorical PCA) method [6]. It consists of two stages, combined in an iterative process. In the former, a numerical variable corresponding to the initial categorical variable is constructed. Each category receives a numerical value, selected by a process of mathematical optimization, that maximizes the joint covariance of the variables that compose the index. On the latter, a proximity index is processed, as it is done in the classic method of PCA.

Multinomial PCA (MuPCA) is another method that can be found in the literature [2]. It is analogue to the standard PCA, with the difference that Multinomial PCA can handle categorical and discrete variables. However, this method is based on a parametric model, and it is difficult to represent and extract the estimated result.

Also, the Multiple Correspondence Analisys (MCA) can be cited among the methods developed to cope with categorical data. This method works like PCA with categorical and discrete data [1]. MCA is not based on a parametric model and have an easier representation of results. Nevertheless, with this method the covariance or correlations coefficients can not be directly obtained, making difficult to interpret which variables will be selected as principal components. This method is also known as homogeneity analysis, dual scaling, or reciprocal averaging.

The method RS-PCA can be analogous to MCA method. The difference between them is the way to interpret the results. The principal components in MCA can not be obtained directly, whereas in RS-PCA the covariance matrix is obtained directly. This has been the selected method for reducing the dimensionality of our medical database.

6. Conclusions and Future Work

One of the main problems of CBR is to deal with attribute dimensionality. In this paper we propose the use of RS-PCA to reduce the number of categorical data, resulting in a hybrid PCA-CBR system. The results show that the classification of patients with PCA-CBR method have an improvement with the classification through original data. This allow us to classify new patients data with a smaller amount of data, obtaining a better classification than the original data.

In future works, the possibility of including the non categorical data on the PCA has to be studied. This inclusion can allow a better classification of new patient's data, and maybe the reduction to a space with less dimensions.

As it has been mentioned, the CBR system requires further work in order to improve the results. So, different distance functions and attribute weights have to be tested for the retrieve step, and several reuse algorithms need to be experimented. Also, the numeric attributes should be included in the overall process to analyze their impact on the final results.

Acknowledgments

This work has been made possible due to the support of the Spanish MEC project DPI2005-08922-C02-02, Girona Biomedical Research In- stitute (IdiBGi) project GRCT41 and DURSI AGAUR SGR 00296 (AEDS).

References

[1] H. Abdi and D. Valentin. Multiple correspondence analysis. *Encyclopedia of Measurement and Statistics*, 2007.

[2] W. Buntine and S. Perttu. Is multinomial pca multi-faceted clustering or dimensionality reduction? *Proceedings of the*

Ninth International Workshop on Artificial Intelligence and Statistics, page 300307, 2003.

[3] J. Corchado, E. Corchado, J. Aiken, C. Fyfe, F. Fernandez, and M. Gonzalez. Maximum likelihood hebbian learning based retrieval method for cbr systems. *Proceedings of the International Conference on Case-Based Reasoning ICCBR 2003*, pages 107–121, 2008.

[4] J. Jackson. *A User's Guide to Principal Components*. John Wiley and Sons, ISBN: 978-0-471-47134-9, 2003.

[5] R. Lopez de Mantaras, , and E. Plaza. Case-based reasoning: An overview. *AI Communications*, pages 21–29, 1997.

[6] J. J. Meulman. Optimal scaling methods for multivariate categorical data. *SPSS white paper*, 2003.

[7] H. Niitsuma and T. Okada. Covariance and pca for categorical variables. 2005.

[8] M. Ruiz, J. Colomer, and M. J. Monitoring a sequencing batch reactor for the treatment of wastewater by a combination of multivariate statistical process control and classification technique. *Frontiers in Statistical Quality Control.*, 2006.

[9] J. Yuan, L. Qu, W. Zhang, and L. Li. Case-based reasoning combined with information entropy and principal component analysis for short-term load forecasting. *2007 International Conference on Computational Intelligence and Security*, pages 446–450, 2008.