Outlier overcoming using re-sampling techniques

Bruno Baruque[1], Bogdan Gabrys[2], Emilio Corchado[1], Álvaro Herrero[1], Jordi Rovira[3], Javier Gonzalez[3]

[1]Department of Civil Engineering, University of Burgos, Spain.
bbaruque@ubu.es, escorchado@ubu.es, ahcosio@ubu.es
[2]Computational Intelligence Research Group, Bournemouth University, United Kingdom.
bgabrys@bournemouth.ac.uk
[3]Department of Biotechonolgy and Food Science
jrovira@ubu.es

Abstract. Machine learning has extensively and successfully used statistical re-sampling techniques for generation of classifier and predictor ensembles. It has been frequently shown that combining so called unstable predictors has a stabilizing effect on and improves the performance of the prediction system generated in this way. In this paper we use the re-sampling techniques in the context of Principal Component Analysis (PCA). We show that the proposed PCA ensembles exhibit a much more robust behaviour in the presence of outliers which can seriously affect the performance of an individual PCA algorithm. The performance and characteristics of the proposed approaches are illustrated on a number of experimental studies where an individual PCA is compared to the introduced PCA ensemble.

1. Introduction

Projectionist methods are those based on the identification of "interesting" directions in terms of any one specific index or projection. Such indexes or projections are, for example, based on the identification of directions that account for the largest variance of a data set as in the Principal Component Analysis (PCA) method [1], [2]. Having identified the interesting projections, the data is then projected onto a lower dimensional subspace in which it is possible to examine its structure visually, which normally involves plotting the projection in two or three dimensions. The remaining dimensions are discarded as they are mainly related to a very small percentage of the information or the data set structure. In that way, the structure identified through a multivariable data set may be easily analyzed with the naked eye. This visual analysis may be distorted by the presence of outliers [3], [4]. Outliers are observations that lie an abnormal distance from other values in a set of data. In a sense, this definition leaves it up to the analyst (or a consensus process) to decide what will be considered abnormal. The presence of outliers can be caused by a number of different reasons and usually indicates faulty data, erroneous procedures, or areas where a certain theory might not be valid. In this study we analyse the use of statistical re-sampling theory [7], [9], [10], [12] in generation of PCA ensembles as a way of reducing or removing the influence of outliers on the generated principal components as well as identifying outliers which in themselves could be very interesting for the data analyst. The ideas explored in this paper are similar to those that have been employed in

generation of multiple classifier systems (classifier ensembles) [7]-[13] where the so called unstable classifiers (i.e. classifiers like decision trees or some neuro-fuzzy classifiers, the performance of which can be significantly affected by the presence of outliers) have been stabilized through the use of classifier ensembles. It has been frequently observed that PCA is also very sensitive to the outliers and the principal directions found can be significantly affected by their presence which in turn can lead to much more difficult analysis of the projected data or wrong conclusions.

The proposed approach is based on voting and averaging with the principal directions selected from the multiple PCA runs on sub-samples of the data set. Firstly the most frequently occurring principal directions are identified and as they can be somewhat different a further stabilizing effect is achieved through the averaging of the relevant eigenvectors. The hypothesis related to the presence or absence of harmful significant outliers is tested through the analysis of the consistency of the generated principal directions and the relative spread of the percentages of the variance explained. The significant shift in the directions of the principal components and large variation of the explained variance by different principal components obtained from different subsets of the original data set is used as indicators of the presence of the possible outliers.

The remaining parts of this paper are organised as follows. Basic PCA algorithm is summarised in section 2. Statistical re-sampling techniques and PCA ensembles are discussed in section 3. This is followed by the experimental analysis and results in section 4. And finally, conclusions and future work are described in section 5.

2. Analysis based on the variance

PCA originated in work by Pearson (1901) [1], and independently by Hotelling (1933) [2] to describe the variation in a set of multivariate data in terms of a set of uncorrelated variables each of which is a linear combination of the original variables. Its goal is to derive new variables, in decreasing order of importance, that are linear combinations of the original variables and are uncorrelated with each other. PCA can be implemented by means of some connectionist models [5], [6].

The disadvantage of this technique, both employing statistical or connectionist models is that this process is accomplished in a global way. This means that every data point that is situated far from the majority of the other cases belonging to the dataset can influence the final result, as it introduces a high variance compared with the rest, although it could be very small in number and could be considered as anecdotic or dispensable case. Almost in every mid-size non-artificial dataset a number of these outlier cases appear, distorting its variance and hence hindering its analysis.

3. Ensemble creation based on unsupervised learning

The technique utilised in this study to resist or detect the presence of outliers in a multidimensional dataset, is based on statistical re-sampling theory. One of the most widely known approaches utilizing statistical re-sampling techniques introduced by Breiman [7] is called "bootstrap aggregation" or "bagging".

In our case, the idea is to employ the bagging technique [7], [9] in combination with the PCA analysis in order to have more than one independent analysis performed over the same dataset. It is expected that, if any significant perturbation of the statistical characteristics of the dataset is produced only by a few of its components it

will be more evident in analysis of some data subsets than in others. Firstly, it is necessary to obtain different subsets of the dataset. This is achieved by randomly selecting several cases from the dataset and considering them as if they were a complete dataset. This process simulates the obtaining of several replications of the dataset we are working with. By doing this operation *n* times, *n* different datasets will be available, although they are really subsets of the main dataset. The next step consists of performing an individual PCA analysis on each one of the *n* subsets obtained by re-sampling the original one (Re-sampling PCA or Re-PCA). If the whole dataset does not include elements that alter drastically its statistical properties (i.e. in this case, its second statistical moment: the variance), the set of results obtained on the analysis of different subsets should be similar within a small margin. On the other hand, if few cases that alter these statistical properties are included in the main dataset, it is expected to generate different results in terms of directions of the principal components obtained. While re-sampling the data it is easy to imagine that one of those infrequent outlier data points can be included in a minority of the subsets, but will not be present in a majority of the other subsets. It can also be intuitively expected that the PCA performed on subsets containing outliers will be more influenced by the outliers if the ratio of the outliers to the number of other data points is high.

It is stated in [10] that bagging is especially recommended when applied to unstable algorithms or learning methods. As PCA can be considered as such an unstable algorithm an application of bagging for stabilizing of PCA in presence of outliers is one of the main premises of this investigation.

The description of the Re-PCA model proposed in this work can be summarized in the following two major steps:

I. Re-sampling and Principal Components Calculation. In this step first *n* subsets of the original data set are generated by re-sampling without replacement. This is followed by application of the standard PCA to each of the subsets. For further analysis the set of eigenvectors representing the directions of the first 3 principal directions and the percentages of variance explained by each of these principal components are recorded.

II. Voting and averaging. To perform voting and averaging of directions in order to obtain the final principal components the following steps are performed. A) For each of *n* subsets of eigenvectors we first identify the similar directions by performing pair wise similarity test by calculating the scalar product between the eigenvectors; B) All the vectors with their respective scalar products below certain threshold are then clustered together; C) The cluster with the largest number of the eigenvectors is selected and the sum of only these eigenvectors is calculated giving the final averaged direction for a respective principal component.

4. Application and Results

A real life dataset has been used to test the performance of the technique proposed on this work. The dataset used consists on measures obtained from several brands of seven types of Spanish ham, available in the Spanish market. The types of ham selected for this objective (with their codification in this experiment between brackets) are the following: the fat of hams cured for the same period of time (JCTC), cured ham of superior quality (JCCS), cured ham cured under a traditional speciality

guaranteed called TSG (JCTSG), cured ham with protected designation of origin (PDO) from Teruel, a region in Spain (JCTE), cured ham cured during seven months (JC7C), cured ham with different defects, off-flavours or taints, (JCCA and JCBO) and standard cured ham (JCNO). The commercial brands from where the samples were extracted are not taken into account in this study.

Several samples of different parts of each type of ham where cut and measures where made over each of these samples, by an electronic nose α FOX 4000 (Alfa MOS, Toulouse, France) with a sensor array of 18 metal oxide sensors. Our final dataset consists in a total of 176 samples of ham, each of them composed of 18 different variables measured over it.

In order to test the efficiency of the Re-PCA in dealing with outlier measures, 4 outlier samples have been added to three of the experiments detailed below.

Experiment 1: Performing a single PCA analysis over the original data and representing the data into the axes obtained in the analysis gives a quite interesting result. In that projection it can clearly be seen the samples of highest quality types of ham (JC7C, JCCS and partially JCTE) in the right of the image, the altered parts completely concentrated in the center of the image and the standard and low quality types of ham situated in the left of the image (JCNO, JCTC). Even the particular fact that some of the samples of high quality ham (JCCS and JCTE) where more rancid than others, is reflected in the image as it can be seen several of them mixed in the group of the standard quality types. Attaching an identifier to each of the points it can be verified that those samples are precisely the more rancid ones.

This information can be clearly seen in the projection of the data over the first and second Principal Components as well as in the first and third Components and showed in Figs. 1 and 2.

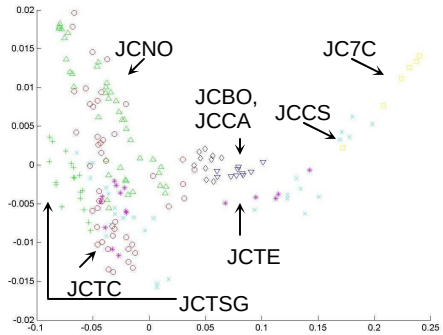

Fig. 1. Projection of dataset (without outliers) over the 1st and 2nd Principal Components extracted form a Simple PCA analysis.

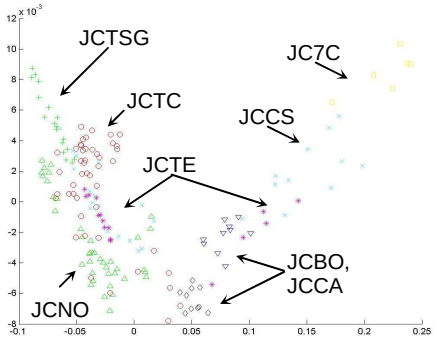

Fig. 2. Projection of dataset (without outliers) over the 1st and 3rd Principal Components extracted form a Simple PCA analysis.

Table 1. Percentage of information captured by each of the principal components in the first part of the experiment (without outliers) including.

Principal component	*Percentage of information captured*
First	86.58 %
Second	8.74 %
Third	4.66 %

Table 1 shows the percentage of information captured by each one of the first three principal components. This information can be compared with the same information in following experiments, where the inclusion of few outliers makes vary the whole analysis.

Performing a Re-PCA analysis over this dataset does not reveal further noticeable information. As no outlier points are included in the dataset all the independent tests composing the Re-PCA give almost the same result, so the average of their results coincides almost completely with the result of a simple PCA.

Experiment 2: In order to observe the effect that the inclusion of outliers has in the PCA analysis, four outlier measures have been added to the original dataset for this experiment. Performing a simple PCA analysis and projecting the data in the two axes determined by the principal components, in exactly the same way as in the previous experiments gives the results showed in Figs 3 and 4.

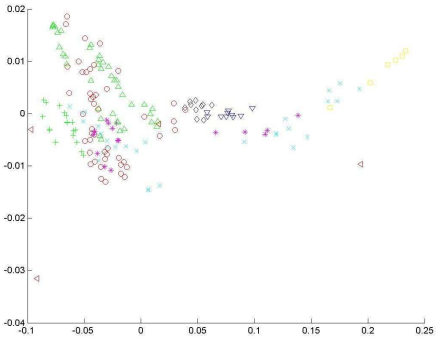

Fig. 3. Projection of dataset (including outliers) over the 1st and 2nd Principal Components obtained from a Simple PCA analysis.

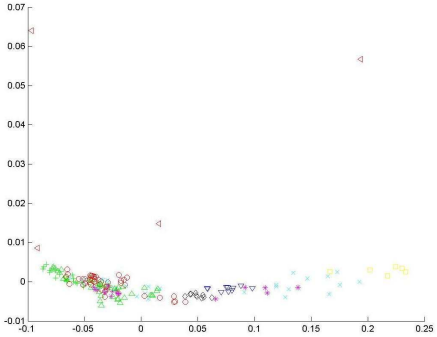

Fig. 4. Projection of dataset (including outliers) over the 1st and 2nd Principal Components obtained from a Simple PCA analysis.

As it can be seen, although the first and second components are affected by the inclusion of outliers, the groups described in experiment 1 can still be distinguished. On the contrary, the third principal component is completely different, as the projection over it and the first one no longer shows those groups.

Table 2. Percentage of information captured by each of the principal components in the second part of the experiment (including outliers).

Principal component	*Percentage of information captured*
First	83.78 %
Second	8.4 %
Third	7.8 %

Regarding to Table 2 it can be observed that the percentage of information captured by the first principal component is lower in this case, while that percentage rises for the third PC. In plain language, this means that the first component is no longer able to capture as much information as before (as the variance of the dataset has been altered) and is the third component the "one in charge" of dealing with the information that was before captured by the first one.

This situation verifies empirically, that the presence of very few outliers can alter the results of the PCA analysis in a significant way.

Experiment 3: In this experiment the Re-PCA has been applied to the dataset including outliers. In the first part of the experiment, 80 samples from the whole dataset have been randomly selected (without replacement) for each PCA analysis. Ten different analyses have been performed and averaged to obtain the results showed in Figs 5 and 6.

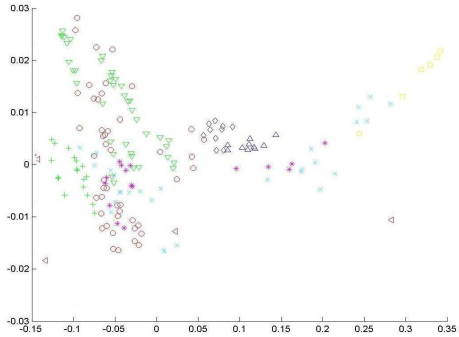

Fig. 5. Projection of dataset (including outliers) over the 1st and 2nd Principal Components obtained from a Re-PCA analysis using 80 samples.

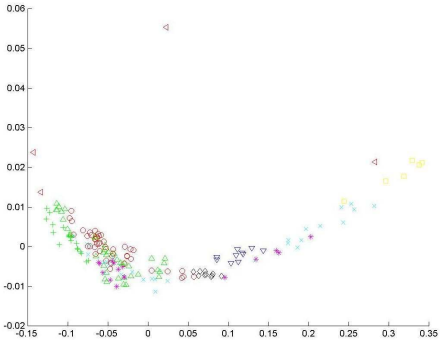

Fig. 6. Projection of dataset (including outliers) over the 1st and 2nd Principal Components obtained from a Re-PCA analysis using 80 samples.

Inspecting Fig. 5 it is patently obvious that the first and second principal components are able to display almost the same information they did in experiment 1, that is the experiment done without outliers in the dataset. Even inspecting Fig. 6 (corresponding to first and third principal components) a slightly more clear structure can be seen.

This could be interpreted as a very good improvement over the experiment 2, but checking Table 3 it can be seen that the process is still too unstable, as the single PCA analyses that compose the Re-PCA perform even better than the single PCA in some occasions (87% of information for the 1st PC) but very poorly in other (79.9% of information for the 1st PC). This can be explained by the fact that less than the half of dataset is used for each of the analyses.

Table 3. Percentage of information captured by each of the principal components in the experiment (with outliers) including the maximum and minimum percentage of information (variance) from the analysed 10 subsets.

Principal component	*Percentage of information captured*	
	Max	Min
First	87.1 %	79.9 %
Second	11.03 %	8.51 %
Third	9.2 %	4.35 %

Experiment 4: In this occasion we repeat the procedure followed in the experiment 3, but this time we include 120 samples randomly selected from the whole dataset in each of the ten analyses performed over the Re-PCA process. The results obtained after projecting the dataset over the three main principal components found in this experiment are showed in Figs. 7and 8.

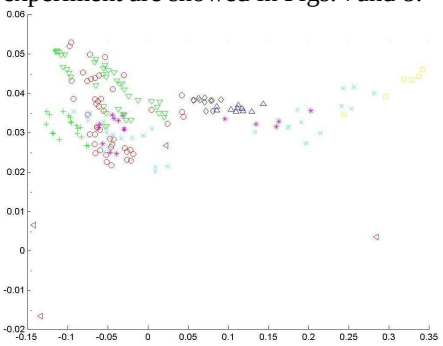

Fig. 7. Projection of dataset (including outliers) over the 1st and 2nd Principal Components obtained from a Re-PCA analysis using 120 samples.

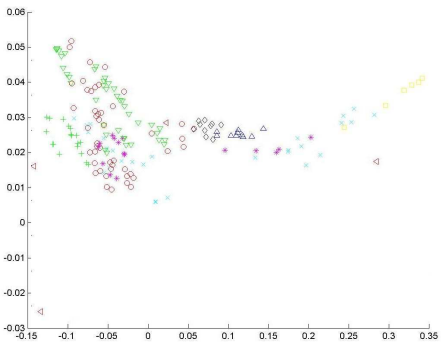

Fig. 8. Projection of dataset (including outliers) over the 1st and 2nd Principal Components obtained from a Re-PCA analysis using 120 samples.

Table 4. Percentage of information captured by each of the principal components in the experiment (with outliers) including the maximum and minimum percentage of information (variance) from the analysed 10 subsets.

Principal component	*Percentage of information captured*	
	Max	Min
First	86.52 %	82.5 %
Second	9.473 %	8.06 %
Third	7.9 %	4.60 %

The results in this case are similar to the ones obtained in experiment 1, having in account that now the outliers have to be represented too (so images get a bit distorted). In both the axes formed by the first and second and by the first and third principal components the group information described in experiment 1 are clearly distinguishable. In Table 4 it is displayed the difference of the percentage of information captured by each one of the principal components. As it can be verified, the difference is much lower know, showing that in this experiments the coincidence of the directions found by each of the ten tests is much higher. This means the directions of maximum variance found in this experiment are quite more consistent than the ones that were found in experiment 3.

5. Conclusions

In this study we have applied a simple projectionist model (PCA) as a powerful technique to identify the existence of outliers in a dataset by using statistical re-sampling techniques in combination with voting and averaging.

We have observed that in absence of outliers, the re-sampling technique gives very similar Principal Components (PCs) as a result of a number of independent runs. However, when outliers are present in the dataset the situation is different. The

smaller the number of points included in a subset, the bigger the difference in the response of the variance obtained due to a greater influence of the outliers in the subset. A higher ratio of the outliers to the normal points significantly affects the directions of maximum variance of the dataset and thus the directions of the principal components. The proposed Re-PCA algorithm has shown a very robust behaviour in presence of outliers consistently finding the right principal directions while the individual PCA was significantly affected. The use of re-sampling in the context of PCA has had an additional benefit by allowing analysing the variance and its differences from different runs which in itself proved to be a very useful tool for detection of the presence of outliers.

Acknowledgments

This research has been supported by the MCyT project TIN2004-07033 and the project BU008B05 of the JCyL.

References

1. Pearson, K. On lines and planes of closest fit to systems of points in space. Philosophical Magazine 2:559-572. (1901).
2. Hotelling, H. Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, 24:417-441,498-520. (1933).
3. Cook, R. D. Detection of influential observations in linear regression. Technometrics 19, 15-18. (1977).
4. Dixon, W. J. Analysis of extreme values, Ann. Math. Stat., 21, 488-506. (1950)
5. Oja, E. Neural networks, principal components and subspaces. International Journal of Neural Systems 1(1):61-68. (1989).
6. Sanger, D. Contribution analysis: A technique for assigning responsibilities to hidden units in connectionist networks. Connection Science, 1:115--138. (1989).
7. Breiman, L. Bagging predictors. Machine Learning, 24:123–140. (1996).
8. Schapire, R.E; Freud, Y; Bartlett, P. and Lee, W.S. Boosting the margin: a new explanation for the effectiveness of voting methods. The Annals of Statistics, 26(5):1651–1686, 1998.
9. Gabrys, B. Combining neuro-fuzzy classifiers for improved generalisation and reliability. In Proceedings the Int. Joint Conference on Neural Networks (IJCNN'2002) a part of the WCCI'2002 Congress, pages 2410–2415, Honolulu, USA, 2002.
10. Kuncheva, L, Combining Pattern Classifiers: Methods and Algorithms. Wiley-Interscience, 2004.
11. Ruta, D. and B.Gabrys, Classifier Selection for Majority Voting, Special issue of the journal of information fusion on Diversity in Multiple Classifier Systems, vol. 6, issue 1, pp. 63-81, 1 March 2005.
12. Gabrys, B., Learning Hybrid Neuro-Fuzzy Classifier Models From Data: To Combine or not to Combine?, Fuzzy Sets and Systems, vol. 147, pp. 39-56, 2004.
13. Ruta, D. and B. Gabrys, A Theoretical Analysis of the Limits of Majority Voting Errors for Multiple Classifier Systems, Pattern Analysis and Applications, vol. 5, pp. 333-350, 2002.