Outlier Resistant PCA Ensembles

Bogdan Gabrys[1], Bruno Baruque[2], and Emilio Corchado[2]

[1] Computational Intelligence Research Group, Bournemouth University, United Kingdom
bgabrys@bournemouth.ac.uk
[2] Department of Civil Engeneering, University of Burgos, Spain
escorchado@ubu.es, bbaruque@ubu.es

Abstract. Statistical re-sampling techniques have been used extensively and successfully in the machine learning approaches 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

B. Gabrys, R.J. Howlett, and L.C. Jain (Eds.): KES 2006, Part III, LNAI 4253, pp. 432 – 440, 2006.

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 are 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 Principal Component Analysis

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 Statistical Re-sampling Techniques and PCA Ensembles

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 Experimental Analysis

The artificial data set used in this series of experiments is made of one cloud of points and several points spread far above the main cloud of points which will be considered as *outliers*. The main cloud is an elongated cluster which moves within the axis delimited by the line defined by the points [1,1,0], [2,1.6,0], [3,2.2,0] and [4,2.8,0]. By employing this dataset we expect to obtain 3 clear principal components, as the

variance of each direction is different in comparison to the other two. The outlier points are spread over the same axis but displaced 5 units above in the vertical axis. There are 118 points in the main cluster and 8 outliers.

In order to test various characteristics of the proposed Re-PCA algorithm with regard to different proportions of outliers to the other data points and various sizes of the data sets, the experiments with 30, 50, 70 and 100 randomly selected points have been carried out. The experiments have been repeated 10 times for each one of those cases and the comparative analysis is presented below.

Dataset 1. In this set of experiments 10 subsets of 30 points randomly selected from the entire dataset (without replacement) are generated and PCA performed on each of them. Firstly, the method described above is applied to the dataset formed only by the main elongated data cluster (i.e. without the outliers). The results of PCA obtained from those 10 subsets are represented in Fig. 1.

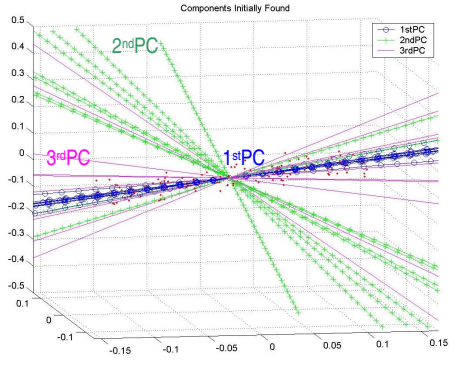

Fig. 1. Projections of Re-PCA using 30 points (excluding outliers)

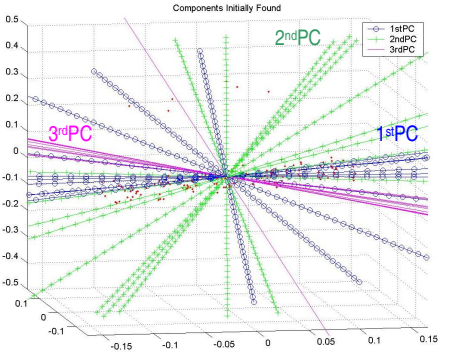

Fig. 2. Projections of Re-PCA using 30 points (including outliers)

Examining Fig. 1 it is easy to observe that the Re-PCA method has found almost the same direction for the first principal component, as it was expected. For the direction of the second and third principal components, they are clearly more dissimilar in the different tests, but still they all follow a consistent direction, except in one case.

The percentage of information (in form of the explained variance) that is represented by each one of the principal components is detailed in Table 1.

Table 1. Percentage of information captured by each of the principal components in the first part of the experiment (without outliers) including the maximum and minimum percentage of information (variance) from the analysed 10 subsets

Principal component	*Percentage of information captured*	
	Max	Min
First	72 %	68 %
Second	18 %	14 %
Third	14 %	11 %

Fig. 2 represents the results obtained performing exactly the same experiment but including now the 8 outliers in the sampled dataset. As it can be seen, the distribution of the directions corresponding to the principal components, produced when outliers are taken into account, are much more spread than in Fig. 1 (data without outliers). This means that the direction found in each case is rather dissimilar to the other corresponding ones. We can even consider that in 3 cases out of 10 (30 % of the cases), the method has found opposed directions for the first and second principal components. Looking at Fig 2, it can be seen that the first principal component appears in an almost horizontal direction on 7 occasions, while it appears in the diagonal that goes from the bottom right corner to the upper left one on the other 3 occasions. This three deviated directions will not be taken in account in the average calculation stage as the majority cluster consists of the 7 cases where the 1st principal component appears in the horizontal direction. The "percentage of information" that is represented by each of the principal components in this case is detailed in Table 2.

Table 2. Percentage of information captured by each one of the principal components in the second part of the experiment (including outliers) including the maximum and minimum percentage of information (variance) from the analysed 10 subsets

Principal component	*Percentage of information captured*	
	Max	Min
First	69 %	49 %
Second	41 %	17 %
Third	13 %	8 %

The results presented in Table 2 are quite different from the results obtained without including the outliers. Comparing both tables (Table 1 and Table 2), the influence of the presence or absence of outliers in a dataset in terms of the direction of the largest variance and relative difference between the Max and Min values for the principal components can be clearly seen.

The amount of information associated with the first principal component is different depending on whether outliers are included or not into the analysed data. The presence of these outliers makes the amount of information detected by the first component (Table 2) to be inferior to the situation without outliers (Table 1). In this case, due to the shape of the artificial dataset used and due to the fact that the outliers are situated in the direction associated with the second principal component, the amount of information represented by this second component (Table 2) is a lot higher than in the case that does not include outliers (Table1). As it was expected, the inclusion of the outliers brings a great instability to the dataset, making different individual PCAs behave in an inconsistent way and resulting in very different results where really the analysis is made over subsets of the same dataset. The use of PCA ensemble in cases like this is of particular use as the 70% of the cases where "the true" principal component is found represents the majority which is selected and then further stability is added through averaging of the eigenvectors from these 70% of majority similar principal directions. The final averaged directions are shown in Fig 3

Dataset 2. In this case the same experiment is performed for 50 points randomly selected from the entire dataset (without replacement). This is also performed 10 times.

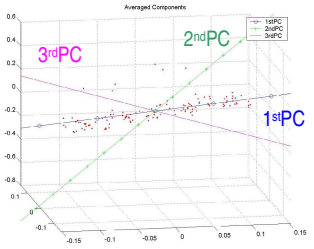

Fig. 3. Resulting average for each of the principal components by voting between the directions shown in Fig 2 (excluding the 30% of the directions strongly influenced by the outliers)

We have noted that increasing the number of samples included in each of the subsets analysed by Re-PCA, brings stability to the performance of the experiments. The “fans” formed (for the case of 50 points data set) by the directions corresponding to the three principal components of the ten tests are far closer than the ones obtained in an analogous experiment including only 30 samples. The "percentage of information" that is represented by each of the principal components is shown in Table 3.

Table 3. Percentage of information captured by each of the principal components (selecting 50 points but excluding outliers) including the maximum and minimum percentage of information from the analysed 10 subsets

Principal component	*Percentage of information captured*	
	Max	Min
First	72 %	68 %
Second	16 %	14 %
Third	14 %	12 %

Performing the same operations but including now the 8 outliers, we have obtained the following. Although including the additional 20 data points has had a stabilizing effect on the individual PCAs, there are still two occasions out of ten (20% of the cases) where the first and second principal components appear in an almost perpendicular direction to the other eight occasions, indicating some instability which may be due to the presence of outliers in the dataset. The second principal component is always very unstable because all the outliers are situated in its direction. Table 4 shows the “percentage of information” for each of the principal components.

Table 4. Percentage of information captured by each of the principal components (selecting 50 points and including outliers) including the maximum and minimum percentage of information from the analysed 10 subsets

Principal component	*Percentage of information captured*	
	Max	Min
First	70 %	44 %
Second	44 %	15 %
Third	13 %	9 %

Datasets 3 and 4. To generate datasets 3 and 4, again 10 subsets of 70 and 100 points respectively have been used to test the stability of the PCA analysis performed over them. The results obtained for data set 4 are shown in Fig. 4 and Fig. 5.

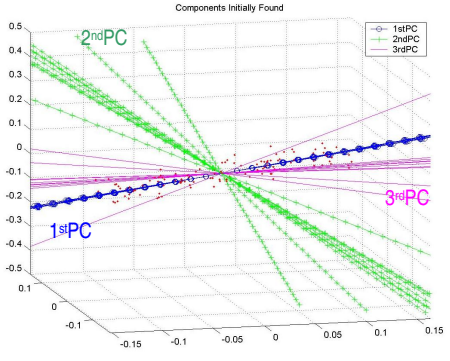

Fig. 4. Projections of Re-PCA using 100 points (excluding outliers)

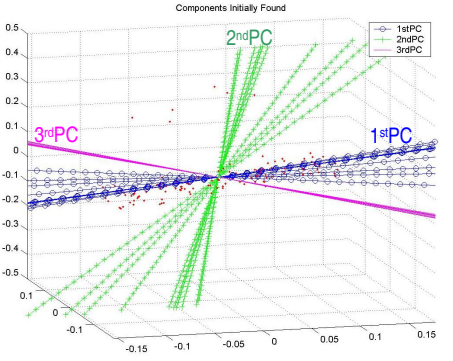

Fig. 5. Projections of Re-PCA using 100 points (including outliers)

As it can be seen in all the above experiments, the more samples are included into the analysis, the more stable behaviour of the individual PCA. Comparing Fig. 2, Fig. 5 (and the results obtained for data sets 2 and 3) gives a visual prove of that, as the directions found using 100 points are slightly more consistent than when using only 30, 50 or 70 points. It can also be seen (Fig. 4 and Fig. 5) that including outliers in the analysed dataset brings a substantial degree of instability, giving as a result more spread "fans" (less consistent results) or even completely different directions for its principal components.

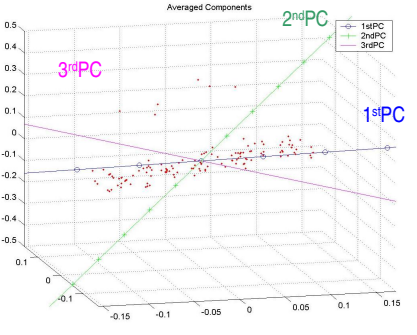

Fig. 6. Resulting average for each of the principal components by voting and averaging of the directions shown in Fig5

Calculating the average directions (Fig. 6) as explained above, we can obtain approximately the same main directions for the three principal components, as when

we have only used 30 points for calculations. This can be considered as an empirical proof of the robustness of the proposed Re-PCA method.

5 Conclusions and Future Work

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. The first principal component is the same almost in 100% of the cases. The second principal component could be different in a larger percentage of cases, due to our particular artificial dataset. 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.

Future work will also investigate this and other neural and statistical methods, based on higher order statistics, on a larger range of data sets which have been impossible to include in this paper due to the space limitations.

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.