Analysis of the dependence between PCA and classifiers

Sijie Guo and Andrea Visentin

Insight
Centre for Data Analytics

Introduction

Principal component analysis (PCA) is one of the most widely used data analysis techniques. One of its main application is dimensionality reduction in classification. A great number of different PCA algorithms has been developed to examine and improve the robustness, smoothness and sparsity[1]. It is not clear which of them should be use in preprocessing of classification tasks.

The most common benchmark in PCA algorithms literature is based on classification of real world dataset.This test is based on the idea that a better PCA solution will preserve more useful information and get rid of noise, outliers and correlation. So a classifier trained on this dataset should perform better than one trained on a reduced dataset with less useful information.

In paper related to PCA research, benchmark performance is tested on only one classifier. In this project, we investigate if the differences of performance in preprocessing data reduction techniques of classification tasks are classifier dependent. There are no work in literature that investigate this dependence.

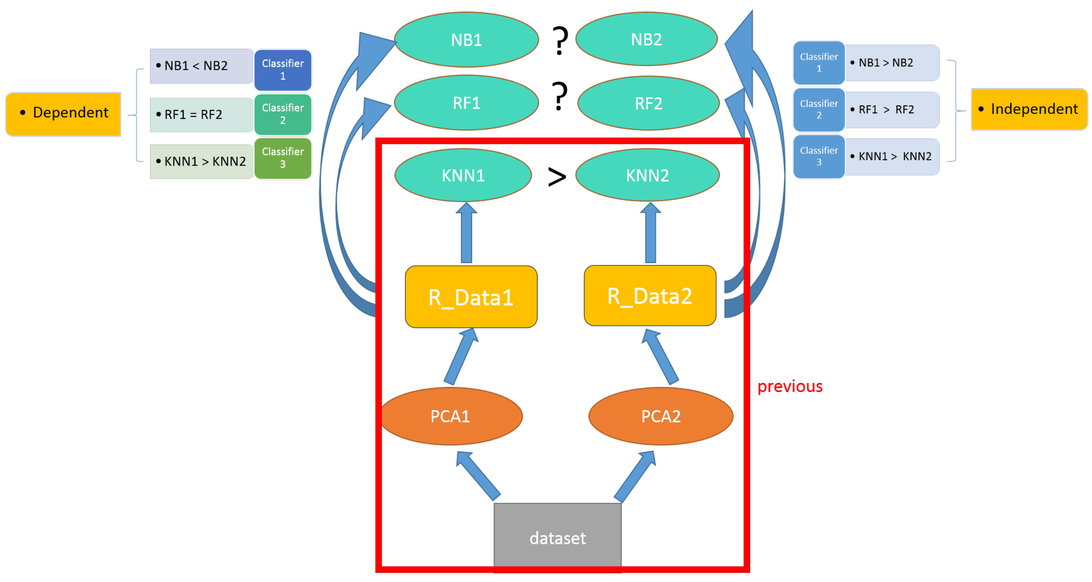

Figure: Flow chart of comparison on different classifiers

Method

We apply pairwise comparison of PCA on several classifiers to verify if reduction results vary among classifiers.

To achieve the goal, our procedure contains three main components:

- Dataset preparation
- PCA comparison
- Compare PCA on different Classifiers

Dataset preparation

All the data used to analyse in this project are from UCI Machine Learning Repository online dataset in various range of fields, such as: biology, health and medicine, automotive and weather. The datasets are heterogeneous on number of attributes and instances.

PCA comparison

We load algorithms from library pcaL1 and pcaPP. We will compare different PCA based on benchmark performance with the different dataset on a fixed classifier.

Compare PCA on different Classifiers

We apply the same comparison of PCA on different classifiers and save the results. Then we analyse the results to draw the conclusion.

Results

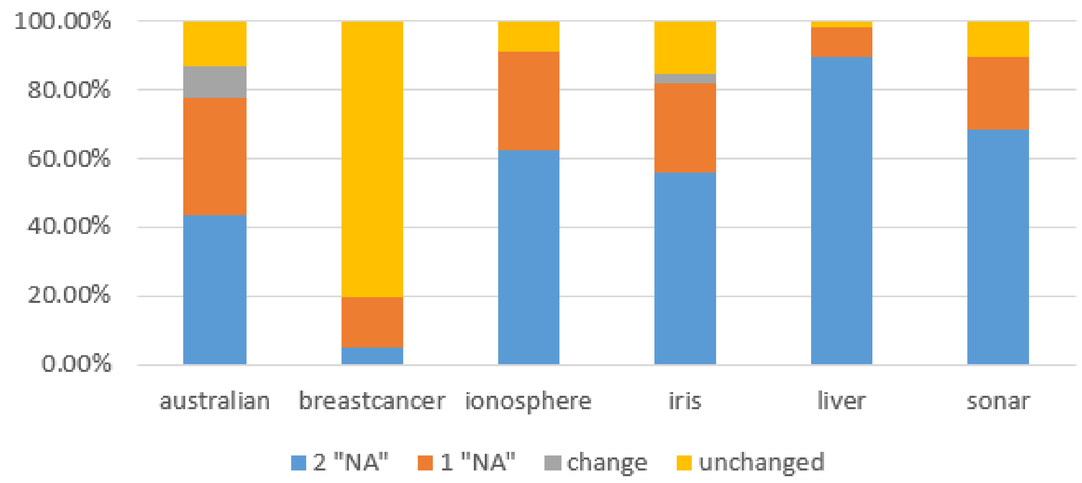

Figure: Flow chart of comparison on different datasets

This plot shows the results of the comparison divided per dataset. Each column is divided in: *NA* if we can not reject the null hypothesis in one of the two test, *unchanged* if the comparison is equivalent in both the test and *changed* if the comparison differs. Most of our comparison lead to an "NA" result, with exception for the "breastcancer" dataset. We think this is due to a high fluctuation on the classification results. We observe *changes* in the comparison only in "australian" and "iris" datasets.

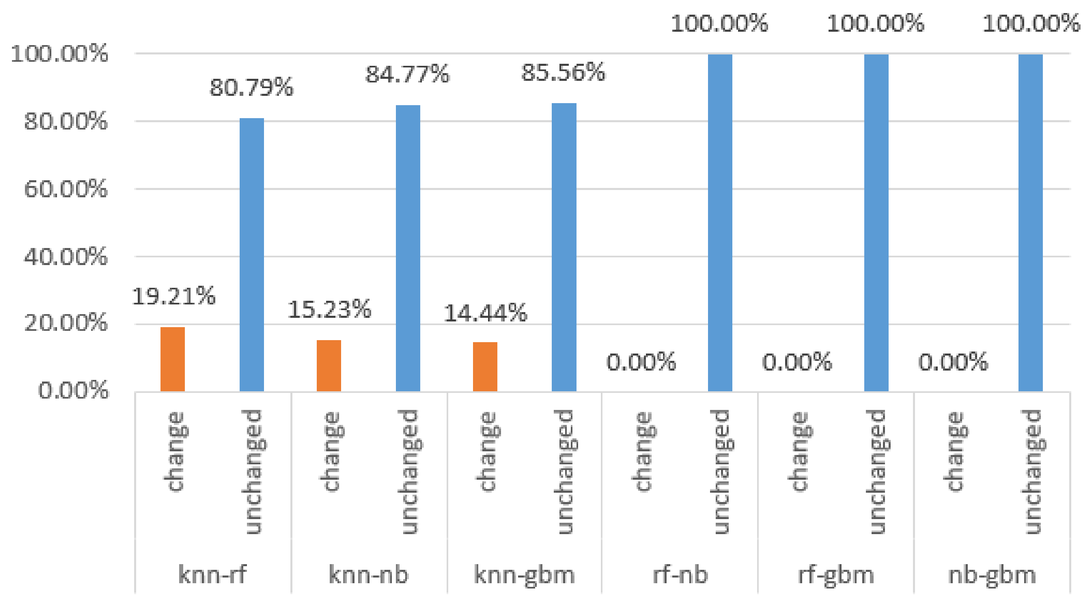

Figure: Plot of comparison between pairwise classifiers

This plot shows the percentage of "changed" and "unchanged" comparison averaged over every pair of classifiers. We can observe that observing a variation in the classifiers performances is rare. The changes only appear between knn and other classifiers.

These changes always are between one classifier being better than the other to be equal. We never observe any inversion of the performances.

Conclusion

From the experiments so far, we can get the conclusion that generally PCA are classifiers independent on most datasets. The fact that we observed no inversion on the performances strength this hypothesis. More data are required to draw a final conclusion.

References

Reris, R., and Brooks, J. P.
Principal component analysis and optimization: A tutorial.

This work has been supported by a research grant by Science Foundation Ireland under grant number SFI/12/RC/2289