Cover letter

for the paper "Principal Component Analysis for Supervised Leaning: a Minimum Classification Error Approach".

Dear Editors,

Please find enclosed the manuscript entitled Principal Component Analysis for Supervised Leaning: a Minimum Classification Approach to be considered for publication in the JOURNAL OF INFORMATION AND DATA MANAGEMENT. This is a selected paper from KDMILE 2016. We believe that our paper is relevant because it presents a novel method to apply PCA using Bayes error rate.

As asked in the invitation letter, we response reviewers's comments bellow.

Best regards,
Tiago B. A. de Carvalho.

Response reviewers's comments

REVIEW 1: *"This paper presents the idea of using PCA for classification tasks, choosing only a subset of the Principal Components based not on the eigenvectors, as usually is done, but looking at the influence of each one on Bayes error rate. The presentation is clear and the experiments straightforward; results support the claim and indicates that*
the proposal deserves further attention."
-- This reviewer did not ask any modification.

REVIEW 2: *"This is a well written paper and the ideas are clearly presented. My main issue with the paper is the very small number of datasets. I don't think authors can draw such conclusion based on the results of only 2 datasets. At most, the results show potential and indicate that it might require a smaller number of features, but this has to be verified in a large test bed.*
Authors state that minimising the Bayes error may be more suitable than PCA. Can the authors elaborate on this?
What is the evidence that suggests that PCA is not the best strategy for classification?
Provide details of the decision tree algorithm used in the experiments."
-- We perform experiments using four more datasets. The computation of the Bayes error rate imposes some restrictions that limit the number of suitable datasets. That is one of the reasons why we prefer to add experiments with artificial datasets. Another important motivation for using synthetic data is the possibility of a deeper analysis by controlling dataset parameters.
-- It is known that PCA is not appropriate for classification because the directions of maximal variance are not the better class discrimination. This is discussed in the Introduction section, and the statement is supported by Bishop (2006), as cited in the text.
-- We added details about the decision tree algorithm: we used a "pruned Decision Tree with Gini's diversity index and a minimum of 10 nodes per leaf".

REVIEW 3: *"The paper presents a modified PCA based feature extraction method that selects features that minimizes the Bayes error rate instead of features that maximizes the variance.*
The authors presented the accuracy of the proposed method With four classifiers: Nearest Neighbor (NN), Naive Bayes, Decision Tree and Linear Discriminant and two data sets using a small number of features and the proposed method had a higher accuracy than traditional PCA.
From the results, it is not clear if the accuracy always converge when the number of features increases or not, since the number is not high (<20). It would be interested to observe this with

more data sets and with more features.
The paper does not present related work, it seems that this idea could have been already proposed.
Another question is: what if instead of choosing between the two methods (the proposed one that selects features that minimizes the Bayes error rate and the original one that selects features that maximizes the variance), both of them were combined with a weight to be learned or given as a parameter? Would it be possible? Would the results be better, the same?"
-- We answered for review 2 that we added more datasets.
-- We can not increase the number of features because each dataset has a fixed number of features. But it is possible using datasets with more features. However, it is not easy to find dataset the met the restrictions imposed by the proposed technique. In a future extension, we will remove some of the restrictions.
-- We presented related work, the only two papers that propose supervised PCA are described in the Introduction section: Barshan et al. (2011) and Bair et al. (2006). We are very familiar with PCA and have confidence that there are not other supervised PCA proposal. Nonlinear supervised dimensionality reduction methods, such as Isomap (Balasubramanian and Schwartz, 2002), are not discussed because they belong to a different domain: nonlinear methods. Other supervised linear feature extraction methods, such LDA (linear discriminant analysis) (Martinez and Kak, 2001), could be commented but we are not sure if this would turn reader's attention for a direction that is not the focus of the paper: better usage of PCA in classification.

References

M. Balasubramanian, E. L. Schwartz, The Isomap Algorithm and Topological Stability. Science 4 January 2002: Vol. 295 no. 5552 p. 7

A. M. Martinez and A. C. Kak, "PCA versus LDA," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 2, pp. 228-233, Feb 2001.