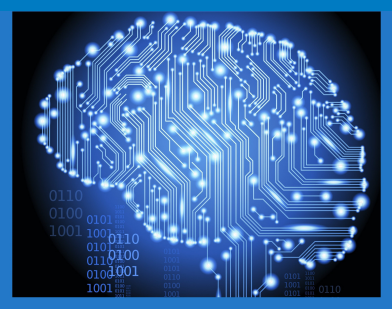
Department of Machine Learning

Institute of Computer Science, CAS Prague



Robust classification for high-dimensional data

Jan Kalina

Motivation

Classification analysis has the aim to learn a decision rule over a training data set, which is able to automatically assign new data to one of K given groups. If the number of observed variables p exceeds (perhaps largely) the number of observations n, then various approaches to regularized linear discriminant analysis (RLDA) represent a suitable methodology for the classification task. Nevertheless, available methods are vulnerable to the presence of outlying values (outliers) in the data. So far, only a few attempts have been made to combine robustness with regularization for the analysis of high-dimensional multivariate data [2,4].

Regularized linear discriminant analysis

Various versions of RLDA are based on regularized estimates of the covariance matrix Σ (common across groups), which is guaranteed to be regular and positive definite even for $n \ll p$. Currently, there exist already dozens of fast algorithms for its computation, while some of them possess a reasonable numerical stability.

The need for robustness

RLDA (like various other data mining procedures) is sensitive to outliers in the data and its robust counterparts are highly desirable. We must namely point out that the common belief that

$$\text{Regularization} = \text{Robustness}$$

is not valid, if we speak about robustness in terms of the breakdown point, which can be characterized as a global measure of robustness of an estimator against severe outliers in the data.

Aim of the work

The aim is to combine principles of robust statistics with regularization to obtain classification methods for high-dimensional data, which are highly robust in terms of the breakdown point. One aim is to investigate the novel robust RLDA of [4]. It is based on the Minimum Weighted Covariance Determinant Estimator (MWCD) of Σ with implicitly given weights assigned to individual observations. It also allows to obtain a sparse solution (i.e. to perform variable selection). Computational aspects and classification performance will be studied for various weighting schemes, especially those ensuring a high robustness (for data contaminated by severe outliers) together with high efficiency (for non-contaminated data).

Further, an original idea seems to consider an estimator of Σ yielding the Minimum Product of s Largest Eigenvalues (therefore abbreviated as MPsLE) for a given (small) s. Within the classification method of [4], the MWCD may be replaced by the MPsLE to obtain a novel robust regularized classifier with much lower computational demands. Properties of the new method will be studied.

Outline of the work

1. Basics of regularized linear discriminant analysis. Efficient algorithms based on tools of numerical linear algebra.
2. Aims and principles of robust statistical estimation in the multivariate model.
3. Minimum weighted covariance determinant (MWCD) estimator. Implementation.
4. Proposal of the MPsLE estimator (Minimum Product of s Largest Eigenvalues), which will be an alternative form of the MWCD based on the largest eigenvalues.
5. Proposal of a novel classification method based on MPsLE.
6. Performance in classification tasks on real and simulated data sets.
7. Various weighting schemes for the MPsLE estimator. Comparison of their performance in the classification task.
8. Conclusions. Advantages and disadvantages of the robust regularized classification based on the MPsLE estimator.

Possible applications

All possible classification tasks of function approximation with high-dimensional data contaminated by outlying values, e.g. in engineering, medicine, bioinformatics, image analysis, chemometrics, or econometrics, while their fast computation and numerical stability are crucial.

References

1 Duintjer Tebbens J., Schlesinger P. (2007): Improving implementation of linear discriminant analysis for the high dimension/small sample size problem. Computational Statistics & Data Analysis 52, 423 – 437.

2 Gschwandtner M., Filzmoser P. (2013): Outlier detection in high dimension using regularization. Advances in Intelligent Systems and Computing 190, 237 – 244.

3 Hastie T., Tibshirani R., Wainwright M. (2015): Statistical learning with sparsity. The lasso and generalizations. CRC Press, Boca Raton.

4 Kalina J., Hlinka J. (2017): Implicitly weighted robust classification applied to brain activity research. Communications in Computer and Information Science 690, 87 – 107.

5 Kalina J. (2014). Classification analysis methods for high-dimensional genetic data. Biocybernetics and Biomedical Engineering 34, 10 – 18.

6 Roelant E., van Aelst S., Willems G. (2009): The minimum weighted covariance determinant estimator. Metrika 70, 177 – 204.

7 Xu H., Mannor S. (2012): Robustness and generalization. Machine Learning 86, 391 – 423.

http://ml.cs.cas.cz kalina@cs.cas.cz