Robust Learning from Bites

A. Christmann[1]

[1] University of Dortmund, Department of Statistics, 44221 Dortmund, Germany

Keywords: Influence function, Kernel based methods, Large data sets, Robustness, Support vector machine.

Abstract

Data sets with millions of observations occur nowadays in different areas. An insurance company or a bank collects many variables to develop tariffs and scoring methods for credit risk management, respectively. Other examples are data mining projects and micro-arrays. For such data sets parametric assumptions are often violated, outliers are present, or some variables can only be measured in an imprecise manner. The application of robust statistical methods is important in such situations. However, some robust methods have the following drawbacks which are serious limitations for the application of robust methods. They are computer-intensive such that they can hardly be used for massive data sets, say for several millions of observations with many explanatory variables. Robust confidence intervals for the estimated parameters or robust confidence intervals for the predictions are often unknown. Some statistical software packages like S-PLUS or R contain state-of-the-art algorithms for robust statistical methods, but the implemented numerical algorithms often require that the whole data set fits into the memory of the computer.

The talk has two goals. Firstly, it is shown that some kernel based regression estimators based on the convex risk minimization principle have good robustness properties. Kernel based methods are flexible non-parametric methods and can be applied to large high-dimensional data sets. The influence function and the sensitivity curve are investigated. These results complement those given Christmann and Steinwart (2004) who considered such estimators for classification problems. Secondly, a simple but quite general method for robust estimation in the context of huge data sets is proposed. The goal of the proposal is to broaden the application of robust methods for data sets which are too large for currently available algorithms. The idea is to split the data set S by random into disjoint subsets. Then the robust method is applied to each subset, and the results are summarized in a robust manner. We call this robust learning from bites (RLB). The proposal yields robust predictions for the median together with distribution-free confidence intervals. The method is scalable to the memory of the computer and the computation can easily be distributed on several processors which helps to reduce the computation time substantially.

Among the very best robust algorithms for large data sets are FAST-LTS for linear regression and FAST-MCD for multivariate location and scatter problems, see Rousseeuw and Van Driessen (1999,2002). The main difference between these algorithms and RLB is the following. FAST-LTS and FAST-MCD also splits the data set into sub-samples, but gives the exact solution or a good approximation for the robust estimate for the whole data set. RLB aggregates the results based on robust estimates obtained from many disjoint sub-samples and offers distribution-free confidence intervals for the median of the predictions. Therefore, RLB and FAST-LTS/FAST-MCD have different goals.

References

L. Breiman (1999). Pasting bites together for prediction in large data sets. *Machine Learning*, 36, 85–103.

N. Chawla, L. Hall, K. Bowyer and W. Kegelmeyer (2004). Learning ensembles for bites: a scalable and accurate approach. *Journal of Machine Learning Reseach*, 5, 421–451.

A. Christmann (2005). Robust Learning from Bites. University of Dortmund, Department of Statistics, TR 07/05, SFB 475.

A. Christmann and I. Steinwart (2004). On robust properties of convex risk minimization methods for pattern recognition. *Journal of Machine Learning Research*, 5, 1007–1034.

A. Christmann and I. Steinwart (2005). Robustness properties of kernel based regression. University of Dortmund, Department of Statistics, TR 01/05, SFB 475, submitted.

S. Keerthi, K. Duan, S.K. Shevade and A.N. Poo (2002). A fast dual algorithm for kernel logistic regression. In: Proceedings of the 19^{th} International Conference on Machine Learning. pp. 299–306, Morgan Kaufmann Publishers Inc., San Francisco.

P. Rousseeuw and K. Van Driessen (2002). Computing lts regression for large data sets. *Estadistica*, 54, 163–190.

P. Rousseeuw and K. Van Driessen (2002). A Fast Algorithm for the Minimum Covariance Determinat Estimator. *Technometrics*, 41, 212–223.

B. Schölkopf and A. Smola (2002). *Learning with Kernels. Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, Massachusetts.