Abstract

Many scientific investigations require data-intensive research where big data are collected and analyzed. To get big insights from big data, we need to first develop our initial hypotheses from the data and then test and validate our hypotheses about the data. We propose FS-S, a flexible and modular Scala based implementation of the Fixed Size Least Squares Support Vector Machine (FS-LSSVM) for large data sets. The framework consists of a set of modules for (gradient and gradient free) optimization, model representation, kernel functions and evaluation of FS-LSSVM models. A kernel based Fixed-Size Least Squares Support Vector Machine (FSLSSVM) model is implemented in the proposed framework, while heavily leveraging the parallel computing capabilities of Apache Spark. Global optimization routines like Coupled Simulated Annealing (CSA) and Grid Search are implemented and used to tune the hyper-parameters of the FS-LSSVM model. Finally, we carry out experiments on benchmark data sets like Magic Gamma, Forest Cover, Susy and higgs etc. and evaluate the performance of various kernel based FS-LSSVM models, all these combine to reveal an effective and efficient way to perform closed-loop big data analysis with visualization and scalable computing.
References

22. Giorgio Valentini, D S I Dipartimento, and Thomas G Dietterich. Bias-Variance Analysis
of Support Vector Machines for the Development of SVM-Based Ensemble Methods. Journal of
simulated annealing. IEEE Transactions on Systems, Man, and Cybernetics, Part B:

Index Terms
Computer Science Information Systems

Keywords
FS-LSSVM, Big Data, Large Scale Models, Non-linear SVMs