Incorporating Prior Knowledge into SVM Algorithms in Analysis of Multidimensional Data

Marcin Orchel

Abstract

In this thesis, we present results for research conducted by us regarding the regression method, called δ support vector regression (δ-SVR), the method of incorporating knowledge about margin of an example, called φ support vector classification (φ-SVC), implementation of support vector machines (SVM) and application of SVM to executing stock orders. In this report, we propose a method, called δ-SVR, that replaces a regression problem with binary classification problems which are solved by SVM. We analyze statistical equivalence of a regression problem with a binary classification problem. We show potential possibility to improve generalization error bounds based on Vapnik-Chervonenkis (VC) dimension, compared to SVM. We conducted experiments comparing δ-SVR with ε-insensitive support vector regression (ε-SVR) on synthetic and real world data sets. The results indicate that δ-SVR achieves comparable generalization error, fewer support vectors, and smaller generalization error over different values of ε and δ. The δ-SVR method is faster for linear kernels while using sequential minimal optimization (SMO) solver, for nonlinear kernels speed results depend on the data set. In this report, we propose a method called φ-SVC for incorporating knowledge about margin of an example for classification and regression problems. We propose two applications for φ-SVC: decreasing the generalization error of reduced models while preserving the similar number of support vectors, and incorporating the nonlinear constraint of a special type to the problem. The method was tested for SVM classifier and ε-SVR. Experiments on real world data sets show decreased generalization error of reduced models for linear and polynomial kernels. In this report, we propose two implementation improvements, the first one for speed of training of SVM, the second one for simplifying implementation of SVM solver. The first improvement, called heuristic of alternatives (HoA), regards a new heuristic for choosing parameters to the working set. It checks not only satisfaction of Karush-Kuhn-Tucker (KKT) conditions, but also growth of an objective function. Tests on real world data sets show, that HoA leads to decreased time of training of SVM, compared to the standard heuristic. The second improvement, called Sequential Multidimensional Subsolver (SMS), regards a new method of solving subproblems with more than two parameters, instead of using complicated quadratic programming solvers, we use SMO method. We achieve simpler implementation with similar speed performance. In this report, we propose an application of support vector regression (SVR) for executing orders on stock markets. We use SVR for predicting a function of volume participation. We propose the improvement of predicting participation function by using SVM with incorporated additional nonlinear constraint to the problem. We show that quality of the prediction influences execution costs. Moreover, we show how we can incorporate knowledge about stock prices. We compared ε-SVR and δ-SVR with simple predictors such as the average price of execution from previous days. The tests were performed on data for stocks from NASDAQ-100 index. For both methods we achieved smaller variance of execution costs. Moreover, we decreased costs of order execution by using prediction of stock prices.