SCALING UP
SUPPORT VECTOR MACHINES

by

WAI-HUNG TSANG

Department of Computer Science

The Hong Kong University of Science and Technology

ABSTRACT

Kernel methods, such as support vector machines (SVMs), have been successfully used in various aspects of machine learning problems, such as classification, regression, and ranking. Many of them are formulated as quadratic programming (QP) problems, which take $O(m^3)$ time and $O(m^2)$ space complexities, where m is the training set size. It is, thus, computationally infeasible on massive data sets. By observing that practical SVM implementations only *approximate* the optimal solution by an iterative strategy, I scale up kernel methods by exploiting such "approximateness" in this thesis.

First, I show that SVM classification problems can be equivalently formulated as minimum enclosing ball (MEB) problems in computational geometry. Then, by adopting an efficient approximate MEB algorithm, I obtain provably approximately optimal solutions with the idea of core-sets. My proposed Core Vector Machine (CVM) algorithm can be used with nonlinear kernels and has a time complexity that is linear in m and a space complexity that is independent of m for a fixed approximation factor $(1 + \epsilon)^2$. Experiments on large real-world data sets demonstrate that the CVM is as accurate as existing SVM implementations but is much faster and can handle much larger data sets than existing scale-up methods.

By generalizing the underlying MEB problem as the center-constrained minimum enclosing ball (CCMEB) problem, I extend the CVM algorithm to the regression and ranking setting. Moreover, the condition on the kernel function is relaxed. Thus, the enhanced CVM algorithm can be used with any linear/nonlinear kernels.

Finally, I introduce the enclosing ball (EB) problem where the ball's radius is fixed and thus does not have to be minimized. I develop efficient $(1+\epsilon)$-approximation algorithms that are simple to implement and do not require any sophisticated numerical solver. For the Gaussian kernel in particular, a suitable choice of this (fixed) radius is easy to determine, and the center obtained from the $(1+\epsilon)$-approximation of this EB problem is close to the center of the corresponding MEB. Experimental results show that the proposed algorithm has accuracies comparable to the other large-scale SVM implementations, but can handle very large data sets and is even faster than the CVM in general.