FAST AND KNOWLEDGE BASED SVM ALGORITHMS WITH APPLICATIONS

by

M. ARUN KUMAR
Department of Electrical Engineering

Submitted
in fulfillment of the requirements of the degree of
DOCTOR OF PHILOSOPHY

to the

INDIAN INSTITUTE OF TECHNOLOGY, DELHI

NOVEMBER 2009

Certificate

This is to certify that the thesis entitled, "**Fast and Knowledge based SVM Algorithms with Applications**", being submitted by **M. Arun Kumar** to the Department of Electrical Engineering, Indian Institute of Technology, Delhi, for the award of the degree of **Doctor of Philosophy** is the record of bona-fide research work carried out by him under my supervision. In my opinion, the thesis has reached the standards of fulfilling the requirements of the regulations relating to the degree.

The results obtained here in have not been submitted to any other University or Institute for the award of any degree or diploma.

(Prof. M. Gopal)
Thesis Supervisor
Department of Electrical Engineering
Indian Institute of Technology, Delhi
Hauz Khas, New Delhi, 110016
India

Acknowledgements

I owe my deepest gratitude to Prof. M. Gopal for introducing me to the world of pattern recognition. More than being my thesis advisor, he has been my friend, philosopher and guide in the truest sense. He has been very generous with his time and has been a constant source of inspiration and support.

My heartfelt thanks to Prof. Suresh Chandra for teaching me the fundamentals of optimization as part of my course work. His valuable inputs thereafter have contributed immensely to this thesis. Chapter 4 is a joint work with him. I would like to record my debt of gratitude to Dr. J. Prakash for being my mentor, and for giving me the unforgettable opportunity to work and learn from him while I was at Annamalai University. I am extremely grateful to my SRC members Dr. Brejesh Lall and Dr. M. Nabi, for their helpful suggestions and advice. I do convey my sincere gratitude and respect to Prof. B. Chandra, and Prof. A.N. Jha, who have taught me all the relevant courses at IITD.

I am grateful to the staffs of PG section, Electrical Engineering, and Central library, for their valuable co-operation. I am indebted to Shri Jaipal Singh, and Shri Virender Singh, for providing me immense facilities and assistance to carry out my research work. A very special thanks to former and present research scholars for all their help – Dr. Rajen Bhatt, Dr. Nischal Verma, A.K. Dwivedi, Dr. S. Gopinath, Dr. Reshma Khemchandani, Hitesh Shah, Bharat Sharma, Deepak Adhyaru, Mahendra Kumar, Ethesham Hassan, Vivek, Shivaram, Tapasee, Madhan Mohan, Ravi Kumar Pandi and Packiam.

And finally, deepest felt thanks to my parents for their unconditional love and support.

M. Arun Kumar

Abstract

This thesis deals with the development of novel algorithms for the problems of binary and multiclass classification. These algorithms are in the support vector machines framework, which aims at minimizing an upper bound on the generalization error through maximizing the margin between two disjoint parallel hyperplanes. Support vector machines (SVM) have emerged to become a powerful paradigm for pattern classification in recent years. The proposed algorithms in this thesis aim at fast training, knowledge incorporation and fast testing of SVM and its variants.

Two formulations namely smooth twin SVM and least squares twin SVM have been proposed to speed-up the training of twin SVMs. Twin SVM is a recently proposed variant of conventional SVM that does binary classification using two non-parallel hyperplanes and has received increased attention because of improved generalization and reduced computational complexity. In smooth twin SVM and least squares twin SVM we attempt to solve primal quadratic programming problems (QPPs) instead of dual QPPs usually solved in twin SVM. Smooth twin SVM solves the primal QPPs by converting them into smooth unconstrained minimization problems using Newton-Armijo algorithm. Least squares twin SVM solves the modified primal QPPs as systems of linear equations. Both algorithms achieve significant speed-up in training when compared to twin SVM. Applications to text categorization have been reported with least squares twin SVM.

The problem of incorporating prior knowledge into SVM variants based on two non-parallel hyperplanes has been addressed in the proposed knowledge based twin/least squares twin SVM formulations. Here prior knowledge in the form of multiple polyhedral sets, each belonging to one of the two classes is introduced into twin/least squares twin SVM

formulations with the use of theorems of alternative. Both formulations are capable of generating non-parallel hyperplanes based on real-world data and prior knowledge. Computational comparisons on applications such as breast cancer diagnosis and DNA promoter recognition demonstrate the versatility of the proposed algorithm over other existing approaches.

A novel idea is the introduction of hybrid SVM based decision tree classifiers to speed-up testing of binary and multiclass SVMs. Here, using probabilistic outputs of binary SVM classifiers, two algorithms namely decision tree based one-against-all for multiclass SVM classification and hybrid SVM based decision tree for binary classification have been proposed. Both algorithms reduce the average number of binary SVMs to be evaluated in classifying a test dataset and thereby decreasing the testing time. A threshold parameter introduced based on probabilistic outputs of binary SVM acts as a trade-off parameter between classification speed and accuracy. Extensive computational comparisons show the remarkable reductions achieved by both these algorithms.

Contents