

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

Certificate	i
Acknowledgements	ii
Abstract	iii
Contents	v
List of Figures	ix
List of Tables	xii
1. Introduction	1
1.1 Notations	3
1.2 Support Vector Machines	4
1.3 Generalized Eigenvalue Proximal Support Vector Machines	7
1.4 Twin Support Vector Machines	10
1.5 Multiclass SVM Classification	13
1.6 Text Categorization using SVMs	16
1.7 Decision Trees	18
1.8 Thesis Structure	21
2. Smooth Twin Support Vector Machines	25
2.1 Introduction	26
2.2 Smooth Twin Support Vector Machines	29
2.3 STSVM with Nonlinear Kernel	33
2.4 Experimental Results and Discussion	34
2.5 Chapter Conclusions	40
3. Least Squares Twin Support Vector Machines	43
3.1 Introduction	44

3.2	Least Squares Twin Support Vector Machines	45
3.3	LSTSVM with Nonlinear Kernel	48
3.4	Experimental Results on Standard Datasets	52
3.5	Application to Multilabel Text Categorization	56
3.5.1	Document representation	56
3.5.2	Data collections	57
3.5.3	Evaluation methodology	58
3.5.4	Experimental results	59
3.6	Chapter Conclusions	61
4	Knowledge Based Least Squares Twin Support Vector Machines	63
4.1	Introduction	64
4.2	Knowledge based Proximal SVM	65
4.2.1	Original formulation and Solution	65
4.2.2	An alternate solution	69
4.3	Knowledge Based Twin SVM	71
4.4	Knowledge Based Least Squares Twin SVM	75
4.5	Experimental Results and Discussion	79
4.6	Chapter Conclusions	83
5	One-Against-All Multiclass SVM Classification: Comparison Studies and	85
	Improvements	
5.1	Introduction.	86
5.2	Improvements to OAA and OAO	88
5.2.1	DAGSVM	88
5.2.2	BTS and c-BTS	88

5.2.3 AO	89
5.3 Empirical Comparison on Unilabel Text Categorization	90
5.3.1 Document representation	90
5.3.2 Data collections	91
5.3.3 SVM hyper-parameter optimization	92
5.3.4 Experimental results and discussion	93
5.4 Decision Tree Based OAA	103
5.5 Experimental Results and Discussion	109
5.6 Chapter Conclusions	114
6 Hybrid SVM Based Decision Tree	117
6.1 Introduction	118
6.2 A Comparison Study on SVM and DT	119
6.3 Hybrid SVM Based Decision Tree	122
6.3.1 Motivation and formulation	122
6.3.2 SVMMDT	125
6.3.3 Closeness measure	128
6.3.4 SVMMDT algorithm	129
6.3.5 Significance of the threshold δ	130
6.3.6 Over-sampling for improved representation of δ -region	131
6.4 Experimental Results	133
6.4.1 Adult datasets	133
6.4.2 Checkerboard dataset	142
6.4.3 Other binary datasets	143
6.5 Chapter Conclusions	144

7	Thesis Conclusions and Future Research	147
	7.1 Smooth Twin Support Vector Machines	148
	7.2 Least Squares Twin Support Vector Machines	148
	7.3 Knowledge Based Least Squares Twin Support Vector Machines	149
	7.4 One-Against-All Multiclass SVM Classification: Comparison, Studies and Improvements	150
	7.5 Hybrid SVM Based Decision Tree	151
	7.6 Summary	152
	Appendix I: Datasets Description	155
	References	165
	List of Publications.	177
	Technical Biography of Author.	179