

CERTIFICATE

It is certified that the work contained in the thesis titled “*Robust SVM Variants and Their Applications*” by *Manisha Singla* has been carried out under my supervision and that this work has not been submitted elsewhere for a degree.

It is further certified that the student has fulfilled all requirements of Comprehensive Examination, Candidacy, and SOTA for the award of Ph.D. Degree.

Supervisor

Prof. K K Shukla

Professor,

Department of Computer
Science and Engineering,
Indian Institute of Technology
(BHU) Varanasi,
Uttar Pradesh, INDIA 221005.

Co-Supervisor

Dr. Debdas Ghosh

Assistant Professor,

Department of Mathematical
Sciences,
Indian Institute of Technology
(BHU) Varanasi,
Uttar Pradesh, INDIA 221005.

DECLARATION BY THE CANDIDATE

I, **Manisha Singla**, certify that the work embodied in this Ph.D. thesis is my own bonafide work carried out by me under the supervision of **Prof. K K Shukla** and co-supervision of **Dr. Debdas Ghosh** from **July 2017** to **March 2021** at **Department of Computer Science and Engineering**, Indian Institute of Technology (BHU) Varanasi. The matter embodied in this thesis has not been submitted for the award of any other degree/diploma. I declare that I have faithfully acknowledged and given credits to the research workers wherever their works have been cited in my work in this thesis. I further declare that I have not willfully copied any other's work, paragraphs, text, data, results, *etc.* reported in journals, books, magazines, reports, dissertations, theses, *etc.*, or available at websites and have not included them in this thesis and have not cited as my own work.

Date:

Place: Varanasi

(**Manisha Singla**)

CERTIFICATE BY THE SUPERVISORS

It is certified that the above statement made by the candidate is correct to the best of our knowledge.

Supervisor

(**Prof. K K Shukla**)

Professor,

Department of Computer Science
and Engineering,
Indian Institute of Technology
(BHU) Varanasi.

Co-Supervisor

(**Dr. Debdas Ghosh**)

Assistant Professor,

Department of Mathematical
Sciences,
Indian Institute of Technology
(BHU) Varanasi.

Signature of Head of Department

COPYRIGHT TRANSFER CERTIFICATE

Title of the Thesis: Robust SVM Variants and Their Applications

Name of the Student: Manisha Singla

Copyright Transfer

The undersigned hereby assigns to the Indian Institute of Technology (Banaras Hindu University) Varanasi all rights under copyright that may exist in and for the above thesis submitted for the award of the *Doctor of Philosophy*.

Date:

Place: Varanasi

(Manisha Singla)

Note: However, the author may reproduce or authorize others to reproduce material extracted verbatim from the thesis or derivative of the thesis for author's personal use provided that the source and the Institute's copyright notice are indicated.

Dedicated to my parents,

Mrs. Sunita Singla

and

Mr. Puran Chand Singla

ACKNOWLEDGEMENT

Though, only my name appears on the cover of this dissertation, so many great people have contributed to its production. I owe my gratitude to all those people who have made this thesis possible and because of whom my post graduate experience has been one that I will cherish forever.

I take this opportunity to express my profound gratitude and deep regards to my supervisor Dr. K. K. Shukla, Professor, Department of Computer Science and Engineering, IIT(BHU), Varanasi and co-supervisor Dr. Debdas Ghosh, Assistant Professor, Department of Mathematical Sciences, IIT(BHU), Varanasi for their exemplary guidance, monitoring and constant encouragement throughout the course of this dissertation. I am obliged to faculty members, in particular Prof. S. K. Singh, HoD, Department of Computer Science and Engineering, IIT(BHU), Varanasi and Prof. T. Som, HoD Department of Mathematical Sciences, IIT(BHU), Varanasi. In addition, I want to express my deepest gratitude to Mrs. Mridula Verma who supported me throughout my thesis work. I am grateful to Mr. Gaurav Bansal , who have provided me through technical, moral and emotional support during my research program.

A very special gratitude goes out to my colleagues and friends with special mention to Mrs. Anviti Pandey, Ms. Vandana Bharti, Ms. Pratishta Verma, and Mr. Jitender Kumar.

Last but not the least, I specially thank my parents, Mr. Puran Chand Singla and Mrs. Sunita Singla, my parents-in-law Mr. Yogender Bansal and Mrs. Beena Bansal and my brother Ravi Singla for their constant support and encouragement, without which this assignment would have not been completed at all. I am grateful to my other family members who have supported me along the way.

(Manisha Singla)

Contents

List of Figures	xii
List of Tables	xiv
List of Symbols	xv
List of Abbreviations	xvi
Preface	xix
1 Introduction	1
1.1 Variants of SVM	3
1.2 Hinge Loss Function	12
1.3 Robust Statistics	14
1.4 Motivation	16
1.5 Research Contributions	16
1.6 Organization of the Thesis	18
2 Related work	19
2.1 Robust Statistics based SVM	19
2.2 Robust Statistics based Variants of SVM	20
2.3 Robust Loss Functions in SVM and its variants	23
2.4 Applications of Robust SVM	28
3 RSVM-PDProx: Adding Robustness to SVM Using Rescaled α-hinge Loss Function	31
3.1 Introduction	31
3.1.1 Motivation and Contribution	33
3.1.2 Outline	34
3.2 Related Concepts	34

3.3	RSVM-PDProx	37
3.3.1	Rescaled α -hinge Loss Function	37
3.3.2	Non-smooth Regularizer, $\ \cdot\ _2$	38
3.4	Analysis of RSVM-PDProx	41
3.4.1	Rate of Convergence of RSVM-PDProx	42
3.4.2	Time Complexity of RSVM-PDProx	45
3.5	Numerical Experiments and Results	46
3.5.1	Experiments on Synthetic Data Sets	46
3.5.2	Experiments on Real-world Data Sets	48
3.6	Discussion about dual variable and regularization parameter	55
3.7	Summary of the Work	57
4	Res-TSVR: Robust Twin Support Vector Regression based on Rescaled Hinge Loss Function	59
4.1	Introduction	59
4.1.1	Motivation Behind This Work	61
4.1.2	Contribution of This Work	62
4.1.3	Outline	62
4.2	Brief Introduction to ε -SVR and TSVR	63
4.2.1	Support Vector Regression	63
4.2.2	Twin Support Vector Regression	64
4.3	Res-TSVR	66
4.3.1	Problem Formulation	67
4.3.2	Linear Res-TSVR	70
4.3.3	Non-Linear Res-TSVR	73
4.3.4	Algorithm Res-TSVR	74
4.4	Analysis of Res-TSVR	76
4.4.1	Convergence Proof of Algorithm 2	77
4.4.2	An Estimate of k_{\max} in Algorithm 2 for an ε -Precision Solution	78
4.5	Numerical Experiments and Results	79
4.5.1	Performance Criteria	80
4.5.2	Performance on Synthetic Data Sets	81
4.5.3	Performance on Real-world Data Sets	85
4.6	Discussion About the Rescaling Parameter, $\hat{\eta}$	90
4.7	Summary	92

5	Robust Diabetic Retinopathy Detection Using Twin Support Vector Machines	95
5.1	Introduction	96
5.1.1	Motivation and Contributions	97
5.2	Related work	99
5.3	Dataset Used	101
5.3.1	Feature Extraction	101
5.3.2	Number and Area of Microaneurysms With a Density of Hard Exudates	102
5.3.3	Blood Vessels Extraction	103
5.3.4	Entropy	103
5.3.5	Standard Deviation	105
5.3.6	Morphological Transformations	105
5.4	Classifiers Used	105
5.5	Numerical Experiments and Results	109
5.5.1	Results On DIARETDB1 Data Set (Without Noise)	109
5.5.2	Results On DIARETDB1 Data Set (With Gaussian Noise)	110
5.6	Summary	114
6	$\overline{\text{pin}}$-TSVM: A robust transductive support vector machine and its application to the detection of COVID-19 infected patients	115
6.1	Introduction	115
6.1.1	A Brief Introduction of TSVM	116
6.1.2	Motivation and Contribution	119
6.1.3	Outline	120
6.2	$\overline{\text{pin}}$ -TSVM: A Robust Transductive Support Vector Machine with Truncated Pinball Loss Function	120
6.3	Numerical Experiments	129
6.3.1	Experiments on Real-world Data Sets	131
6.4	Application to the Detection of Novel Coronavirus (COVID-19) Infected Patients using Chest X-ray Images	140
6.5	Summary	142
7	Conclusion and Future Directions	145
7.1	Conclusion	145
7.2	Future Directions	147

References	150
Appendices	166
List of Publications	173

List of Figures

1.1	Types of Machine Learning Algorithms	2
1.2	Support Vector Machine	3
1.3	Variants of SVM [1]	4
1.4	Hinge Loss Function	13
1.5	Generalized Framework of Robust Statistics [1]	15
2.1	Robust Loss Functions Used with SVM and its Variants	27
3.1	Rescaled Hinge Loss Function with Different η Values	36
3.2	Rescaled α -hinge Loss Function with Different η Values	37
3.3	Comparison of Rescaled α -hinge Loss Function and Rescaled Hinge Loss Function at $\eta=4$	38
3.4	Comparison of SVM, RSVM-RHHQ and RSVM-PDProx over Synthetic Data Set with 0% Noise	47
3.5	Comparison of SVM, RSVM-RHHQ and RSVM-PDProx over Synthetic Data Set with 15% Noise	47
3.6	Comparison of SVM, RSVM-RHHQ and RSVM-PDProx over Synthetic Data Set with 30% Noise	48
3.7	Comparison of RSVM-RHHQ and RSVM-PDProx Based on Support Vector Ratio over Real-World Data Sets	55
3.8	Effect of λ on the Accuracy at $\eta = 3$	56
3.9	Effect of λ on the Support Vector Ratio at $\eta = 3$	57
4.1	ε -Insensitive Loss Function	63
4.2	Loss Functions used in TSVR	65
4.3	Hinge Loss and Rescaled Hinge Loss Function with Different $\hat{\eta}$ Values	68
4.4	Comparison of the Proposed Technique with Other Methods in the Absence and Presence of Noise (Outliers are Indicated by Arrows in Case of Noisy Data)	84

4.5	Effect of Increasing $\hat{\eta}$ Values on RMSE and Computational Time . . .	91
4.6	The RMSE of Different Regressors Against SNR	92
5.1	Hard Exudates [2]	97
5.2	Microaneurysms [2]	98
5.3	Soft Exudates [2]	99
5.4	Figure Showing the Eye Fundus Images of the Diseased Person and the Normal Person [3]	101
5.5	Figure Showing the Steps Followed To Extract The Features From the Eye Fundus Images; The Bottom Circles Represent The Features Used In This Work	102
5.6	Figure Showing Effect of The Steps Applied to the Original Image to Obtain the Final Image	104
5.7	The Conventional Hinge Loss Function	107
5.8	Pinball Loss Function	108
5.9	Plot of Accuracy versus Regularization parameter Without Adding Noise to The Data Set	110
5.10	Plot of Accuracy versus Regularization Parameter After Adding Noise To The Data Set	111
5.11	Box Plot of Accuracies on Linear SVM	112
5.12	Box Plot of Accuracies on TWSVM with Hinge Loss	112
5.13	Box Plot of Accuracies on TWSVM with Pinball Loss	113
6.1	Hinge Loss Function for Labeled and Unlabeled Samples	117
6.2	Truncated Pinball Loss Function with $\tau = s = 0.5$	121
6.3	Chest X-ray Images of Humans with a) Bacterial Infection b) Coronavirus c) Normal X-ray	141
6.4	Steps to Extract Features From the COVID-19 Data Set and Training $\overline{\text{pin}}$ -TSVM	142

List of Tables

1.1	Summary of Variants of SVM	10
3.1	Data Sets Used for Experimentation Purposes	49
3.2	Results of RSVM-PDProx and the Existing Methods with 0% Noise in the Real-World Data Sets	50
3.3	Results of RSVM-PDProx and the Existing Methods with 15% Noise in the Real-World Data Sets	52
3.4	Results of RSVM-PDProx and the Existing Methods with 30% Noise in the Real-World Data Sets	53
3.5	Mean Support Vector Ratios Over All the Data Sets	55
4.1	Performance Metrics	80
4.2	Synthetic Data Sets Used For Experimentation Purposes	81
4.3	Comparison of Various Techniques over Synthetic Data Sets using RBF Kernel (Without Noise)	82
4.4	Comparison of Various Techniques over Synthetic Data Sets using RBF Kernel (With Gaussian Noise $\mathcal{N}(0, 0.2)$)	82
4.5	Comparison of Various Techniques over Synthetic Data Sets using RBF Kernel (With Uniform Noise $\mathcal{U}(0, 0.2)$)	83
4.6	Real-world Data Sets Used for Experimentation Purposes	85
4.7	Comparison of Various Techniques over Real-World Data Sets using RBF Kernel (With Gaussian Noise $\mathcal{N}(0, 0.25)$)	86
4.8	Comparison of Various Techniques over Real-World Data Sets using RBF Kernel (With Uniform Noise $\mathcal{U}(0, 0.25)$)	87
4.9	Ranks of Various Techniques over All the Data Sets with Gaussian Noise	87
4.10	Average Ranks of all the Techniques	89
4.11	Results of p -Significance Test Comparing Res-TSVR with Existing Methods at 1% Significance Level	90

5.1	Accuracy and Computational Time Of Different SVM Variants On DIARETDB1 Data Set	113
6.1	Literature Survey of TSVM robust towards noise	118
6.2	Comparison of Various Techniques on Synthetic Data Set Using Linear Kernel	130
6.3	Small Data Sets Used for Experimentation Purposes	130
6.4	Comparison of Various Techniques with 0% Noise in the Real-World Data Sets	131
6.5	Comparison of Various Techniques with 15% Noise in the Real-World Data Sets	132
6.6	Comparison of Various Techniques with 30% Noise in the Real-World Data Sets	134
6.7	Used Large Data Sets for Experimentation Purposes	136
6.8	Comparison of Various Techniques over Large Real-World Data Sets using Linear Kernel	137
6.9	Comparison of Various Techniques over COVID-19 Data Set	143
B.1	Results of RSVM-PDProx and the Existing Methods on Synthetic Data Sets	168
D.1	Comparison of Different SVM Variants On DIARETDB1 Data Set	170
D.2	Comparison of Different SVM Variants On Messidor Data Set	170
E.1	Comparison of $\overline{\text{pin}}$ -TSVM and the Existing Methods over COVID-19 Data Set	172

List of Symbols

Symbol	Description
w	Weight Vector
b	Bias Term
λ	Regularization Parameter
e	Vector of ones
x	Input Vector
m	number of Instances
y	Target Vector
L_1	Absolute-value Norm
L_2	Euclidean Norm
L_K	Laplace Kernel Induced Loss Function
z	Margin Variable
\hat{w}	Optimal Weight Vector
\hat{b}	Optimal Bias Term
η	Rescaling Parameter of RSVM-PDProx
T	Total number of iterations
$\hat{\eta}$	Rescaling Parameter of Res-TSVR
τ	Parameter Used in Pinball Loss Function
α	Dual Variable
ϵ	Tolerance Value
R^2	Coefficient of Determination
\hat{Y}	Predicted Target Values
\bar{Y}	Mean of Y Values
L	Labeled Instances
U	Unlabeled Instances
ϕ	Kernel Function

Abbreviations

Abbreviation	Description
SVM	Support Vector Machine
PDPprox	Primal-dual Proximal
TSVR	Twin Support Vector Regression
TWSVM	Twin Support Vector Machine
TSVM	Transductive Support Vector Machine
SGD	Stochastic Gradient Descent
ML	Machine Learning
SVR	Support Vector Regression
ECG	Electrocardiography
LS-SVM	Least-square SVM
FSVM	Fuzzy SVM
GEPSVM	Generalized Eigen Proximal SVM
PSVM	Proximal Support Vector Machine
OCSVM	One-class Support Vector Machine
CCCP	Concave-convex Procedure
DR	Diabetic Retinopathy
RSVR	Robust Support Vector Regression
DC	Difference of Convex
VAD	Voice Activity Detection
MRI	Magnetic Resonance Imaging
LRMD	Low-rank Matrix Decomposition
SV	Support Vectors
W-ETSVR	Weighted ϵ -TSVR
KKT	Karush-kuhn Tucker
NMSE	Normalized Mean Square Error
RMSE	Root Mean square Error
SSE	Sum Square Error

TSS	Total Sum Squares
SSR	Sum of Squares of Residuals
RBF	Radial Basis Function
SNR	Signal Noise Ratio
Ma	Microaneurysms
NPDR	Non-proliferative Diabetic Retinopathy
MPDR	Moderately Proliferative Diabetic Retinopathy
PDR	Proliferative Diabetic Retinopathy
SE	Soft Exudates
GMM	Gaussian Mixture Model
KNN	k- nearest Neighbor
ADHE	Adaptive Histogram Equalization
PDF	Probability Density Function
CNN	Convolutional Neural Network
PCA	Principal Component Analysis

PREFACE

Although support vector machines (SVMs) are one of the oldest machine learning approaches, both the regression and classification variants still help solve real-world tasks. However, the model has a limitation of sensitivity towards noise in the data set. The performance of SVM gets adversely affected when the model is trained with noisy data. This is due to the presence of unbounded hinge loss function. Since the other SVM variants also comprise hinge loss function, the limitation is inherited in all the SVM variants.

In this thesis, the robust formulations of SVM and its variants, which can handle sensitivity towards the noise, are proposed. This comes under the category of robust statistics. For classification tasks, the robustness against label noise in the data sets is added, while the robustness against Gaussian and uniform noise in the data sets is added for regression tasks.

First, the robustness against label noise in the conventional SVM is added. In this literature, the rescaled α -hinge loss function with a non-smooth L_2 regularizer is used. To solve the non-smooth optimization technique, the primal dual proximal (PDProx)-dual technique is implemented. The proposed approach is observed to be more sparse than the existing robust SVMs. The model converges at the rate of $O(1/T)$ where T denotes the number of iterations.

Next, the robustness against the uniform and Gaussian noise in a regression model, twin support vector regression (TSVR) has been added. The rescaled hinge loss func-

tion is used in this literature. To solve the non-convex problem, the half-quadratic optimization technique is used. Subsequently, an algorithm, Res-TSVR, has been devised to implement the proposed approach. It is found that the maximum number of iterations required to achieve an ϵ -precision solution is $O(\log(1/\epsilon))$.

Next, the twin support vector machine (TWSVM) is used in diabetic retinopathy detection using eye fundus images. In this work, the pinball loss function is used with TWSVM to add robustness to the model. It also helped in reducing the computational time. In all the experiments, the proposed approaches are compared with the existing approaches to prove the superiority of the proposed works.

The above contributions discussed are in the field of supervised machine learning. A semi-supervised machine learning framework has also been made robust to the label noise. Next contribution belongs to the addition of robustness to the conventional transductive support vector machine (TSVM). The semi-supervised learning model is robustified against label noise using a truncated pinball loss function. The model is tested on both small and large scale real-world data sets. The proposed approach is implemented using both the stochastic gradient descent (SGD) method and the dual problem solver, `mlcv_quadprog` (name is based on the machine learning-computer vision lab). The robust model is also applied to the detection of COVID-19 infected patients using chest X-ray images.