

FEATURE SELECTION IN HIGH DIMENSIONS WITH APPLICATIONS IN BIO-INFORMATICS

YAMUNA PRASAD



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY DELHI

JULY 2016

©Indian Institute of Technology Delhi (IITD), New Delhi, 2016

FEATURE SELECTION IN HIGH DIMENSIONS WITH APPLICATIONS IN BIO-INFORMATICS

by

YAMUNA PRASAD

Department of Computer Science and Engineering

Submitted

in fulfillment of the requirements of the degree Doctor of Philosophy

to the



Indian Institute of Technology Delhi
JULY 2016

Certificate

This is to certify that the thesis titled **Feature Selection in High Dimensions with Applications in Bio-informatics** being submitted by **Mr. Yamuna Prasad** for the award of **Doctor of Philosophy** in Computer Science and Engineering is a record of bona-fide work carried out by him under our guidance and supervision at the Computer Science and Engineering Department, Indian Institute of Technology Delhi. To the best of our knowledge, the work presented in this thesis has not been submitted elsewhere, either in part or full, for the award of any other degree or diploma.

K. K. Biswas

Professor

Department of Computer Science and Engineering

Indian Institute of Technology Delhi

New Delhi 110 016

M. Hanmandlu

Professor

Department of Electrical Engineering

Indian Institute of Technology Delhi

New Delhi 110 016

Acknowledgments

First of all, I wish to express my deepest and sincere gratitude to my supervisor Prof. K. K. Biswas for his enlightening guidance and suggestions in the development and completion of the thesis work. I have been fortunate enough to have a guide who gave me the freedom to explore on my own, and at the same time guide me to recover when my steps faltered. I will always be thankful to him for his wisdom, knowledge, and deep concern for me.

My co-advisor, Prof. M. Hanmandlu, has been always there to listen and give advice. I am grateful to him for the long discussions that helped me sort out the technical details of my work. Prof. Hanmandlu always inspired me to improve my knowledge for soft computing with his sound and clear concepts in this domain.

I would also like to extend my sincere thanks to Dr. Parag Singla for his invaluable suggestions which helped me a lot in complimenting my thesis work. Further, I would like to thank my research committee members Prof. Amit kumar and Prof. Niladri Chatterjee for their valuable feedbacks. I would also like to thank Prof. S. K. Gupta and Prof. Saroj Kaushik for their motivation and encouragement.

I would like to thank my friends from IIT Delhi - Arun Parakh, Shibhasis Guha and Chinmay Narayan for their help in coding; Manoj Gupta, Nisha Jain, Mona Jain, Syamantak Das, Swati Sharma, Dinesh Khandelwal, Yashoteja Prabhu, Happy Mittal, Suvam Patra, Ram Pyare and friends from IIIT Noida - Dr. Chakresh Kumar Jain and Dr. Manish Thakur for their cooperation and support. I would also like to thank the staff members of Computer Science & Engineering department especially our lab instructors Mr. K. R. Kaushik and Mr. Surendra S. Negi; office staff members Renu Bhatia, Rekha Rathi, Abhisek Sharma and Hemant Singh for their help.

A special gratitude and love goes to my family for their support. I thank my parents

for their abiding love and support. My brother Saryu Prasad, sister Archana and uncle C. P. Ojha also deserve a special mention for their support. Finally, I would like to dedicate this thesis and everything I do to my wife Ankita for her support and sharing the hard work during the most difficult times.

Yamuna Prasad

Abstract

Many machine learning problems in Bioinformatics, Vision and Natural Language Processing (NLP) need to deal with very high dimensional data. Generally, the analysis of these datasets is carried out through classification/regression. The presence of a large number of features in a dataset leads to poor generalization accuracy and high execution time. Several methods have been proposed in the literature to reduce the dimensionality through feature selection. These methods can be broadly categorized into filter based, wrapper based and embedded methods. The time complexity of these methods vary from quadratic to cubic in the number of dimensions which limits their applicability for the datasets in high dimensions.

In this thesis, we have addressed this issue by developing a faster feature selection approaches in all the three categories. In the filter based category, we have proposed two new approaches. The first one integrates a clustering scheme with Quadratic Programming Feature Selection (QP) method in an interleaved manner, reaping the advantage of clustering at various levels of granularity. The second approach involves a max-margin framework for feature selection posed as a One class SVM problem which optimizes both redundancy and relevance. Further, we have also proposed a Sequential Minimal Optimization (SMO) strategy for speeding up QP. We also show that the feature selection can be solved faster by incorporating a regularizer on feature weights in the objective of the QP method and solving it through a single variable decomposition

strategy. The time complexity of our proposed methods vary from linear to quadratic in the number of dimensions.

In the wrapper category, we have proposed two approaches namely, a recursive particle swarm optimization approach (RPSO) and a divide and conquer PSO approach (PPSO) followed by a Map-Reduce framework. We have also integrated linear SVM weight vector, Mutual Information, F-score and Wilcoxon's rank test with PSO for feature selection. These methods yield a minimal set of features without losing on accuracy.

In the embedded category, we have proposed a guided initialization of kernel weight vector in Generalized Multiple Kernel Learning (GMKL) by incorporating Fuzzy Rough Set scores and feature weights using QP method for faster convergence. Further, we have also proposed a multi-task learning framework using GMKL for siRNA efficacy prediction.

Exhaustive experiments have been carried out to validate our proposed methods on small, medium and large sized datasets from Bioinformatics, Vision and NLP domains. In most of the cases, our proposed approaches have been shown to be more efficient than the state-of-the-art feature selection methods while maintaining comparable level of accuracy in all the three categories.

Contents

Certificate	i
Acknowledgments	iii
Abstract	v
List of Figures	xiii
List of Tables	xvii
List of Abbreviations	xxi
1 Introduction	1
1.1 Introduction	1
1.1.1 Feature Selection using Filter based Methods	3
1.1.2 Feature Selection using Wrapper based Methods	5
1.1.3 Feature Selection using Embedded based Methods	7
1.2 Contribution of the Thesis	8
1.3 Thesis Organization	9
2 Speeding up Quadratic Programming Feature Selection	11
2.1 Introduction	11

2.2	Sequential Minimal Optimization for QP	12
2.2.1	Working Set Selection	15
2.2.2	The QPS Method	19
2.3	Variants of QPS	20
2.3.1	Speed-up of QPS with Nyström	20
2.3.2	Memory Reduction with QPS Computation on the Fly	21
2.3.3	Speeding up QPS through Parallelization	22
2.4	Regularized Quadratic Feature Selection	25
2.5	Experiments	29
2.5.1	Datasets	29
2.5.2	Methodology	29
2.5.3	Results	32
2.5.4	Results with Nyström Sampling	41
2.5.5	Biological Significance of Top Selected Genes	44
2.5.6	Comparative Study with proposed and state-of-the-art Methods	46
2.6	Summary	48
3	Integrated K-means Quadratic Programming Feature Selection	51
3.1	Introduction	51
3.2	Interleaved K-Means Feature Selection (IKQP)	52
3.2.1	Distance Metric for Similarity	53
3.2.2	Modified K-means for Features Clustering	54
3.2.3	The IKQP Algorithm	55
3.2.4	Convergence Issues	58
3.2.5	Time Complexity Analysis	59
3.2.6	Distance Computations	60

3.2.7	Interleaved K-Means Aggressive Feature Selection (IKAQP) . . .	60
3.3	Two-pass KN-means Feature Selection	61
3.3.1	The TKQP Algorithm	62
3.4	Experiments	66
3.4.1	Datasets	66
3.4.2	Methodology	66
3.4.3	Results	67
3.4.4	Biological Significance of Top Selected Genes	73
3.4.5	Comparative Study with proposed and state-of-the-art Methods	76
3.5	Summary	78
4	Max-Margin Framework for Feature Selection	81
4.1	Introduction	81
4.2	Proposed Max-Margin Framework	83
4.2.1	Formulation	84
4.2.2	Dual Formulation	85
4.2.3	Choice of Metrics	88
4.2.4	Dual Coordinate Descent for MMFS	89
4.3	Relationship to QP Method	91
4.4	Experiments	93
4.4.1	Datasets	93
4.4.2	Methodology	94
4.4.3	Results	95
4.4.4	Biological Significance of Top Selected Genes	96
4.4.5	Comparative Study with proposed and state-of-the-art Methods	100
4.5	Summary	102

5	Improving Accuracy with Wrapper based Approaches	105
5.1	Introduction	105
5.2	Background	107
5.2.1	Binary PSO	107
5.3	Proposed Recursive PSO	109
5.3.1	Integrated Approach for Feature Selection	112
5.4	Partition based PSO	113
5.5	Application: siRNA Efficacy Prediction	115
5.5.1	Feature Subset Selection Criteria	117
5.5.2	Ants Traversal Stopping Criteria in ACO	118
5.6	Experiments	118
5.6.1	Dataset	118
5.6.2	Results	120
5.6.3	Biological Significance of Selected Genes	130
5.6.4	Comparative Study with proposed and state-of-the-art Methods	130
5.7	Summary	132
6	Improving Convergence in Embedded Approaches	135
6.1	Introduction	135
6.2	Improvements to GMKL	136
6.2.1	QP-GMKL	137
6.2.2	FR-GMKL	139
6.2.3	l_{12} Regularization	142
6.3	Experiments	143
6.3.1	Dataset	143
6.3.2	Methodology	144

6.3.3	Effect of Initialization on Convergence	146
6.3.4	Effect of Regularization on Accuracy	150
6.3.5	Biological Significance of Selected Genes	151
6.3.6	Comparative Study with proposed and state-of-the-art Methods	152
6.4	Multi-Task Learning using GMKL	154
6.4.1	Efficacy Prediction of siRNA	155
6.4.2	Multi-task Multiple Kernel Learning	157
6.5	Experiments	158
6.6	Summary	161
7	Comparative Studies and Conclusions	163
7.1	Complexity Analysis of Proposed and state-of-the-art Methods	163
7.2	Observations on Proposed Feature Selection Approaches	164
7.3	Summary of Research Contributions	169
7.4	Future Research Directions	171
8	Conclusions	173
8.1	Summary of Research Contributions	173
8.2	Future Research Directions	175
	Bibliography	177
	Appendix A	185
	Publications From the Thesis	197

List of Figures

2.1	Parallel computation of max/min using map-reduce	23
2.2	Time Complexity	34
2.3	Figures 2.3(a)-2.3(f) represent accuracy for binary class datasets with varying number of top features. For RAOA, T2D and RAC datasets, instead of QPS and QP, accuracies with QPSN and QPN are presented.	37
2.4	Figures 2.4(a)-2.4(c) represent accuracy for multi-class datasets with varying number of top features. For GCM dataset, instead of QPS and QP, accuracies with QPSN and QPN are presented.	38
2.5	Figures 2.5(a)-2.5(f) represent Tuning for theta, l_1 and l_2 for Colon and SRBCT datasets	40
2.6	Figures 2.6(a)-2.6(d) represent speedup and efficiency for all datasets with parallel QPSFP and parallel RQFS	42
2.7	Box plot for Gene 765 in Colon dataset.	45
2.8	Box plot for Gene 4847 in Leukemia dataset.	45
3.1	Sub-clustering process at level i	58
3.2	Plots of Error rates for each methods with varying number of top $k(1-100)$ features for bioinformatics datesets	69

3.3	Plot of accuracies (%) for Colon dataset using IKQP with varying τ and varying number of top $k(1-100)$ features at fixed initial clusters $k'=15$.	72
3.4	Plot of accuracies (%) for Colon dataset using IKQP with varying number of initial clusters k' and varying number of top $k(1-100)$ features at fixed $\tau=0.8$	72
3.5	Box plot for Gene 2121 in Leukemia dataset.	76
4.1	Feature representation in sample space. The diagram is conceptual only.	86
4.2	Plots of average accuracies (in %)for each methods with varying number of top K features.	98
4.3	Plots of average execution times (in seconds) for each methods with varying number of top K features.	99
4.4	Accuracy and execution time for REAL-SIM dataset	100
4.5	Accuracy variation across γ and top k features	100
4.6	Box plot for Gene 1834 in Leukemia dataset.	101
5.1	Schematic diagram for Recursive PSO.	111
5.2	Schematic diagram for Partition based PSO.	116
5.3	Leave-one-out cross-validation accuracy at each recursive depths with recursive variants of our methods.	125
5.4	Reduction in number of features at each recursive depths with recursive variants of our methods (LOOCV).	126
5.5	Reduction in number of features at each partition level for partition based PSO methods PP and PPSW for LOOCV.	127
5.6	Comparison of accuracies for siRNA dataset	129
5.7	Box plot for Gene 2288 in Leukemia dataset.	131

6.1	Average number of SVM evaluations in 10 runs with 10 splits for binary classification. Figures 6.1(a) and 6.1(c) represents number of SVM evaluations for product combination and linear combination of kernels respectively for l_1 -regularization. Figures 6.1(b) and 6.1(d) represents average execution times for product combination and linear combination of kernels respectively for l_1 -regularization.	147
6.2	Average number of SVM evaluations in 10 runs with 10 splits for multi-class classification. Figures 6.2(a) and 6.2(c) represents number of SVM evaluations for product combination and linear combination of kernels respectively for l_1 -regularization. Figures 6.2(b) and 6.2(d) represents average execution times for product combination and linear combination of kernels respectively for l_1 -regularization.	148
6.3	Box plot for Gene 1779 in Leukemia dataset.	153

List of Tables

2.1	Datasets: detailed description	29
2.2	Methods and their terminologies	31
2.3	Comparison of average execution times.	33
2.4	Power curve fitting for time complexity	33
2.5	Comparison of memory requirements.	35
2.6	Polynomial curve fitting for memory complexity	35
2.7	Best accuracy (%) and corresponding number of features (NF)	36
2.8	Time required (in seconds) by QPSFP	41
2.9	Time required (in seconds) by the parallel implementation of RQFS	43
2.10	Average time and average accuracy (%) for RAOA, T2D, RAC and GCM datasets using QP and QPS with Nystrom method with Top 50 features.	43
2.11	Best 10CV accuracy (in %) for Colon dataset.	47
2.12	Best 10CV accuracy (in %) for Leukemia dataset.	48
3.1	Datasets: detailed description	66
3.2	Lowest error rates (%)	67
3.3	Number of selected features corresponding to lowest error rates (%)	68
3.4	table	68
3.5	Comparison of average memory requirements(in KB).	68

3.6	Error rates (%) for bioinformatics and vision datasets by each methods	70
3.7	Biological validation of sample clusters for Colon dataset.	74
3.8	Biological validation of sample cluster for Leukemia dataset.	75
3.9	Best 10CV accuracy (in %) for Colon dataset.	77
3.10	Best 10CV accuracy (in %) for Leukemia dataset.	78
4.1	Dataset description	93
4.2	Best Accuracy (in %)	97
4.3	Best 10CV accuracy (in %) for Colon dataset.	102
4.4	Best 10CV accuracy (in %) for Leukemia dataset.	103
5.1	Method notations and their descriptions	119
5.2	Reduction in number of features (NF) without losing on LOOCV accuracy.	121
5.3	Best accuracy (%) in ten runs of LOOCV.	122
5.4	Number of features corresponding to best accuracy in ten runs of LOOCV.	123
5.5	P-values for RPSW method against P, PFS, PMI and PSW methods for ten runs.	123
5.6	Comparison of average execution time during LOOCV.	128
5.7	Time required (in seconds) by the parallel implementation of PP	128
5.8	Parameter values for the models	129
5.9	Comparison with other published methods using 10CV.	133
5.10	Comparison with other published results using LOOCV.	134
6.1	Description of the data sets.	144
6.2	SVM evaluations using product of kernels with l_1 regularization	149
6.3	p-values for FR-GMKL and GMKL methods.	150
6.4	Average accuracy using product of kernels	150

6.5	Average accuracy using sum of kernels	151
6.6	Best 10CV accuracy (in %) for Colon dataset.	154
6.7	Best 10CV accuracy (in %) for Leukemia dataset.	155
6.8	dataset	159
6.9	Average Total RMSE for siRNA-1 dataset	160
6.10	Average Total RMSE for siRNA-2 dataset	160
7.1	Time Complexity of Algorithms	165
7.2	Space Complexity of Algorithms	166
7.3	Execution times by the proposed algorithms.	167
A.1	Annotations of four top genes selected by RQL1 method for Colon dataset.	186
A.2	Annotations of top five genes selected by RQL1 method for Leukemia dataset.	187
A.3	Annotations of top five genes selected by IKAQP method for Colon dataset.	188
A.4	Annotations of top five genes selected by IKAQP method for Leukemia dataset.	189
A.5	Annotations of top five genes selected by MMD method for Colon dataset.	190
A.6	Annotations of top five genes selected by MMD method for Leukemia dataset.	191
A.7	Annotations of five genes selected by RPSW method for colon dataset.	192
A.8	Annotations of five genes selected by RPSW method for Leukemia dataset.	193
A.9	Annotations of top five genes ranked by the weights using FR-GMKL method in Product of Kernel scheme for Colon dataset.	194
A.10	Annotations of top five genes ranked by the weights using FR-GMKL method in Product of Kernel scheme for Leukemia dataset.	195
A.11	List of top five genes for Colon dataset.	196

A.12 List of top five genes for Leukemia dataset.	196
---	-----

List of Abbreviations

ACO	Ant Colony Optimization
BURK/BC	Burkholderia Cepacia GG4
CL	Clustering Large Applications based on Randomized Search (CLARANS)
CL+GO	Clustering Large Applications based on Randomized Search (CLARANS) with Gene Ontology
DCD	Dual Coordinate Decent
DICER	An enzyme name encoded by human DICER1 gene
FCBF	Fast Correlation Based Filter
FCL+GO	Fuzzy Clustering of Large Applications based on Randomized Search (FCLARANS) with Gene Ontology
FGM	Feature Generating Machine
FOL	First Order Logic
FR-GMKL	Fuzzy Rough Set based initialization in GMKL
FRS	Fuzzy Rough Set
FRSVM	Fuzzy Rough Set based SVM
GA	Genetic Algorithm
GDM	Group Discovery Machine
GMKL	Generalized Multiple Kernel Learning
HI	Haemophilus Influenzae
IKAQP	Interleaved K-Means Aggressive Feature Selection
IKQP	Interleaved K-Means Quadratic Programming Feature Selection
KKT	Karush-Kuhn-Tucker

MaxRel	Maximal Relevance
MKL	Multiple Kernel Learning
MLN	Markov Logic Network
MMFS	Max-Margin Feature Selection
mRMR	minimal-Redundancy-Maximal - Relevance
MTLRR	Multi-task Linear Ridge Regression
MTMKL	Multi-task Multiple Kernel Learning
NPSO	Novel Binary PSO
PP	Partition based PSO
PPI	Protein-Protein Interaction
PSO	Particle Swarm Optimization
QP	Quadratic Programming Feature Selection
QP-GMKL	QP based initialization in GMKL
QPN	QP using Nyström Method
QPS	QP method with SMO
QPSF	QP method with SMO on the Fly
QPSFP	QP method with SMO on the Fly using Map-Reduce
QPSN	QP method with SMO using Nyström Method
RISC	RNA-induced silencing complex
RP	Recursive PSO
RQFS	Regularized Quadratic Feature Selection
RQP	Parallel RQFS
SA	simulated annealing
SVM	Support Vector Machine
SVM-RFE	Recursive Feature Elimination using SVM
SVR	Support Vector Regression
TKQP	Two-pass KN-means Feature Selection
α	The feature weight vector (Lagrangian variable)
C	The number of classes
M	The number of Features
N	The number of Samples
Q	The Redundancy Matrix (or Similarity Matrix)
s	The relevance vector
θ	The scaling parameter between redundancy and relevance