# 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

# 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
Deptartment 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 Chaterjee 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 JIIT 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

# List of Figures

# List of Tables

# 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 |
| $\alpha$ | 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 |
| $\theta$ | The scaling parameter between redundancy and relevance |