DEVELOPING EFFICIENT TECHNIQUES FOR FEATURE SELECTION, CLASSIFICATION, SEMI-SUPERVISED CLUSTERING AND THEIR APPLICATIONS

By

MANISH GUPTA

Department of Mathematics

Submitted

in fulfillment of the requirements of the degree

of

Doctor of Philosophy

to the

Indian Institute of Technology, Delhi

New Delhi-110016, India

June, 2010

Certificate

This is to certify that the thesis entitled **Developing Efficient Techniques for Feature Selection, Classification, Semi-Supervised Clustering and Their Applications**, which is being submitted by **Manish Gupta** for the award of the degree of **Doctor of Philosophy in Mathematics** to the **Indian Institute of Technology, Delhi**, is a bona fide research work done under our guidance and supervision.

The thesis has reached the standard fulfilling the requirements of the regulations relating to the degree. The results obtained in the thesis have not been submitted to any other University or Institute for the award of any degree or diploma.

June, 2010

Dr (Mrs.) B.Chandra
Professor
Department of Mathematics
Indian Institute of Technology Delhi
Hauz Khas, New Delhi -110016
India

Dr M. P. Gupta
Professor
Department of Management Studies
Indian Institute of Technology Delhi
Hauz Khas, New Delhi -110016
India

Acknowledgements

I express gratitude to my supervisor Prof. B. Chandra, for her precious guidance and active support during my Ph.D. program. Without her constant encouragement and extraordinary guidance, I could not have finished this thesis. I am grateful for her valuable time which she spent in guiding me. Her valuable advice, discussions, comments and suggestions helped me to complete this thesis. She showed me different ways to approach a research problem. I do not have words to express my feelings and admiration for her inspiration, untiring efforts and guidance.

I thank my co-supervisor Prof. M. P. Gupta, for his support and encouragement. I learnt from him to be persistent to accomplish any goal.

I wish to acknowledge the Indian Institute of Technology, Delhi for providing me with outstanding research facilities. I extend my sincere thanks to my SRC members, Prof. S. G. Deshmukh and Dr. Aparna Mehra for their help and support. I would also like to thank the Head of the Department for all the help and support provided to me.

I would like to thank Sh. HV Srinivasa Rao, Director, ISSA, DRDO for his active support to pursue my research at IIT Delhi. I express sincere thanks to Sh. Sumant Mukherji, Scientist 'E', ISSA, DRDO for providing me sufficient time for my studies.

I am also thankful to Director, National Crime Records Bureau (NCRB) for providing Indian crime data for carrying out research and analysis.

I am extremely thankful to my parents for their blessings. I am forever indebted to my wife Surabhi and my daughter Ashita for their understanding and endless patience.

Above all, I am grateful to God and His blessings to accomplish my objective.

Manish Gupta

Abstract

Data Mining plays a vital role in wide variety of applications. There are broadly three techniques in Data Mining viz. Classification, Clustering and Association Rule Mining. The thesis focuses on developing novel Classification and Clustering techniques. Classification and Clustering is extensively used for solving many practical problems such as pattern recognition, image processing, information retrieval, data segmentation cancer diagnosis and prediction etc. For efficient Classification and Clustering, feature selection is one of the important data pre-processing tasks. If appropriate features are selected, the efficiency of Classification and Clustering can be drastically improved.

The main contributions of this thesis are as follows. An efficient feature selection and ranking algorithm based on statistically defined effective range has been proposed. In the area of Classification, a robust function for estimating probabilities in Naïve Bayes Classifier has been developed. For Semi-Supervised Clustering, a novel neural network approach has been proposed for finding weights for weighted clustering. An efficient similarity measure has also been proposed for multivariate time series (MTS) clustering. A two phase methodology for performance analysis based on MTS clustering using proposed similarity measure and Data Envelopment Analysis model has been proposed.

The effectiveness of the proposed efficient techniques have been illustrated on benchmark datasets from UCI machine learning repository, well known gene expression datasets and on live project on crime data mining funded by National Crime Records Bureau (NCRB). In order to depict the utility of the proposed algorithms in real world applications, architecture of Intelligent Decision Support System using developed efficient techniques has been designed for Indian Police.

Contents