

**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

<b>Certificate.....</b>	<b>i</b>
<b>Acknowledgements.....</b>	<b>ii</b>
<b>Abstract.....</b>	<b>iii</b>
<b>List of Tables .....</b>	<b>ix</b>
<b>List of Figures .....</b>	<b>xi</b>
<b>Chapter 1 Introduction.....</b>	<b>1</b>
1.1 Overview of Classification and Clustering .....	1
1.2 Overview of Feature Selection Methods.....	6
1.3 Overview of Decision Support Systems .....	7
1.4 Motivation.....	8
1.5 Objective.....	10
1.6 Proposed Efficient Techniques: A Brief Overview .....	11
1.7 Thesis Organization .....	14
<b>Chapter 2 A Statistical Approach for Feature Selection and Ranking.....</b>	<b>15</b>
2.1 Introduction.....	15
2.2 Feature Selection Methods: A Brief Survey .....	19
2.2.1 Branch and Bound (B&B) .....	20
2.2.2 SFG (Sequential Forward Generation) .....	20
2.2.3 Relief-F .....	21
2.2.4 Minimum Redundancy-Maximum Relevance (MRMR).....	22
2.3 ERFS (Effective Range based Feature Selection) .....	23
2.3.1 Effective Range ( $R_{ij}$ ) .....	23
2.3.2 ERFS Algorithm .....	24
2.3.3 Theoretical Analysis .....	26
2.3.3.1 Effect of Factor ( $1-p_j$ ) .....	27
2.3.3.2 Hypothesis Testing.....	28
2.3.3.3 Discriminant Rule .....	31
2.3.3.4 Maximum Likelihood Discriminant Rule.....	31
2.3.3.5 Expected Cost of Misclassification (ECM) .....	31

2.4	Results and Discussion .....	34
2.4.1	Results on UCI Datasets .....	34
2.4.2	Results on Gene Expression Datasets .....	37
2.5	Concluding Remarks.....	41
<b>Chapter 3 A Robust Approach for Estimating Probabilities in Naive-Bayes Classifier.....</b>		<b>43</b>
3.1	Introduction.....	44
3.2	Naive-Bayes Classifier (NBC).....	46
3.2.1	Estimation of Probabilities in Naive-Bayes Classifier.....	47
3.2.1.1	Estimate Approach.....	48
3.2.1.2	Estimation of Probabilities for Numeric Feature.....	48
3.3	Robust Naïve Bayes Classifier (R-NBC).....	49
3.3.1	Robust Function ( $f_j$ ) .....	51
3.3.2	Procedure .....	52
3.3.3	Illustration.....	54
3.4	Results and Discussion .....	58
3.4.1	Results on UCI Datasets .....	58
3.4.2	Results on Gene Expression Datasets .....	61
3.5	Simulation Study.....	63
3.6	Concluding Remarks.....	65
<b>Chapter 4 A Novel Approach for Distance-Based Semi-Supervised Clustering using Functional Link Neural Network.....</b>		<b>66</b>
4.1	Introduction.....	67
4.2	Related Work .....	70
4.2.1	K-means Clustering Algorithm.....	70
4.2.2	Semi Supervised Clustering Algorithms.....	71
4.2.2.1	Pairwise Constraint K-Means (PCKMeans) Algorithm .....	72
4.2.2.2	Metric Pairwise Constraint K-Means (MPCK-Means) Algorithm.....	73
4.3	Functional link Neural Network based Clustering Approach (FNNCA).....	75
4.3.1	Parametric Minkowski Model.....	75
4.3.2	Functional Link Neural Network .....	76

4.3.3	Orthonormal Basis .....	77
4.3.3.1	Gram-Schmidt Procedure.....	78
4.3.4	Proposed Methodology (FNNCA).....	78
4.3.4.1	Threshold Fixation .....	78
4.3.4.2	FNNCA Algorithm .....	79
4.3.4.3	Procedure .....	81
4.3.5	Cluster Evaluation Method .....	82
4.4	Results and Discussion .....	82
4.4.1	Comparative Results with Training Data of Fixed Size .....	82
4.4.2	Comparative Results with Training Data of Varying Sizes.....	85
4.5	Application of FNNCA for Identification of Crime Hot Spots on India Crime Data.....	87
4.6	Concluding Remarks.....	90
<b>Chapter 5</b>	<b>An Efficient Similarity Measure based Multivariate Time Series Clustering Approach for Performance Analysis .....</b>	<b>92</b>
5.1	Introduction.....	93
5.2	Existing Similarity Measures for MTS .....	95
5.2.1	Dynamic Time Wrapping (DTW).....	95
5.2.2	Extended Frobenius Norm (Eros).....	97
5.3	Proposed Similarity Measure.....	97
5.3.1	Advantages of the Proposed Similarity Measure.....	101
5.4	Comparative Performance Evaluation of Proposed Similarity Measure .....	101
5.4.1	Datasets used for comparative evaluation.....	101
5.4.2	Comparative Results .....	102
5.5	MTS Clustering with DEA for Measuring Efficiency of Homogeneous Units.....	104
5.5.1	Two Phase Methodology .....	104
5.5.2	Malmquist DEA Model.....	105
5.5.3	Experimental Results .....	106
5.5.3.1	Identification of crime zones by MTS clustering using proposed similarity measure .....	107

5.5.3.2	Performance Analysis using Malmquist DEA model.....	108
5.6	Concluding Remarks.....	115
<b>Chapter 6</b>	<b>A Real World Application: An Intelligent Decision Support System for Indian Police .....</b>	<b>116</b>
6.1	Introduction.....	116
6.2	Indian Police System: A Brief Overview.....	118
6.2.1	Structure of Indian Police .....	118
6.2.2	Role of Indian Police .....	120
6.3	Existing Police Information System .....	121
6.3.1	Crime Criminal Information System (CCIS).....	122
6.3.2	Common Integrated Police Application (CIPA).....	123
6.4	Proposed Architecture of Intelligent Police System (IPS).....	126
6.4.1	User/Decision-Maker .....	127
6.4.2	Crime Analysis Model .....	128
6.4.3	Web Base Management System.....	128
6.4.4	Adaptive Query Interface.....	129
6.4.5	Knowledge Acquisition System.....	129
6.4.6	Data Base Management System for India Crime Database .....	130
6.4.7	Model Base Management System.....	132
6.4.8	Model Base .....	132
6.4.9	Knowledge Base .....	133
6.4.10	Visualization .....	134
6.5	Current Status of IPS Prototype on Indian Crime Records.....	134
6.5.1	Implementation of Crime Analysis Model .....	135
6.5.2	Implementation of Adaptive Query Interface .....	138
6.5.3	Implementation of Data Extraction Module for Indian Crime Database.....	142
6.5.4	Model Base For IPS .....	144
6.5.5	Implementation of Visualization.....	148
6.6	Application of ERFS and FNNCA for Identification of Crime Zones and Crime Hot Spots in IPS.....	149

6.7 Concluding Remarks.....	152
<b>Chapter 7 Conclusions.....</b>	<b>153</b>
<b>References.....</b>	<b>158</b>
<b>Bio-Data.....</b>	<b>170</b>