TABLE OF CONTENTS

Candidate Declaration i
Certificate ii
Acknowledgements iii
List of Publications iv
Abstract v
List of Tables xii
List of Figures xvi

Contents Page No.
Chapter 1: Introduction to Data Mining 1-17

1.1. Overview of Data Mining 1
1.2 Knowledge Discovery from Data 3
1.3 Data Mining Techniques 4
1.3.1 Frequent Pattern Mining and Association Rule Analysis 4
1.3.2 Classification 5
1.3.3 Clustering 6
1.4 Introduction to K-Means Algorithm 8
1.5 K-Modes Algorithm 9
1.6 K-Prototype Algorithm 10
1.7 Challenges 12
1.8 Motivation 14
1.9 Objectives 14
1.10 Methodology 15
1.11 Contribution 16
1.12 Organization of Thesis 16

viii
Chapter 2: Literature Survey

2.1 Initial centroids Selection Sensitivity Problem 18
2.2 Dissimilarity Measure for Categorical and Mixed Data 21
2.3 Providing the value of K in K-Means algorithm for Numerical Datasets 24
2.4 Providing the value of K in K-Means algorithm for Categorical and Mixed Datasets 25
2.5 Efficiency of K-Means Algorithm for High Dimension Datasets 26
2.6 Gaps in Study 28
2.7 Problem Statement 28

Chapter 3: An Extended K-Means Algorithm for Generating Clusters Dynamically 30-54

3.1 Introduction 30
3.2 Existing Methods for Clustering a Dataset without Inputting Number of Clusters 31
3.3 The Proposed K-Means Algorithm 32
3.4 Illustrative Examples 34
   3.4.1 Iris Dataset 34
   3.4.2 Marks Dataset 36
   3.4.3 Ruspini Dataset 39
3.5 Comparative Analysis of the Proposed Algorithm 43
   3.5.1 Results on Iris Dataset 43
   3.5.2 Results on Wine Dataset 47
   3.5.3 Results on Breast Tissue Dataset 50
   3.5.4 Results on Yeast Dataset 52
3.6 Conclusions 54
Chapter 4: An Extended K-Modes Algorithm for Automatic Generation of Clusters

4.1 Introduction 55
4.2 The Proposed K-Modes Algorithm 58
  4.2.1 The Proposed Objective Function 58
  4.2.2 Generation of Initial Clusters 59
  4.2.3 Finding Centroids of the Clusters 62
  4.2.4 Computation of Distance of an Object from its Centroid 62
  4.2.5 The Pseudocode of the Proposed K-Modes Algorithm 63
4.3 Illustrative Examples 64
  4.3.1 Balance Scale Dataset 66
4.4 Comparative Analysis of the Proposed Algorithm 70
  4.4.1 Results on Car Evaluation Dataset 70
  4.4.2 Results on Credit Approval Dataset 72
4.5 The Proposed K-Modes Algorithm for High Dimension Datasets 75
4.6 Illustrative Examples 77
  4.6.1 Academic Dataset 77
4.7 Comparative Analysis of the Proposed Algorithm 82
  4.7.1 Results on Credit Approval Dataset 82
  4.7.2 Results on Zoo Dataset 84
4.8 Conclusions 87

Chapter 5: Dynamic Clustering of Data with Modified K-Prototype Algorithm 88-113

5.1 Introduction 88
5.2 The Proposed K-Prototype Algorithm 89
  5.2.1 The Proposed Objective Function 89
  5.2.2 Generation of Initial Clusters 90
5.2.3 Finding Centroids of the Clusters 93
5.2.4 Computation of Distance of an Object from its Centroid 94
5.2.5 The Pseudocode of the Proposed K-Prototype Algorithm 95
5.3 Illustrative Examples 96
5.3.1 Illustration on Dataset2 97
5.3.2 Illustration on Dataset3 102
5.4 Comparative Analysis of the Proposed Algorithm 104
5.4.1 Results on Lymphography Dataset 105
5.4.2 Results on Post-Operative Patient Dataset 109
5.5 Conclusions 112

Chapter 6: Conclusions and Future Scope 114-117
6.1 Relevance and Conclusions 114
6.2 Scope for Future Work 117
References 118-126

APPENDICES
Appendix A: Snapshots of Application 127-128
Brief Profile of Scholar 129