TABLE OF CONTENTS

ABSTRACT ……..... iv

LIST OF FIGURES ……..... vi
LIST OF TABLES ……..... vii

LIST OF ACRONYMS ……... viii

CHAPTER 1: INTRODUCTION 1 - 20

1.1 Data Clustering ……... 6

1.2 Major Clustering Approaches …… 8

1.2.1 Hierarchical Clustering …… 8

1.2.2 Partitional Clustering …… 9

1.2.3 Density-based Clustering …… 10

1.2.4 Grid-based Clustering …… 10

1.2.5 Model-based Clustering …… 10

1.3 Successful algorithms for data clustering ……..…… 11

1.3.1 Clustering numerical attributes …… 11

1.3.2 Clustering categorical attributes …… 13

1.3.3 Limitations of existing approaches …… 15

1.4 Incremental Clustering ……... 15

1.5 Motivation …….. 17

1.6 Problem Statement ……. 18

1.7 Original contributions of the research work …….. 19

1.8 Organization of the thesis …….. 19

CHAPTER 2: LITERATURE REVIEW 21 - 47

2.1 Incremental clustering algorithms for mining dynamic and large datasets ……... 25

2.2 Incremental clustering of numerical data …….. 27

2.3 Incremental clustering of categorical data …….. 29

2.4 Swarm intelligence based incremental clustering algorithms …… 34

2.5 Incremental clustering of streaming data ……... 36

2.6 Incremental clustering algorithms for other applications ……... 38

2.7 Handling mixed data …….. 42

2.8 Clustering of mixed data points incrementally …….. 46

CHAPTER 3: INCREMENTAL CLUSTERING APPROACH FOR NUMERICAL DATA 48 - 74

3.1 The functionality of CFICA ……... 49

3.2 Initial clustering of the static database …….. 52

3.2.1 K-means clustering algorithm …… 53

3.2.2 Steps of k-means algorithm …….. 54

3.3 Computation of Cluster Feature (CF) …… 55

3.4 Insertion of a new data point ……. 58

3.4.1 Justification for Bias …….. 60

3.4.2 Proposed Distance Metric – Inverse Proximity Estimate (IPE) ……... 62

3.4.3 Extendibility of the cluster boundary towards Sparse area Vs Dense area …… 63

3.5 Incremental Clustering Approach with CFICA ……. 64

3.6 Finding the farthest point in the vicinity of incoming data point … 65

3.7 Finding the appropriate cluster ……..……. 66

3.8 Updating of Cluster Feature ……..……. 66

3.9 Merging of closest cluster pair ……..……. 68

3.10 Need for Cluster Refresh ……..…….. 70

3.11 The Pseudo-code for CFICA ……..…… 71

CHAPTER 4: INCREMENTAL CLUSTERING APPROACH TO MIXED DATA 75 - 106

4.1 Distance Estimation for Mixed type of data …….. 76

4.1.1 Specificity of an Attribute-value pair …… 77

4.1.2 Prominence of attribute-value pairs …….. 79

4.2 Mixed Distance ……..……. 80

4.3 Initial cluster formation of the static database …… 83

4.3.1 Modified k-means algorithm …….. 84

4.4 Clustering of incremental database …… 85

4.5 Computation of Cluster Feature (CF) …… 85

4.6 Proximity Estimation of a new data point ……..…… 86

4.7 Insertion of a new data point ……..…… 87

4.8 Finding p-farthest points of a cluster ...…….. …… 88

4.9 Finding farthest point in the vicinity of Δy …… 89

4.10 Updating of Cluster Feature ……...…… 90

4.11 Merging of closest cluster pair ……..…… 91

4.12 Cluster Refresh ……..... 94

4.13 Hypothetical dataset ……... 95

CHAPTER 5: EXPERIMENTAL ANALYSIS AND RESULTS 107 - 125

5.1 Preprocessing …….. 108

5.1.1 Numerical data ……... 108

5.1.2 Mixed data ……. 109

5.2 Dimensionality reduction ……... 110

5.3 Cluster formation …… 115

5.3.1 Numerical data ……... 115

5.3.2 Mixed data ……. 116

5.4 Metrics in which performance is estimated …… 117

5.5 Performance Evaluation ……. 118

5.5.1 Cluster Purity (CFICA) ……. 118

5.5.2 Execution time (CFICA) …… 121

5.5.3 Cluster Purity (M-CFICA) ……. 122

5.6 System Configuration ……. 125

CHAPTER 6: CONCLUSIONS 126 - 129

6.1 Significance of the Research Work …… 127

6.2 Future Extension ……. 128

REFERENCES …… 130

PUBLICATIONS …… 142