

**A MULTI-OBJECTIVES GENETIC ALGORITHM
CLUSTERING ENSEMBLES BASED
APPROACH TO SUMMARIZE
RELATIONAL DATA**

GABRIEL JONG CHIYE

**FACULTY OF COMPUTING AND INFORMATICS
UNIVERSITI MALAYSIA SABAH
2015**

**A MULTI-OBJECTIVES GENETIC ALGORITHM
CLUSTERING ENSEMBLES BASED
APPROACH TO SUMMARIZE
RELATIONAL DATA**

GABRIEL JONG CHIYE

***SUBMITTED IN PARTIAL FULFILLMENT FOR
THE REQUIREMENTS FOR THE BACHELOR
OF COMPUTER SCIENCE
(SOFTWARE ENGINEERING)***

**FACULTY OF COMPUTING AND INFORMATICS
UNIVERSITI MALAYSIA SABAH
2015**

SUPERVISOR'S CONFIRMATION

NAME : **GABRIEL JONG CHIYE**
Matrik No. : **BK11110089**
TITLE : **A MULTI-OBJECTIVES GENETIC ALGORITHM
CLUSTERING ENSEMBLES BASED APPROACH TO
SUMMARIZE RELATIONAL DATA**
DEGREE : **DEGREE IN BACHELOR OF COMPUTER SCIENCE
(SOFTWARE ENGINEERING)**
VIVA DATE : **29 JUNE 2015**

CERTIFIED BY

1. SUPERVISOR

Associate Professor Dr. Rayner Alfred

Signature

DECLARATION

I hereby declare that the material in this thesis is my own except for quotations, excerpts, equations, summaries and references. All information from these other sources has been duly acknowledged.

June 29, 2015

Gabriel Jong Chiye

BK11110089

ACKNOWLEDGEMENT

First of all, I would like to express my very great appreciation to my main supervisor Associate Professor Dr. Rayner Alfred for provided me with valuable suggestions during the planning of this project, constructive guidance during the development of this project work. His willingness to give his time and share his knowledge so generously has been very much appreciated. I would also like to offer my special thanks to my examiner Dr. Lau Hui Keng and panel Dr. Mohd Hanafi Ahmad Hijazi for their encouragement, value recommendations and advise to keep my progress on schedule.

I would also like to extend my thanks to senior lecturer Dr. Chin Kim On for his help in offering me the research laboratory and necessary resources to run the experiments. My special thanks extended to my doctorate degree and master degree colleagues from the Center of Excellence of Semantic Agents for their professional guidance.

Finally, I wish to thank my parents and brother for their continuous love, support and encouragement throughout my study.

Gabriel Jong Chiye

June 2015

ABSTRACT

K-means algorithm is one of the well-known clustering algorithms that promise to converge to a local optimum in few iterative. However, traditional k-means algorithm is designed to cluster data of single target table. Due to the nature of data collected in real life applications, many data have been collected and stored in relational databases. Traditional clustering and classification learning algorithms cannot be applied directly in learning multi-relational databases. Several approaches have been designed and proposed to learn relational data which includes Inductive Logic Programming based approaches, Graph based approaches, Multi-View approaches and also Dynamic Aggregation of Relational Attributes approach. Dynamic Aggregation of Relational Attributes approach is very effective in learning relational data set. Dynamic Aggregation of Relational Attributes summarizes relational data by clustering records exist in non-target tables. However, the quality of summarization of data depends highly on the position of initial centroids selected. Thus, it may affect the overall classification task. Thus, this project proposes a Genetic Algorithm-based Clustering Ensembles in learning relational datasets by combining the results obtained from several k-means clustering runs with different values of number of clusters, in which the location of centroids are optimal for every sets of clusters. The effects of using different similarity measurements and applying different fitness functions for the genetic algorithm on the predictive accuracies of the classifiers are also studied. Based on the results obtained, it can be concluded that using the consensus result of several clustering results can increase the predictive accuracy of classification task. It can be concluded that the Euclidean distance has better performance on mutagenesis datasets and cosine similarity has better performance on hepatitis datasets when evaluated with Weka C4.5 classifier, but the other way round when Naïve Bayes classifier is used for evaluation.

ABSTRAK

Bersama-samaan: Kekelompokan Berdasarkan Susunan

Algoritma k-means adalah satu algoritma pengelompokan yang berjanji untuk menghasilkan satu optima tempatan dalam beberapa lalaran. Walau bagaimanapun, algoritma k-means yang tradisional telah direka untuk mengelompok data jadual sasaran tunggal. Oleh kerana aplikasi kehidupan sebenar, kita mengumpul dan menyimpan data di dalam pangkalan data berhubung. Algoritma pembelajaran secara pengelompokan dan klasifikasi tidak boleh diguna secara langsung dengan pangkalan data berhubung. Beberapa cara telah direka dan dicadangkan untuk belajar data berhubung, termasuk Inductive Logic Programming, Graph, Multi-View, dan Dynamic Aggregation of Relational Attributes. Dynamic Aggregation of Relational Attributes merupakan cara yang berkesan untuk belajar data berhubung. Dynamic Aggregation of Relational Attributes meringkaskan data berhubung dengan mengelompok rekod dalam jadual bukan sasaran. Walau bagaimanapun, kualiti ringkasan data sangat bergantung kepada kedudukan pusat awal yang dipilih. Dengan itu, ia boleh menjejaskan kerja keseluruhan klasifikasi. Oleh itu, kerja ini mencadangkan satu cara pengelompokan ensemble berdasarkan algoritma genetik bagi belajar data berhubung dengan menggabungkan keputusan yang didapat dari beberapa larian pengelompokan k-mean dengan bilangan kelompokan yang berbeza, di mana lokasi pusat adalah optima untuk setiap pengelompokan. Kesan ketepatan ramalan oleh pengelasan bagi mengguna pengukuran jarak yang berbeza dan mengguna fungsi kecergasan dalam algoritma genetik juga telah dikaji. Berdasarkan keputusan kajian yang telah diperolehi, ia boleh disimpulkan bahawa keputusan konsensus boleh meningkat ketepatan ramalan oleh tugas pengelasan. Ia juga boleh dikatakan bahawa jarak Euclidean boleh memperoleh prestasi yang lebih tinggi dalam data mutagenesis, persamaan kosinus memperoleh prestasi yang lebih tinggi dalam data hepatitis semasa dinilai dengan pengelasan Weka C4.5, tetapi sebaliknya ketika dinilai dengan pengelasan Naïve Bayes.

TABLE OF CONTENTS

Supervisor’s Confirmation	ii
Declaration.....	iii
Acknowledgement.....	iv
Abstract	v
Abstrak	vi
List of Tables.....	xii
List of Figures.....	xviii
List of Code Listing	xxxi
List of Equations	xxxii
List of Abbreviations.....	xxxiii
CHAPTER 1 Introduction	1
1.1 Problem Background	1
1.2 Problem Statement	4
1.3 Objectives	5
1.4 Project Scope	5
1.5 Organization of Report	7
1.6 Conclusion.....	7
CHAPTER 2 Literature Review	8
2.1 Data Mining and Knowledge Discovery.....	8
2.2 Multi-Relational Data Mining.....	9
2.3 Term Frequency-Inverse Document Frequency	11
2.4 Clustering.....	12
2.4.1 Hierarchy Clustering.....	12

2.4.2	Partition Clustering	13
2.5	Clustering Validity	14
2.5.1	Sum Of Squared Error	14
2.5.2	Davies-Bouldin Index	14
2.5.3	Cluster Entropy.....	15
2.6	Evolutionary Algorithms.....	15
2.6.1	Genetic Algorithm.....	16
2.6.2	Elitism	18
2.7	Multi-Objectives Optimization.....	18
2.8	Ensemble	20
2.8.1	Bagging	20
2.8.2	Boosting	21
2.8.3	Stacking.....	21
2.9	Clustering Ensemble.....	22
2.10	Conclusion.....	23
	CHAPTER 3 Methodology.....	24
3.1	Base learner data.....	25
3.2	K-means Clustering	26
3.3	Genetic Algorithms Hybridization	29
3.4	Multi-Objectives Fitness.....	36
3.5	Stacking Ensembles.....	38
	CHAPTER 4 Experimental Setup	40
4.1	Base Learner	40
4.2	Ensembles Learner.....	41
4.3	Genetic Algorithm	41

CHAPTER 5 Experimental results	43
5.1 Predictive Accuracy generated by Weka C4.5 Classifier on a Summarized Mutagenesis Datasets with Euclidean Distance as Distance Measurement.....	44
5.1.1 Results when using Sum of Squared Error as Fitness Function.....	44
5.1.2 Results when using Davies-Bouldin Index as Fitness Function.....	46
5.1.3 Results when using Cluster Entropy as Fitness Function	49
5.1.4 Results when using Multi-Objectives as Fitness Function	51
5.2 Predictive Accuracy generated by Weka C4.5 Classifier on a Summarized Mutagenesis Datasets with Cosine Similarity as Distance Measurement.....	54
5.2.1 Results when using Sum of Squared Error as Fitness Function.....	54
5.2.2 Results when using Davies-Bouldin Index as Fitness Function.....	56
5.2.3 Results when using Cluster Entropy as Fitness Function	59
5.2.4 Results when using Multi-Objectives as Fitness Function	61
5.3 Predictive Accuracy generated by Weka C4.5 Classifier on a Summarized Hepatitis Datasets with Euclidean Distance as Distance Measurement.....	64
5.3.1 Results when using Sum of Squared Error as Fitness Function.....	64
5.3.2 Results when using Davies-Bouldin Index as Fitness Function.....	66
5.3.3 Results when using Cluster Entropy as Fitness Function	68
5.3.4 Results when using Multi-Objectives as Fitness Function	70
5.4 Predictive Accuracy generated by Weka C4.5 Classifier on a Summarized Hepatitis Datasets with Cosine Similarity as Distance Measurement.....	73
5.4.1 Results when using Sum of Squared Error as Fitness Function.....	73
5.4.2 Results when using Davies-Bouldin Index as Fitness Function.....	75

5.4.3	Results when using Entropy as Fitness Function	77
5.4.4	Results when using Multi-Objectives as Fitness Function	79
5.5	Predictive Accuracy generated by Weka Naïve Bayes Classifier on a Summarized Mutagenesis Datasets with Euclidean Distance as Distance Measurement.....	82
5.5.1	Results when using Sum of Squared Error as Fitness Function.....	82
5.5.2	Results when using Davies-Bouldin Index as Fitness Function.....	84
5.5.3	Results when using Cluster Entropy as Fitness Function	87
5.5.4	Results when using Multi-Objectives as Fitness Function	89
5.6	Predictive Accuracy generated by Weka Naïve Bayes Classifier on a Summarized Mutagenesis Datasets with Cosine Similarity as Distance Measurement.....	92
5.6.1	Results when using Sum of Squared Error as Fitness Function.....	92
5.6.2	Results when using Davies-Bouldin Index as Fitness Function.....	94
5.6.3	Results when using Cluster Entropy as Fitness Function	97
5.6.4	Results when using Multi-Objectives as Fitness Function	99
5.7	Predictive Accuracy generated by Weka Naïve Bayes Classifier on a Summarized Hepatitis Datasets with Euclidean Distance as Distance Measurement.....	102
5.7.1	Results when using Sum of Squared Error as Fitness Function.....	102
5.7.2	Results when using Davies-Bouldin Index as Fitness Function.....	104
5.7.3	Results when using Cluster Entropy as Fitness Function	107
5.7.4	Results when using Multi-Objectives as Fitness Function	109
5.8	Predictive Accuracy generated by Weka Naïve Bayes Classifier on a Summarized Hepatitis Datasets with Cosine Similarity as Distance Measurement.....	112
5.8.1	Results when using Sum of Squared Error as Fitness Function.....	112

5.8.2 Results when using Davies-Bouldin Index as Fitness Function.....	114
5.8.3 Results when using Cluster Entropy as Fitness Function	116
5.8.4 Results when using Multi-Objectives as Fitness Function	118
5.9 Summary of Results	120
5.10 Comparison of Average Predictive Accuracy between the use of Unsummarized, Summarized with Clustering and Summarized with Clustering Ensembles for Classification Task.....	125
CHAPTER 6 Conclusion	128
6.1 Conclusion.....	128
6.2 Limitations and Future Works.....	128
6.2.1 K-means Clustering Algorithm and Non-spherical Data.....	128
6.2.2 Tuning of Weight for Individual Functions in Multi-Objectives Optimization.....	129
6.2.3 Integrate Features Selection into Summarization Stage.....	129
Bibliography	131

LIST OF TABLES

Table 5.1:	Comparison of predictive accuracy generated by Weka C4.5 classifier on mutagenesis dataset after summarization without clustering ensembles and with clustering ensembles using Euclidean distance as distance measurement and sum of squared error as fitness function in genetic algorithm 44
Table 5.2:	Comparison of predictive accuracy generated by Weka C4.5 classifier on mutagenesis dataset after summarization without clustering ensembles and with clustering ensembles using Euclidean distance as distance measurement and Davies-Bouldin index as fitness function in genetic algorithm 46
Table 5.3:	Comparison of predictive accuracy generated by Weka C4.5 classifier on mutagenesis dataset after summarization without clustering ensembles and with clustering ensembles using Euclidean distance as distance measurement and cluster entropy as fitness function in genetic algorithm..... 49
Table 5.4:	Comparison of predictive accuracy generated by Weka C4.5 classifier on mutagenesis dataset after summarization without clustering ensembles and with clustering ensembles using Euclidean distance as distance measurement and multi-objectives as fitness function in genetic algorithm 51
Table 5.5:	Comparison of predictive accuracy generated by Weka C4.5 classifier on mutagenesis dataset after summarization without clustering ensembles and with clustering ensembles using cosine similarity as distance measurement and sum of squared error as fitness function in genetic algorithm..... 54
Table 5.6:	Comparison of predictive accuracy generated by Weka C4.5 classifier on mutagenesis dataset after summarization without clustering ensembles and with clustering ensembles using cosine similarity as distance measurement and Davies-Bouldin index as fitness function in genetic algorithm. 56

Table 5.7:	Comparison of predictive accuracy generated by Weka C4.5 classifier on mutagenesis dataset after summarization without clustering ensembles and with clustering ensembles using cosine similarity as distance measurement and cluster entropy as fitness function in genetic algorithm.....	59
Table 5.8:	Comparison of predictive accuracy generated by Weka C4.5 classifier on mutagenesis dataset after summarization without clustering ensembles and with clustering ensembles using cosine similarity as distance measurement and multi-objectives as fitness function in genetic algorithm.....	61
Table 5.9:	Comparison of predictive accuracy generated by Weka C4.5 classifier on hepatitis dataset after summarization without clustering ensembles and with clustering ensembles using Euclidean distance as distance measurement and sum of squared error as fitness function in genetic algorithm	64
Table 5.10:	Comparison of predictive accuracy generated by Weka C4.5 classifier on hepatitis dataset after summarization without clustering ensembles and with clustering ensembles using Euclidean distance as distance measurement and Davies-Bouldin index as fitness function in genetic algorithm	66
Table 5.11:	Comparison of predictive accuracy generated by Weka C4.5 classifier on hepatitis dataset after summarization without clustering ensembles and with clustering ensembles using Euclidean distance as distance measurement and cluster entropy as fitness function in genetic algorithm.....	68
Table 5.12:	Comparison of predictive accuracy generated by Weka C4.5 classifier on hepatitis dataset after summarization without clustering ensembles and with clustering ensembles using Euclidean distance as distance measurement and multi-objectives as fitness function in genetic algorithm	70
Table 5.13:	Comparison of predictive accuracy generated by Weka C4.5 classifier on hepatitis dataset after summarization without	

	clustering ensembles and with clustering ensembles using cosine similarity as distance measurement and sum of squared error as fitness function in genetic algorithm.....	73
Table 5.14:	Comparison of predictive accuracy generated by Weka C4.5 classifier on hepatitis dataset after summarization without clustering ensembles and with clustering ensembles using cosine similarity as distance measurement and Davies-Bouldin index as fitness function in genetic algorithm.....	75
Table 5.15:	Comparison of predictive accuracy generated by Weka C4.5 classifier on hepatitis dataset after summarization without clustering ensembles and with clustering ensembles using cosine similarity as distance measurement and cluster entropy as fitness function in genetic algorithm.....	77
Table 5.16:	Comparison of predictive accuracy generated by Weka C4.5 classifier on hepatitis dataset after summarization without clustering ensembles and with clustering ensembles using cosine similarity as distance measurement and multi-objectives as fitness function in genetic algorithm.....	79
Table 5.17:	Comparison of predictive accuracy generated by Weka Naïve Bayes classifier on mutagenesis dataset after summarization without clustering ensembles and with clustering ensembles using euclidean distance as distance measurement and sum of squared error as fitness function in genetic algorithm.	82
Table 5.18:	Comparison of predictive accuracy generated by Weka Naïve Bayes classifier on mutagenesis dataset after summarization without clustering ensembles and with clustering ensembles using euclidean distance as distance measurement and Davies-Bouldin index as fitness function in genetic algorithm.	84
Table 5.19:	Comparison of predictive accuracy generated by Weka Naïve Bayes classifier on mutagenesis dataset after summarization without clustering ensembles and with clustering ensembles	

	using euclidean distance as distance measurement cluster entropy as fitness function in genetic algorithm.....	87
Table 5.20:	Minimum and maximum predictive accuracy and their respective number of clusters using Mutagenesis dataset and multi-objectives as fitness function in genetic algorithms using Euclidean distance as distance measurement.	89
Table 5.21:	Comparison of predictive accuracy generated by Weka Naïve Bayes classifier on mutagenesis dataset after summarization without clustering ensembles and with clustering ensembles using cosine similarity as distance measurement and sum of squared error as fitness function in genetic algorithm.	92
Table 5.22:	Comparison of predictive accuracy generated by Weka Naïve Bayes classifier on mutagenesis dataset after summarization without clustering ensembles and with clustering ensembles using cosine similarity as distance measurement and Davies-Bouldin index as fitness function in genetic algorithm.	94
Table 5.23:	Comparison of predictive accuracy generated by Weka Naïve Bayes classifier on mutagenesis dataset after summarization without clustering ensembles and with clustering ensembles using cosine similarity as distance measurement and cluster entropy as fitness function in genetic algorithm.....	97
Table 5.24:	Comparison of predictive accuracy generated by Weka Naïve Bayes classifier on mutagenesis dataset after summarization without clustering ensembles and with clustering ensembles using cosine similarity as distance measurement and multi-objectives as fitness function in genetic algorithm.	99
Table 5.25:	Comparison of predictive accuracy generated by Weka Naïve Bayes classifier on hepatitis dataset after summarization without clustering ensembles and with clustering ensembles using Euclidean distance as distance measurement and sum of squared error as fitness function in genetic algorithm.	102

Table 5.26:	Comparison of predictive accuracy generated by Weka Naïve Bayes classifier on hepatitis dataset after summarization without clustering ensembles and with clustering ensembles using Euclidean distance as distance measurement and Davies-Bouldin index as fitness function in genetic algorithm.	104
Table 5.27:	Comparison of predictive accuracy generated by Weka Naïve Bayes classifier on hepatitis dataset after summarization without clustering ensembles and with clustering ensembles using Euclidean distance as distance measurement and cluster entropy as fitness function in genetic algorithm.....	107
Table 5.28:	Comparison of predictive accuracy generated by Weka Naïve Bayes classifier on hepatitis dataset after summarization without clustering ensembles and with clustering ensembles using Euclidean distance as distance measurement and multi-objectives as fitness function in genetic algorithm	109
Table 5.29:	Comparison of predictive accuracy generated by Weka Naïve Bayes classifier on hepatitis dataset after summarization without clustering ensembles and with clustering ensembles using cosine similarity as distance measurement and sum of squared error as fitness function in genetic algorithm	112
Table 5.30:	Comparison of predictive accuracy generated by Weka Naïve Bayes classifier on hepatitis dataset after summarization without clustering ensembles and with clustering ensembles using cosine similarity as distance measurement and Davies-Bouldin index as fitness function in genetic algorithm.	114
Table 5.31:	Comparison of predictive accuracy generated by Weka Naïve Bayes classifier on hepatitis dataset after summarization without clustering ensembles and with clustering ensembles using cosine similarity as distance measurement and cluster entropy as fitness function in genetic algorithm.....	116
Table 5.32:	Comparison of predictive accuracy generated by Weka Naïve Bayes classifier on hepatitis dataset after summarization	

	without clustering ensembles and with clustering ensembles using cosine similarity as distance measurement and multi-objectives as fitness function in genetic algorithm.	118
Table 5.33:	Increment of average predictive accuracy in percentage generated by different from classifiers using summarized datasets with clustering ensembles of different combination of experimental settings.....	121
Table 5.34:	Number of cases of improved, unchanged or worst average predictive accuracy with respect to proximity type for clustering ensembles	122
Table 5.35:	Number of cases of improved, unchanged or worst average predictive accuracy with respect to different classifier used	123
Table 5.36:	Number of cases of improved, unchanged or worst average predictive accuracy with respect to different fitness function used in the genetic algorithm	123
Table 5.37:	Number of cases of improved, unchanged or worst average predictive accuracy with respect to different dataset used in the experiment	124
Table 5.38:	Comparison of average predictive accuracy produced by Weka C4.5 Classifier in term of different types of data produced at different stage.....	125
Table 5.39:	Comparison of average predictive accuracy produced by Weka Naïve-Bayes Classifier in term of different types of data produced at different stage	127

LIST OF FIGURES

Figure 2.1:	Single point crossover example in Genetic Algorithms.	18
Figure 3.1:	The proposed genetic algorithm-based stacking-based clustering ensembles overall framework illustration to summarize relational data.	25
Figure 3.2:	Sample of chromosome design in the genetic algorithms.	29
Figure 5.1:	Graph of predictive accuracy generated by Weka C4.5 classifier against number of clusters for B1 dataset when using sum of squared error as fitness function in genetic algorithm using Euclidean distance as distance measurement.	45
Figure 5.2:	Graph of predictive accuracy generated by Weka C4.5 classifier against number of clusters for B2 dataset when using sum of squared error as fitness function in genetic algorithms using Euclidean distance as distance measurement.	45
Figure 5.3:	Graph of predictive accuracy generated by Weka C4.5 classifier against number of clusters for B3 dataset when using sum of squared error as fitness function in genetic algorithms using Euclidean distance as distance measurement.	46
Figure 5.4:	Graph of predictive accuracy generated by Weka C4.5 classifier against number of clusters for B1 dataset when using Davies-Bouldin index as fitness function in genetic algorithm using Euclidean distance as distance measurement.	47
Figure 5.5:	Graph of predictive accuracy generated by Weka C4.5 classifier against number of clusters for B2 dataset when using Davies-Bouldin index as fitness function in genetic algorithm using Euclidean distance as distance measurement.	48
Figure 5.6:	Graph of predictive accuracy generated by Weka C4.5 classifier against number of clusters for B3 dataset when using Davies-Bouldin index as fitness function in genetic algorithm using Euclidean distance as distance measurement.	48

Figure 5.7:	Graph of predictive accuracy generated by Weka C4.5 classifier against number of clusters for B1 dataset when using cluster entropy as fitness function in genetic algorithm using Euclidean distance as distance measurement.	50
Figure 5.8:	Graph of predictive accuracy generated by Weka C4.5 classifier against number of clusters for B2 dataset when using cluster entropy as fitness function in genetic algorithm using Euclidean distance as distance measurement.	50
Figure 5.9:	Graph of predictive accuracy generated by Weka C4.5 classifier against number of clusters for B3 dataset when using cluster entropy as fitness function in genetic algorithm using Euclidean distance as distance measurement.	51
Figure 5.10:	Graph of predictive accuracy generated by Weka C4.5 classifier against number of clusters for B1 dataset when using multi-objectives as fitness function in genetic algorithm using Euclidean distance as distance measurement.	52
Figure 5.11:	Graph of predictive accuracy generated by Weka C4.5 classifier against number of clusters for B2 dataset when using multi-objectives as fitness function in genetic algorithm using Euclidean distance as distance measurement.	53
Figure 5.12:	Graph of predictive accuracy generated by Weka C4.5 classifier against number of clusters for B3 dataset when using multi-objectives as fitness function in genetic algorithm using Euclidean distance as distance measurement.	53
Figure 5.13:	Graph of predictive accuracy generated by Weka C4.5 classifier against number of clusters for B1 dataset when using sum of squared error as fitness function in genetic algorithm using cosine similarity as distance measurement.	55
Figure 5.14:	Graph of predictive accuracy generated by Weka C4.5 classifier against number of clusters for B2 dataset when using sum of squared error as fitness function in genetic algorithm using cosine similarity as distance measurement.	55

Figure 5.15:	Graph of predictive accuracy generated by Weka C4.5 classifier against number of clusters for B3 dataset when using sum of squared error as fitness function in genetic algorithm using cosine similarity as distance measurement.	56
Figure 5.16:	Graph of predictive accuracy generated by Weka C4.5 classifier against number of clusters for B1 dataset when using Davies-Bouldin index as fitness function in genetic algorithm using cosine similarity as distance measurement.	57
Figure 5.17:	Graph of predictive accuracy generated by Weka C4.5 classifier against number of clusters for B2 dataset when using Davies-Bouldin index as fitness function in genetic algorithm using cosine similarity as distance measurement.	58
Figure 5.18:	Graph of predictive accuracy generated by Weka C4.5 classifier against number of clusters for B3 dataset when using Davies-Bouldin index as fitness function in genetic algorithm using cosine similarity as distance measurement.	58
Figure 5.19:	Graph of predictive accuracy generated by Weka C4.5 classifier against number of clusters for B1 dataset when using cluster entropy index as fitness function in genetic algorithms using cosine similarity as distance measurement.	60
Figure 5.20:	Graph of predictive accuracy generated by Weka C4.5 classifier against number of clusters for B2 dataset when using cluster entropy index as fitness function in genetic algorithms using cosine similarity as distance measurement.	60
Figure 5.21:	Graph of predictive accuracy generated by Weka C4.5 classifier against number of clusters for B3 dataset when using cluster entropy index as fitness function in genetic algorithms using cosine similarity as distance measurement.	61
Figure 5.22:	Graph of predictive accuracy generated by Weka C4.5 classifier against number of clusters for B1 dataset when using multi-objectives as fitness function in genetic algorithms using cosine similarity as distance measurement.	62

Figure 5.23:	Graph of predictive accuracy generated by Weka C4.5 classifier against number of clusters for B2 dataset when using multi-objectives as fitness function in genetic algorithms using cosine similarity as distance measurement.	63
Figure 5.24:	Graph of predictive accuracy generated by Weka C4.5 classifier against number of clusters for B3 dataset when using multi-objectives as fitness function in genetic algorithms using cosine similarity as distance measurement.	63
Figure 5.25:	Graph of predictive accuracy generated by Weka C4.5 classifier against number of clusters for H1 dataset when using sum of squared error as fitness function in genetic algorithm using Euclidean distance as distance measurement.	65
Figure 5.26:	Graph of predictive accuracy generated by Weka C4.5 classifier against number of clusters for H2 dataset when using sum of squared error as fitness function in genetic algorithm using Euclidean distance as distance measurement.	65
Figure 5.27:	Graph of predictive accuracy generated by Weka C4.5 classifier against number of clusters for H3 dataset when using sum of squared error as fitness function in genetic algorithm using Euclidean distance as distance measurement.	66
Figure 5.28:	Graph of predictive accuracy generated by Weka C4.5 classifier against number of clusters for H1 dataset when using Davies-Bouldin index as fitness function in genetic algorithm using Euclidean distance as distance measurement.	67
Figure 5.29:	Graph of predictive accuracy generated by Weka C4.5 classifier against number of clusters for H2 dataset when using sum of squared error as fitness function in genetic algorithm using Euclidean distance as distance measurement.	67
Figure 5.30:	Graph of predictive accuracy generated by Weka C4.5 classifier against number of clusters for H3 dataset when using sum of squared error as fitness function in genetic algorithm using Euclidean distance as distance measurement.	68

Figure 5.31:	Graph of predictive accuracy generated by Weka C4.5 classifier against number of clusters for H1 dataset when using cluster entropy as fitness function in genetic algorithm using Euclidean distance as distance measurement.	69
Figure 5.32:	Graph of predictive accuracy generated by Weka C4.5 classifier against number of clusters for H2 dataset when using cluster entropy as fitness function in genetic algorithm using Euclidean distance as distance measurement.	69
Figure 5.33:	Graph of predictive accuracy generated by Weka C4.5 classifier against number of clusters for H3 dataset when using cluster entropy as fitness function in genetic algorithm using Euclidean distance as distance measurement.	70
Figure 5.34:	Graph of predictive accuracy generated by Weka C4.5 classifier against number of clusters for H1 dataset when using multi-objectives as fitness function in genetic algorithm using Euclidean distance as distance measurement.	71
Figure 5.35:	Graph of predictive accuracy generated by Weka C4.5 classifier against number of clusters for H2 dataset when using multi-objectives as fitness function in genetic algorithm using Euclidean distance as distance measurement.	71
Figure 5.36:	Graph of predictive accuracy generated by Weka C4.5 classifier against number of clusters for H3 dataset when using multi-objectives as fitness function in genetic algorithm using Euclidean distance as distance measurement.	72
Figure 5.37:	Graph of predictive accuracy generated by Weka C4.5 classifier against number of clusters for H1 dataset when using sum of squared error as fitness function in genetic algorithms using cosine similarity as distance measurement.	74
Figure 5.38:	Graph of predictive accuracy generated by Weka C4.5 classifier against number of clusters for H2 dataset when using sum of squared error as fitness function in genetic algorithms using cosine similarity as distance measurement.	74

Figure 5.39:	Graph of predictive accuracy generated by Weka C4.5 classifier against number of clusters for H3 dataset when using sum of squared error as fitness function in genetic algorithms using cosine similarity as distance measurement.	75
Figure 5.40:	Graph of predictive accuracy generated by Weka C4.5 classifier against number of clusters for H1 dataset when using Davies-Bouldin index as fitness function in genetic algorithms using cosine similarity as distance measurement.	76
Figure 5.41:	Graph of predictive accuracy generated by Weka C4.5 classifier against number of clusters for H2 dataset when using Davies-Bouldin index as fitness function in genetic algorithms using cosine similarity as distance measurement.	76
Figure 5.42:	Graph of predictive accuracy generated by Weka C4.5 classifier against number of clusters for H3 dataset when using Davies-Bouldin index as fitness function in genetic algorithms using cosine similarity as distance measurement.	77
Figure 5.43:	Graph of predictive accuracy generated by Weka C4.5 classifier against number of clusters for H1 dataset when using cluster entropy as fitness function in genetic algorithms using cosine similarity as distance measurement.	78
Figure 5.44:	Graph of predictive accuracy generated by Weka C4.5 classifier against number of clusters for H2 dataset when using cluster entropy as fitness function in genetic algorithms using cosine similarity as distance measurement.	78
Figure 5.45:	Graph of predictive accuracy generated by Weka C4.5 classifier against number of clusters for H3 dataset when using cluster entropy as fitness function in genetic algorithms using cosine similarity as distance measurement.	79
Figure 5.46:	Graph of predictive accuracy generated by Weka C4.5 classifier against number of clusters for H1 dataset when using multi-objectives as fitness function in genetic algorithms using cosine similarity as distance measurement.	80

BIBLIOGRAPHY

- [1] R. Cattral, F. Oppacher, and K.J.L. Graham, "Techniques for evolutionary rule discovery in data mining," in *Evolutionary Computation, 2009. CEC '09. IEEE Congress*, Trondheim, May 2009, pp. 1737-1744.
- [2] Lei Xu, Chunxiao Jiang, Jian Wang, Jian Yuan, and Yong Ren, "Information Security in Big Data: Privacy and Data Mining," in *Access, IEEE*, 2014, pp. 1149-1176.
- [3] Arno J. Knobbe, Marc de Haas, and Arno Siebes, "Principles of Data Mining and Knowledge Discovery," in *Principles of Data Mining and Knowledge Discovery*.: Springer Berlin Heidelberg, 2001, pp. 277-288.
- [4] Jiawei Han, "Data Mining," in *Encyclopedia of Creativity, Invention, Innovation and Entrepreneurship (2013)*., pp. 595-597.
- [5] Dipankar Dutta, Paramartha Dutta, and Jaya Sil, "Data Clustering with Mixed Features by Multi Objective Generic Algorithm," in *12th International Conference on Hybrid Intelligent Systems*, Pune, 2012, pp. 336-341.
- [6] Saso Dzeroski, "Relational Data Mining," in *Data Mining and Knowledge Discovery Handbook*, Oded Maimon and Lior Rokach, Eds.: Springer US, 2010, pp. 887-911.
- [7] Ping Ling and Xiangsheng Rong, "Double-Phase Locality Sensitive Hashing of neighborhood development for multi-relational data," *Computational Intelligence (UKCI), 2013 13th UK Workshop*, pp. 206-213, September 2013.
- [8] Urvashi Mistry and Amit R Thakkar, "Link-based classification for Multi-Relational database," *Recent Advances and Innovations in Engineering (ICRAIE), 2014*, pp. 1-6, May 2014.

- [9] Wei Zhang, "Multi-Relational Data Mining Based on Higher-Order Inductive Logic," in *Intelligent Systems, 2009. GCIS '09. WRI Global Congress*, Xiamen, 2009, pp. 453-458.
- [10] Jingfeng Guo, Lizhen Zheng, and Tieying Li, "An Efficient Graph-based Multi-relational Data Mining Algorithm," in *2007 International Conference on Computational Intelligence and Security*, Harbin, 2007, pp. 176-180.
- [11] Hongyu Guo and Herna L. Viktor, "Mining Relational Databases with Multi-view Learning," in *Proceedings of the 4th international workshop on Multi-relational mining*, New York, 2005, pp. 15-24.
- [12] Dan Roth and Wen-tau Yih, "Propositionalization of Relational Learning: An Information Extraction Case Study," in *Seventeenth International Joint Conference on Artificial Intelligence*, Seattle, 2001.
- [13] Chung Seng Kheau, Rayner Alfred, and Lau Hui Keng, "Dimensionality Reduction in Data Summarization Approach to Learning Relational Data," in *Intelligent Information and Database Systems Lecture Notes in Computer Science*. Berlin, Germany: Springer, 2013, pp. 166-175.
- [14] Rayner Alfred, "Summarizing Relational Data using Semi-Supervised Genetic Algorithm-based Clustering Techniques," *Journal of Computer*, vol. 6, no. 7, pp. 775-784, 2010.
- [15] Rao M. Kotamarti, Douglas W. Raiford, Michael L. Raymer, and Margaret H. Dunham¹, "A Data Mining Approach to Predicting Phylum for Microbial Organisms using Genome-Wide Sequence Data," in *Bioinformatics and BioEngineering, 2009. BIBE '09. Ninth IEEE International Conference*, Taichung, June 2009, pp. 161-167.
- [16] W. Segretier, M. Clergue, M. Collard, and L. Izquierdo, "An evolutionary data mining approach on hydrological data with classifier juries," in *Evolutionary*

Computation (CEC), 2012 IEEE Congress, Brisbane, June 2012, pp. 1-8.

- [17] Ming Xue and Changjun Zhu, "The Application of Data Mining In the Decision of Supermarket Extension and Businesses Expansion Based on Evolutionary Computation," in *Circuits, Communications and Systems, 2009. PACCS '09. Pacific-Asia Conference*, Chengdu, May 2009, pp. 778-780.
- [18] Bin Lu and Fangyuan Ju, "An optimized genetic K-means clustering algorithm," in *International Conference on Computer Science and Information Processing*, Xi'an, 2012, pp. 1296-1299.
- [19] Taoying Li and Yan Chen, "A Weight Entropy k-means Algorithm for Clustering Dataset with Mixed Numeric and Categorical Data," in *Fifth International Conference on Fuzzy Systems and Knowledge Discovery*, Shandong, 2008, pp. 36-41.
- [20] David Pettinger and Giuseppe Di Fatta, "Space Partitioning for Scalable K-Means," in *Machine Learning and Applications (ICMLA), 2010 Ninth International Conference*, Washington, December 2010, pp. 319-324.
- [21] Xiaohui Cui, Thomas E. Potok, and Paul Palathingal, "Document Clustering using Particle Swarm Optimization," in *Swarm Intelligence Symposium, 2005. SIS 2005. Proceedings 2005 IEEE*, 2005, pp. 185-191.
- [22] R.F. Abdel-Kader, "Genetically Improved PSO Algorithm for Efficient Data Clustering," in *Machine Learning and Computing (ICMLC), 2010 Second International Conference*, Bangalore, February 2010, pp. 71-75.
- [23] Nisha M. N., Mohanavalli S., and Swathika R., "Improving the quality of Clustering using Cluster Ensembles," in *Proceedings of 2013 IEEE Conference on Information and Communication Technologies*, JeJu Island, 2013, pp. 88-92.

- [24] Tan Pang-Ning, Michael Steinbach, and Vipin Kumar, *Introduction to Data Mining*. Boston, United States: Pearson, 2006.
- [25] Soumen Chakrabarti et al., *Data Mining Know It All*. Burlington, United States: Morgan Kaufmann, 2009.
- [26] Ricardo Baeza-Yates and Berthier Ribeiro-Neto, *Modern Information Retrieval: the concepts and technology behind search*. Harlow, England: Pearson Education Limited, 2011.
- [27] Nileshkumar D. Bharwad and Mukesh M. Goswami, "Proposed efficient approach for classification for multi-relational data mining using Bayesian Belief Network," in *Green Computing Communication and Electrical Engineering (ICGCCCEE), 2014 International Conference*, Coimbatore, 2014, pp. 1-4.
- [28] Stephen Muggleton, "Inductive Logic Programming," *New Generation Computing*, vol. 8, no. 4, pp. 295-318, 1991.
- [29] Lawrence B. Holder and Diane J. Cook, "Graph-Based Relational Mining: Current and Future Directions," *ACM SIGKDD Explorations Newsletter*, vol. 5, no. 1, pp. 90-93, July 2003.
- [30] Rayner Alfred, "The Study of Dynamic Aggregation of Relational Attributes on Relational Data Mining," in *Third International Conference, ADMA*, Harbin, 2007, pp. 214-226.
- [31] Florence Sia, Rayner Alfred, Leau Yu Beng, and Tan Soo Fun, "A Variable Length Feature Construction method for data summarization using DARA," in *Computing and Convergence Technology (ICCCCT), 2012 7th International Conference*, Seoul, 2012, pp. 881-887.
- [32] Neepa Shah and Sunita Mahajan, "Document Clustering: A Detailed Review," *International Journal of Applied Information Systems*, vol. 4, pp.

30-38, October 2012.

- [33] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze, *An Introduction to Information Retrieval*, Online edition ed. Cambridge, England: Cambridge University Press, 2009.
- [34] Chun-Ling Chen, Frank S.C. Tseng, and Tyne Liang, "An integration of WordNet and fuzzy association rule mining for multi-label document clustering," *Data & Knowledge Engineering*, vol. 69, no. 11, pp. 1208-1226, November 2010.
- [35] Cheng-Fa Tsai, Han-Chang Wu, and Chun-Wei Tsai, "A new data clustering approach for data mining in large databases," in *Parallel Architectures, Algorithms and Networks, 2002. I-SPAN '02. Proceedings. International Symposium*, Makati City, 2002, pp. 278-283.
- [36] Stuart P. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129-137, March 1982.
- [37] Hans-Hermann Bock, "Clustering Methods: A History of k-Means Algorithms," in *Selected Contributions in Data Analysis and Classification*. Berlin: Springer Berlin Heidelberg, 2007, pp. 161-172.
- [38] Xin Jin and Jiawei Han, "K-Means Clustering," in *Encyclopedia of Machine Learning*, Claude Sammut and Geoffrey I. Webb, Eds. United States: Springer US, 2010, pp. 563-564.
- [39] Erendira Rendon, Alejandra Arizmendi Itzel Abundez, and Elvia M. Quiroz, "Internal versus External cluster validation indexes," *International Journal of Computers and Communications*, vol. 5, no. 1, pp. 27-32, 2011.
- [40] M. Bilal, S. Masud, and S. Athar, "FPGA Design for Statistics-Inspired Approximate Sum-of-Squared-Error Computation in Multimedia Applications," *Circuits and Systems II: Express Briefs, IEEE Transactions*,

vol. 59, no. 8, pp. 506-510, July 2012.

- [41] Thomas Weise and R. Chiong, "Evolutionary Data Mining Approaches for Rule-based and Tree-based Classifiers," *Cognitive Informatics (ICCI), 2010 9th IEEE International Conference*, pp. 696-703, July 2010.
- [42] Melanie Mitchell, *An Introduction To Genetic Algorithms*. London, England: MIT Press, 1999.
- [43] Noraini Mohd Razali and John Geraghty, "Genetic Algorithm Performance with Different Selection Strategies in Solving TSP," in *Proceedings of the World Congress on Engineering 2011*, vol. II, London, 2011.
- [44] Abdul Wahid, Xiaoying Gao, and Andreae Peter, "Multi-View Clustering of Web Documents using Multi-Objective Genetic Algorithm," in *Evolutionary Computation (CEC), 2014 IEEE Congress*, Beijing, 2014, pp. 2625 - 2632.
- [45] Xiaoyu Wen, Xinyu Li, Liang Gao, Liang Wan, and Wenwen Wang, "Multi-objective genetic algorithm for integrated process planning and scheduling with fuzzy processing time," in *Advanced Computational Intelligence (ICACI), 2013 Sixth International Conference*, Hangzhou, 2013, pp. 293-298.
- [46] Abdullah Konak, David W. Coit, and Alice E. Smith, "Multi-objective optimization using genetic algorithms: A tutorial," *Reliability Engineering & System Safety*, vol. 91, no. 9, pp. 992-1007, September 2006.
- [47] Fatimah Sham Ismail, Rubiyah Yusof, and S Muhd Masduqi Waqiyuddin, "Multi-objective optimization problems: Method and application," in *Modeling, Simulation and Applied Optimization (ICMSAO), 2011 4th International Conference*, Kuala Lumpur, 2011, pp. 1-6.
- [48] Nafissa Zeghichi, Mekki Assas, and Leila Hayet Mouss, "Genetic Algorithm with Pareto Fronts for Multi-Criteria Optimization Case Study Milling

- Parameters Optimization," in *Software, Knowledge Information, Industrial Management and Applications (SKIMA), 2011 5th International Conference*, Benevento, 2011, pp. 1-5.
- [49] K. Atashkari, N. NarimanZadeh, A. R. Ghavimi, M. J. Mahmoodabadi, and F. Aghaienezhad, "Multi-objective optimization of power and heating system based on artificial bee colony," in *Innovations in Intelligent Systems and Applications (INISTA), 2011 International Symposium*, Istanbul, 2011, pp. 64-68.
- [50] Zhi-Hua Zhou, "Ensemble Learning," in *Encyclopedia of Biometrics*, Stan Z. Li and Anil Jain, Eds. United States: Springer, 2009, pp. 270-273.
- [51] Shaohua Wan and Hua Yang, "Comparison among Methods of Ensemble Learning," in *Biometrics and Security Technologies (ISBAST), 2013 International Symposium*, Chengdu, 2013, pp. 286-290.
- [52] S. Kulkarni and V. Kelkar, "Classification of multispectral satellite images using ensemble techniques of bagging, boosting and adaboost," in *Circuits, Systems, Communication and Information Technology Applications (CSCITA), 2014 International Conference*, Mumbai, April 2014, pp. 253-258.
- [53] Xiang Hui and Yang Sheng Gang, "Using clustering-based bagging ensemble for credit scoring," in *Business Management and Electronic Information (BMEI), 2011 International Conference*, vol. 3, Guangzhou, May 2011, pp. 369-371.
- [54] S. Bernard, S. Adam, and L. Heutte, "Using Random Forests for Handwritten Digit Recognition," in *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference*, vol. 2, Parana, September 2007, pp. 1043-1047.
- [55] Jian Xu, Jianhua Chen, and Bin Li, "Random Forest for Relational Classification with Application to Terrorist Profiling," in *Granular Computing*,

2009, *GRC '09. IEEE International Conference*, Nanchang, August 2009, pp. 630-633.

- [56] Peter Buhlmann,,: Springer Handbooks of Computational Statistics 2012, 2011, pp. 985-1022.
- [57] Youqin Pan and Zaiyong Tang, "Ensemble methods in bank direct marketing," in *Service Systems and Service Management (ICSSSM), 2014 11th International Conference*, Beijing, June 2014, pp. 1-5.
- [58] Yijun Chen and Man Leung Wong, "An Ant Colony Optimization Approach for Stacking Ensemble," in *2010 Second World Congress on Nature and Biologically Inspired Computing*, Fukuoka, 2010, pp. 146-151.
- [59] R. Polikar, "Ensemble based systems in decision making," *Circuits and Systems Magazine, IEEE*, vol. 6, no. 3, pp. 21-45, September 2006.
- [60] Gao Wei and Cheng Mingshu, "A new dynamic credit scoring model based on clustering ensemble," in *Computer Science and Network Technology (ICCSNT), 2013 3rd International Conference*, Dalian, 2013, pp. 421-425.
- [61] Mao-ting Gao and Bing-jing Wang, "Text clustering ensemble based on genetic algorithms," in *Systems and Informatics (ICSAI), 2012 International Conference*, Yantai, 2012, pp. 2329-2332.
- [62] Rodrigo A. Coelho, Fabricio R. N. Guimaraes, and Ahmed A. A. Esmin, "Applying Swarm Ensemble Clustering Technique for Fault Prediction Using Software Metrics," in *Machine Learning and Applications (ICMLA), 2014 13th International Conference*, Detroit, 2014, pp. 356-361.
- [63] R Ghaemi, M. Sulaiman, N. Mustapha, and H. Ibrahim, "Improving of Initial Clusters Fitness in Genetic Guided-Clustering Ensembles," in *Information Technology: New Generations (ITNG), 2010 Seventh International*

Conference, Las Vegas, 2010, pp. 227-232.

- [64] Armindokht Hashempour Sadeghian and Hossein Nezamabadi-pour, "Document clustering using gravitational ensemble clustering," in *Artificial Intelligence and Signal Processing (AISP), 2015 International Symposium*, Mashhad, 2015, pp. 240-245.
- [65] Zhenya Zhang, Hongmei Cheng, Shuguang Zhang, Wanli Chen, and Qiansheng Fang, "Clustering aggregation based on genetic algorithm for documents clustering," in *Evolutionary Computation, 2008. CEC 2008. (IEEE World Congress on Computational Intelligence). IEEE Congress*, Hong Kong, 2008, pp. 3156-3161.
- [66] Rayner Alfred and Dimitar Kazakov, "Pattern-Based Transformation Approach to Relational Domain Learning Using Dynamic Aggregation for Relational Attributes," in *The 2006 International Conference on Data Mining*, Las Vegas, 2006, pp. 118-124.
- [67] Chung Seng Kheau, Rayner Alfred, and Lau Hui Keng, "Dimensionality Reduction in Data Summarization Approach to Learning Relational Data," in *Intelligent Information and Database Systems Lecture Notes in Computer Science*. Berlin, Germany: Springer, 2013, pp. 166-175.