UNIVERSITI PUTRA MALAYSIA

FUZZY CLUSTERING METHOD AND EVALUATION BASED ON MULTI CRITERIA DECISION MAKING TECHNIQUE

FADHAA OTHMAN SAMEER

FS 2018 28
FUZZY CLUSTERING METHOD AND EVALUATION BASED ON MULTI CRITERIA DECISION MAKING TECHNIQUE

By

FADHAA OTHMAN SAMEER

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfilment of the Requirements for the Degree of Doctor of Philosophy

January 2018
COPYRIGHT

All material contained within the thesis, including without limitation text, logos, icons, photographs and all other artwork, is copyright material of Universiti Putra Malaysia unless otherwise stated. Use may be made of any material contained within the thesis for non-commercial purposes from the copyright holder. Commercial use of material may only be made with the express, prior, written permission of Universiti Putra Malaysia.

Copyright ©Universiti Putra Malaysia
DEDICATIONS

• To my Prophet Mohamed who taught the humanity. To respectful father and a lovely mother who taught me the meaning of courage and always had confidence in me.

• To my brothers Safhaa, Deyhaa, Muhmad, Ahmed and Alaa also my sister Nuras and for all their charge and support of me a lot and made it all possible for me.

• To all my friends who accompanied me through the different parts of my study.
Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfillment of the requirement for the degree of Doctor of Philosophy

FUZZY CLUSTERING METHOD AND EVALUATION BASED ON MULTICRITERIA DECISION MAKING TECHNIQUE

By

FADHAA OTHMAN SAMEER

January 2018

Chairman: Associate Professor Mohd. Rizam Abu Bakar, PhD
Faculty: Science

In the financial sector, credit scoring is one of the most successful operational research techniques. Credit scoring is an evaluation of the risk connected with lending to clients (consumers) or an organization. In actual credit scoring-related problems, generally inaccurate parameters or input data are used due to incomplete or inaccessible information being provided. Thus, designing a successful credit scoring model is then becoming more complex. Furthermore, the fuzzy approach is more efficient than the others to handle imprecisions and uncertainties. Hence, fuzzy clustering analysis such as the Gustafson-Kessel (GK) algorithm is seen to be a very important tool in the field of credit scoring. In a credit scoring problem with cluster analysis, finding a subset of features from large data sets is a very important issue. In addition, two other important problems are the requiring predefined number of clusters and selecting initial centres of clusters. Thus in this study we intend to overcome these problems by determining a feature subset and the number of the cluster problems after developing an algorithm which simultaneously solved these issues. This proposed algorithm is developed based on heuristic method named modified binary particle swarm optimization (MBPSO) with kernel fuzzy clustering method as a fitness function. The proposed algorithm is used as a pre-processing method for data followed by Gustafson-Kessel (GK) algorithm to classify credit scoring data. For the third problem a modified of Kohonen Network (MKN) algorithm was proposed to select the initial centres of clusters. A similar degree between points was utilized to get similarity density, and then by means of maximum density points selecting them as weights of the Kohonen algorithm. After the optimization of the weights by modified version of the Kohonen Network method these weights will be set as the initial centres of the Gustafson-Kessel (GK) algorithm. Hence, we proposed a complete method by combining MBPSO, MKN and GK (MBPSO+MKN+GK). The new proposed method (MBPSO+MKN+GK) Gustafson-Kessel algorithm (GK) integrated with modified of Kohonen Network algorithm (MKN)
and modified binary particle swarm optimization (MBPSO) was used to classify the credit scoring data. Multi-criteria decision making was used for measuring the overall preference values of these methods and considered all the criteria at the same time. The technique for order preference by similarity to ideal solution (TOPSIS) was used for ranking the fuzzy clustering processes having multiple criteria. Furthermore, the weights of the criteria were determined by using the modified fuzzy analytic hierarchy process (MFAHP) with ranking function. Simulation experiments were carried out to investigate the performance of methods with different number of samples and different number of features. Also these methods were applied on two credit scoring datasets of German and Australian. For a real problem application, we consider the data from Gulf Commercial Bank in Iraq. This study revealed that the GK along with the MBPSO algorithm showed a better performance as compared to the GK algorithm alone. Also, the GK and MKN algorithms together were better than GK alone. But the best performance of all will be the MBPSO+MKN+GK.
Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

KAEDAHD PENGELOMPOKAN KABUR DAN PENILAIAN BERASASKAN TEKNIK MEMBUAT-KEPUTUSAN PELBAGAI-KRITERIA

Oleh

FADHAA OTHMAN SAMEER

Januari 2018

Pengerusi : Associate Professor Mohd. Rizam Abu Bakar, PhD
Fakulti : Science

Dalam sektor kewangan, pemarkahan kredit adalah salah satu daripada teknik-teknik penyelidikan operasi yang paling berjaya. Pemarkahan kredit ialah penilaian risiko berhubung pemberian pinjaman kepada klien (pengguna) atau suatu organisasi. Dalam permasalahan sebenar berkaitan pemarkahan kredit, secara umumnya parameter atau data input yang tidak tepat digunakan oleh kerana maklumat tidak lengkap atau tidak tercapai yang diberikan. Maka mereka bentuk model pemarkahan kredit yang berjaya menjadi bertambah kompleks. Tambahan pula, pendekatan kabur lebih cekap daripada yang lain dalam menangani ketidaktepatan dan ketakpastian. Oleh itu analisis berkelompok kabur seperti algoritma Gustafson-Kessel (GK) dilihat sebagai satu kaedah yang penting dalam bidang pemarkahan kredit. Dalam permasalahan pemarkahan kredit dengan analisis berkelompok, mencari subset ciri daripada set data yang besar adalah isu sangat penting. Tambahan, dua permasalahan penting yang lain adalah keperluan prapertakrif bilangan kelompok dan mencari pusat kelompok permulaan. Maka dalam kajian ini kami mencadangkan satu kaedah untuk mengatasi masalah ini dengan penentuan subset ciri dan masalah bilangan kelompok selepas membangunkan satu algoritma yang mampu menyelesaikan masalah ini secara serentak. Algoritma yang dicadang ini dibangun berdasarkan kepada kaedah heuristik yang dipanggil pengoptimuman kerumunan zarah biner terubahsuai (MBPSO) dengan kaedah pengelompokan intipati kabur sebagai fungsi kebugaran. Algoritma yang dicadang digunakan dalam kaedah praprosesan untuk data dan diikuti dengan algoritma Gustafson-Kessel (GK) untuk mengelaskan data pemarkahan kredit. Untuk masalah ketiga algoritma rangkaian Kohonen (MKN) dicadang untuk memilih pusat permulaan kelompok. Darjah setara antara titik digunakan untuk mendapatkan kepadatan keserupaan dan seterusnya menggunakan min kepadatan titik pemilihan maksimum sebagai pemberat algoritma Kohonen. Selepas pengoptimuman pemberat menggunakan kaedah rangkaian Kohonen terubahsuai,
ACKNOWLEDGEMENTS

First and foremost, I would like to thank GOD for the bounties who has granted me. I thank him for giving me the ability to deal with my challenges during my research and letting me accomplishing this thesis.

I would like to thank Assoc. Prof. Dr. Mohd Rizam Abu Bakar, my supervisor, for the professional, scientific, and personal guidance he has generously given me through this long journey.

My gratitude also goes to the members of my supervisory committee, Prof. Dr. Iden Hassan Hussain Al-Kanani and Assoc. Prof. Dr. Lee Lai Soon and Assoc. Prof. Dr. Leong Wah June for their expertise and important contributions, useful suggestions helpful comments and personal mentor.

Also importantly, my special thanks to my family, my parents, my sister and brothers who encouraged me not to miss my hope in doing my research and supported me a lot mentally.

I would like to thank the Ministry of Higher Education and Scientific Research of Iraq for the financial supporting of the scholarship. Further thanks to Gulf Commercial Bank for supporting my research by providing of data. Much gratitude is also due to the Universiti Putra Malaysia members who created an environment in which Ph.D. students can flourish. My acknowledgement would be incomplete without mentioning of my friends who made wonderful memories for me. Thanks you all.
I certify that a Thesis Examination Committee has met on 23 January 2018 to conduct
the final examination of Fadhaa Othman Sameer on her thesis entitled "Fuzzy
Clustering Method and Evaluation Based on Multi Criteria Decision Making
Technique" in accordance with the Universities and University Colleges Act 1971 and
the Constitution of the Universiti Putra Malaysia [P.U.(A) 106] 15 March 1998. The
Committee recommends that the student be awarded the Doctor of Philosophy.

Members of the Thesis Examination Committee were as follows:

Zarina Bibi binti Ibrahim, PhD
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Chairman)

Habshah binti Midi, PhD
Professor
Faculty of Science
Universiti Putra Malaysia
(Internal Examiner)

Mohd Bakri bin Adam, PhD
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Internal Examiner)

Anthony Bellotti, PhD
Senior Lecturer
Imperial College London
United Kingdom
(External Examiner)

NOR AINI AB. SHUKOR, PhD
Professor and Deputy Dean
School of Graduate Studies
Universiti Putra Malaysia

Date: 28 March 2018
This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfilment of the requirement for the degree of Doctor of Philosophy. The members of the Supervisory Committee were as follows:

**Mohd. Rizam Abu Bakar, PhD**  
Associate Professor  
Faculty of Science  
Universiti Putra Malaysia  
(Chairperson)

**Lee Lai Soon, PhD**  
Associate Professor  
Faculty of Science  
Universiti Putra Malaysia  
(Member)

**Leong Wah June, PhD**  
Associate Professor  
Faculty of Science  
Universiti Putra Malaysia  
(Member)

**Iden Hassan Hussain Al-Kanani, PhD**  
Professor  
Faculty of Science  
Universiti Putra Malaysia  
(Member)

---

**ROBIAH BINTI YUNUS, PhD**  
Professor and Dean  
School of Graduate Studies  
Universiti Putra Malaysia

Date:
Declaration by graduate student

I hereby confirm that:
• this thesis is my original work;
• quotations, illustrations and citations have been duly referenced;
• this thesis has not been submitted previously or concurrently for any other degree at any other institutions;
• intellectual property from the thesis and copyright of thesis are fully-owned by Universiti Putra Malaysia, as according to the Universiti Putra Malaysia (Research) Rules 2012;
• written permission must be obtained from supervisor and the office of Deputy Vice-Chancellor (Research and Innovation) before thesis is published (in the form of written, printed or in electronic form) including books, journals, modules, proceedings, popular writings, seminar papers, manuscripts, posters, reports, lecture notes, learning modules or any other materials as stated in the Universiti Putra Malaysia (Research) Rules 2012;
• there is no plagiarism or data falsification/fabrication in the thesis, and scholarly integrity is upheld as according to the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) and the Universiti Putra Malaysia (Research) Rules 2012. The thesis has undergone plagiarism detection software.

Signature: ___________________________ Date: ___________________________

Name and Matric No: Fadhaa Othman Sameer, GS41271
Declaration by Members of Supervisory Committee

This is to confirm that:
- the research conducted and the writing of the thesis was under supervision;
- supervision responsibilities as stated in Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) are adhered to.

Signature : _____________________________
Name of Chairman of Supervisory Committee : Associate Professor Dr. Mohd Rizam Abu Bakar

Signature : _____________________________
Name of Member of Supervisory Committee : Associate Professor Dr. Lee Lai Soon

Signature : _____________________________
Name of Member of Supervisory Committee : Associate Professor Dr. Leong Wah June

Signature : _____________________________
Name of Member of Supervisory Committee : Professor Dr. Iden Hassan Hussain Al-Kanani
TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
</tr>
<tr>
<td>ABSTRAK</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
</tr>
<tr>
<td>APPROVAL</td>
</tr>
<tr>
<td>DECLARATION</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
</tr>
<tr>
<td>LIST OF ABBREVIATIONS</td>
</tr>
</tbody>
</table>

CHAPTER

1 INTRODUCTION
   1.1 Scope of analysis 1
   1.2 General Review 1
   1.3 Statement of the problem 3
   1.4 Motivation 4
   1.5 Preliminaries 6
   1.6 The objectives of research 6
   1.7 Scope of the present work 7

2 LITERATURE REVIEW OF CREDIT SCORING
   2.1 Introduction 9
   2.2 Credit scoring definition 9
   2.3 Literature Review of Credit Scoring techniques 10
      2.3.1 Single classifier techniques 11
      2.3.2 Hybrid classifier approaches 14
      2.3.3 Clustering analysis 21
      2.3.4 Background of clustering 24
      2.3.5 Determination of clustering model parameters 29
      2.3.6 Fuzzy clustering methods used for credit scoring 30
   2.4 MCDM for evaluating the clustering algorithms 31
   2.5 Limitation Gaps of Existing Work and Research Challenges 33
   2.6 Summary 33

3 METHODOLOGY
   3.1 Introduction 35
      3.1.1 Classification 35
   3.2 Clustering Analysis 35
3.2.1 Hierarchical clustering 36
3.2.2 Non-hierarchical clustering 36
3.3 Particle swarm optimization algorithm 38
  3.3.1 Binary Particle Swarm Optimization (BPSO) 40
  3.3.2 Modified Binary Particle Swarm Optimization (MBPSO) 41
  3.3.3 Steps of the modified algorithm (MBPSO) 42
3.4 Kohonen network method 46
3.5 Modified Kohonen Algorithm 46
3.6 Measures of cluster validity 47
  3.6.1 Internal measures 48
  3.6.2 External measures 49
  3.6.3 Sensitivity and Specificity 52
  3.6.4 Type I and Type II errors 53
  3.6.5 Integrated performance measures 53
3.7 Multi-criteria decision making (MCDM) 53
  3.7.1 Fuzzy sets and fuzzy numbers 54
  3.7.2 Fuzzy analytic hierarchy process (FAHP) 55
  3.7.3 Determining the weights of the criteria 56
  3.7.4 Ranking function 56
  3.7.5 Basic definition and rules of ranking function 57
  3.7.6 The algorithm of ranking function 57
  3.7.7 Determining the weights of the criteria 58
  3.7.8 Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) 59

4 RESULTS AND DISCUSSION 62
4.1 Simulation study 62
  4.1.1 The Monte Carlo Method: Basic Principles 62
  4.1.2 Setup 62
  4.1.3 MCDM of simulation data 66
4.2 Generic data 70
  4.2.1 Data description 70
  4.2.2 Application of BPSO and MBPSO 70
  4.2.3 Application of Kohonen method and modified Kohonen method 72
  4.2.4 Application of external indices for generic data 74
  4.2.5 Sensitivity and Specificity 75
  4.2.6 Related work 77
4.3 Multi-criteria decision making (MCDM) 78
4.4 Case study 84
  4.4.1 Gulf Commercial Bank data 84
  4.4.2 Sensitivity and Specificity 88
  4.4.3 Multi-criteria decision making (MCDM) of Gulf Commercial Bank data 89
4.5 Summary 89
5 CONCLUSION

5.1 Introduction 91
5.2 Main contribution 91
5.3 Conclusion 92
5.4 Recommendations for further research in future 93

BIBLIOGRAPHY 94
APPENDICES 104
BIODATA OF STUDENT 106
LIST OF PUBLICATIONS 107
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 Credit scoring techniques and their application</td>
<td>12</td>
</tr>
<tr>
<td>2.2 Classification of the studies according to the clustering credit scoring techniques.</td>
<td>25</td>
</tr>
<tr>
<td>2.3 Classification of the studies according to the fuzzy clustering credit scoring techniques</td>
<td>31</td>
</tr>
<tr>
<td>3.1 The confusion matrix</td>
<td>52</td>
</tr>
<tr>
<td>4.1 The internal validity measures of simulation data for $p = 20$</td>
<td>63</td>
</tr>
<tr>
<td>4.2 The internal validity measures of simulation data for $p = 25$</td>
<td>64</td>
</tr>
<tr>
<td>4.3 Membership function of linguistic scale of internal indices of simulation data with $n = 500$, $p = 20$</td>
<td>66</td>
</tr>
<tr>
<td>4.4 Dataset description</td>
<td>70</td>
</tr>
<tr>
<td>4.5 The internal validity measures of Australian credit data</td>
<td>71</td>
</tr>
<tr>
<td>4.6 The internal validity measures of German credit data</td>
<td>71</td>
</tr>
<tr>
<td>4.7 The internal validity measures of Australian credit data</td>
<td>73</td>
</tr>
<tr>
<td>4.8 The internal validity measures of German credit data</td>
<td>73</td>
</tr>
<tr>
<td>4.9 The fuzzy Rand validity measure of two credit data</td>
<td>74</td>
</tr>
<tr>
<td>4.10 The fuzzy Fowlkes-Mallow index of two credit data</td>
<td>74</td>
</tr>
<tr>
<td>4.11 The fuzzy DNC index with $\alpha = 0.05$ of two credit data</td>
<td>74</td>
</tr>
<tr>
<td>4.12 The confusion matrix of Australian data set for methods</td>
<td>75</td>
</tr>
<tr>
<td>4.13 The sensitivity and specificity of Australian data set for methods</td>
<td>76</td>
</tr>
<tr>
<td>4.14 The confusion matrix of German data set for methods</td>
<td>76</td>
</tr>
<tr>
<td>4.15 The sensitivity and specificity of German data set for methods</td>
<td>77</td>
</tr>
<tr>
<td>4.16 Membership function of linguistic scale of internal indices of Australian dataset</td>
<td>78</td>
</tr>
</tbody>
</table>
4.17 Membership function of linguistic scale of internal indices German dataset
4.18 Number of good and bad customers for every year
4.19 Number of males and females customers for every year
4.20 Number of good and bad customers for males and females
4.21 Number of customers for warranty provided
4.22 Number of customers for type of loan
4.23 Number of customers for beneficiary sector.
4.24 The internal validity measures of Gulf Commercial Bank data
4.25 The external fuzzy validity measure of Gulf Commercial Bank data
4.26 The confusion matrix of Gulf Commercial Bank data set for methods
4.27 The sensitivity and specificity of Gulf Commercial Bank data set for methods

xiv
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>16</td>
</tr>
<tr>
<td>2.2</td>
<td>25</td>
</tr>
<tr>
<td>3.1</td>
<td>44</td>
</tr>
<tr>
<td>3.2</td>
<td>45</td>
</tr>
<tr>
<td>3.3</td>
<td>51</td>
</tr>
<tr>
<td>3.4</td>
<td>54</td>
</tr>
<tr>
<td>4.1</td>
<td>65</td>
</tr>
<tr>
<td>4.2</td>
<td>65</td>
</tr>
<tr>
<td>4.3</td>
<td>66</td>
</tr>
</tbody>
</table>

2.1 The steps of feature selection process

2.2 Diagram of clustering.

3.1 The binary particle swarm Optimization (BPSO)

3.2 The modified binary particle swarm Optimization (MBPSO)

3.3 Fuzzy sets: good (match), bad (mismatch) and uncertainty (confidence degree lower than $\alpha$.)

3.4 A triangular membership function $\mu_A(x)$

4.1 Internal indices values of simulation data with $n = 500, p = 20$

4.2 Internal indices values of simulation data with $n = 1000, p = 20$

4.3 Internal indices values of simulation data with $n = 4000, p = 20$
**LIST OF ABBREVIATIONS**

<table>
<thead>
<tr>
<th>˜A</th>
<th>Fuzzy set</th>
</tr>
</thead>
<tbody>
<tr>
<td>BPSO</td>
<td>Binary particle swarm optimization</td>
</tr>
<tr>
<td>C</td>
<td>Clusters centres matrix</td>
</tr>
<tr>
<td>d</td>
<td>Distance function</td>
</tr>
<tr>
<td>F</td>
<td>Covariance matrix</td>
</tr>
<tr>
<td>FCM</td>
<td>Fuzzy C-Means clustering algorithm</td>
</tr>
<tr>
<td>FS</td>
<td>Features selection</td>
</tr>
<tr>
<td>GK</td>
<td>Gustafson-Kessel algorithm</td>
</tr>
<tr>
<td>J</td>
<td>Objective function</td>
</tr>
<tr>
<td>KFCM</td>
<td>Kernel fuzzy clustering method</td>
</tr>
<tr>
<td>KN</td>
<td>Kohonen network</td>
</tr>
<tr>
<td>MBPSO</td>
<td>Modified binary particle swarm optimization</td>
</tr>
<tr>
<td>MFAHP</td>
<td>Modified fuzzy analytic hierarchy process</td>
</tr>
<tr>
<td>MKN</td>
<td>Modified Kohonen network</td>
</tr>
<tr>
<td>μ˜A(x)</td>
<td>Membership of x in ˜A</td>
</tr>
<tr>
<td>n</td>
<td>The sample size</td>
</tr>
<tr>
<td>PSO</td>
<td>Particle swarm optimization</td>
</tr>
<tr>
<td>p</td>
<td>Features of data</td>
</tr>
<tr>
<td>SOM</td>
<td>Self-organizing map</td>
</tr>
<tr>
<td>TOPSIS</td>
<td>Technique for order preference by similarity to ideal solution</td>
</tr>
<tr>
<td>U</td>
<td>Partition matrix</td>
</tr>
<tr>
<td>X</td>
<td>Data set</td>
</tr>
</tbody>
</table>
CHAPTER 1

INTRODUCTION

1.1 Scope of analysis

In recent years, finance officers and bankers, all over the world, have faced many challenges. These challenges primarily relate to the complexity of financial markets, as a result of growing demand for financing. This complication in financial environments forces the financiers to find appropriate tools for measuring losses which may arise in the banking sector (Choudhary et al., 2009). This can be done through measuring the major risks named market risk, operational risk and credit risk. As the major operation of banks and financial institutions include lending activities, credit risk is therefore a major risk observed in the banking and financial system. To safeguard against losses which may arise when faced with non-performing loans, measuring the risk and making adequate provisions, based on those calculations, is a must for every financial institution. In this regard, during the recent decades, risk management and especially credit risk management has attracted many academics as well as researchers from the financial sector (Shojai et al., 2010).

1.2 General Review

The credit risk can be described as the risk of the principal loss or the loss if the financial returns which occur due to the borrower’s failure to either fulfil the agreed upon obligations or repay the loan. Credit risk arises and can be expected when a borrower does not meet their obligation in relation to future cash flows. Therefore, there is uncertainty over the borrower’s financial performance in the future. As a result, in recent years, financiers have sought tools and means to enable them to calculate the borrower’s credit worthiness. Credit scoring is a type of risk evaluation process which is seen to be a very important decision made by the financial institutions for preventing any wrong decisions which could potentially lead to huge losses. The various classification models belong to the data mining approach used by the decision makers to decrease the credit risk of the customers. As accuracy is an essential criterion for selecting the appropriate credit scoring model, many studies have been carried out for improving the efficiency of the credit scoring models (Khashei et al., 2013). A majority of the studies have used data mining techniques for investigating problems related to credit scoring. Many data mining models were proposed and applied. Also, in the early 1930-40s, the mail-order companies used the numerical scorecards for credit scoring (Thomas et al., 2005). It is an established fact that the data mining techniques could prove to be helpful to the decision makers and the financial managers to find out the hidden knowledge from the input data regarding their clients and their different requests. Also, this knowledge helps the financial institutions decrease the risk of clients (Khashei et al., 2013). Since the beginning, credit is always has always been for enhancing consumption. Many forms of credit like borrowing, lending, instalments..
and payments before or after the delivery of service or goods, consumption credit etc. already exist and help in enabling smoother transactions and increase economic growth. Even today, these consumer credit alternatives facilitate consumption and improve the economy. Many recent credit scoring models have been proposed based on the operational research, statistical techniques, and artificial intelligence (AI) (Thomas et al., 2005). However, the statistical techniques can be applied when determining a linear relationship between the dependent and the independent variables. Also artificial intelligence techniques have a major limitation which lies in their poor understandability because of the black box nature. It is very complicated for artificial intelligence techniques to make knowledge representation and these techniques require a vast number of training samples and long learning time (Thomas et al., 2005).

Moreover, the drawbacks of the application of mathematical programming techniques for risk assessment of the credit are mathematical programming techniques need a computation of effort that is rarely known by the business and financial analysts (Marques et al., 2013). Though intelligent and statistical techniques are classification models they are crispy models that use the classic logic in their procedures of modelling. However; these two techniques can not be able to get effective models for the uncertainties that exist in the relationships and data (Khashei et al., 2009). Therefore; to overcome these drawbacks, a soft models or fuzzy models have been proposed to the financial sector which is a second category of credit scoring models.

The probable future performance of the customers is changing over time which is humanity behaviour. In the case of credit applications it is difficult to gather relevant information about individuals for determining their credit worthiness score. Also the final decision of the experts based on the applicant characteristics like credit history, age, income, loan required etc. However, in actual credit scoring problems, the input parameters and data are often incorrect as the information could be unattainable or incomplete (Lin, 2010). Therefore, there are many complex difficulties which appear while tackling the uncertainties during the designing of an appropriate credit scoring model. Furthermore, the fuzzy approach is more efficient than the others to handle imprecisions and uncertainties (Safaei et al., 2008).

The data collected provides a further basis for analysing, reasoning, making decisions and finally, for understanding different phenomena and objects. One important data analysis activity involves classifying or grouping the data into sets of clusters or categories. Objects placed in one group showed similar properties depending on a particular criterion (Zakrzewska, 2007). Cluster analysis or clustering is defined as the identification of the subsets of similar objects. This subset generally and intuitively is seen to correspond to points which are similar to one another as compared to points from other clusters. These points in the same cluster display a similar label. Clustering can be conducted in an unsupervised manner by determining the similar subset of points, without using a predefined idea of a cluster (Lim and Sohn, 2007).
Thus, the determination of the clusters having labels is based on data instead of using a specific model or a perspective of this data. There is a huge need for developing reliable models which can help in predicting the defaults (Bezdek et al., 1984).

One of the statistical method that be non-parametric which can be used for credit scoring is clustering analysis. The essential advantages of this type of analysis is that a specification of distribution does not assume for data, so it is a suitable method where sufficient priorly hypotheses does not exist. It is therefore exploratory and identifies the most likely solution which makes it suitable for use in credit risk analysis (Huang and Dun, 2008). The clustering is seen to be a substantial unsupervised learning problem; and like other similar problems, it also deals with classifying and categorising the collection of the unlabelled data (Kaufman and Rousseeuw, 2009). Furthermore, a fuzzy clustering is more flexibly than hard clustering where each object has memberships in all clusters instead of in a single cluster. The membership matrix provides more information to help the users to decide the core and boundary objects of clusters (Bai et al., 2013).

Many of the recently published studies have stated that combining many classifiers (or classifier ensembles) is better than using a single classifier (Tsai and Wu, 2008). Hence, hybrid models which have combined the advantages of many models are a very hot topic of research. The simple hybrid model consists of a process of developing a credit evolution model which is further classified into three steps: feature subset selection (FSS), determining the parameters of model and classification. This approach helps in selecting different methods for the above-mentioned three steps. The feature selection step is very important as it restricted the input feature number to improve the prediction accuracy and further decrease the computational complexity.

Furthermore, credit scoring databases in particular are often large and contain a lot of redundant and irrelevant features, so it is more demanding in terms of computation cost to classify such data. However, this difficulty can be overcome by using a feature selection method. Thus the selection of features is one of the important and most challenging issues in credit scoring.

1.3 Statement of the problem

Loan credit approval or evaluation is defined as the process which is carried out for any individual or business application for credit, which determines the eligibility of the customer for the loan (Louzada et al., 2012). The evaluation systems of credit risk is an important role in the financial decision-making. These systems diminish possible risks to enable faster decisions of credit and to reduce the cost of credit analysis. The need to control and effectively to manage credit risk has led financial institutions to strive to improve the designed techniques for credit. As a result the development of various quantitative models was promulgated by financial institutions.
and consulting companies. Thus, the numbers of academic studies about credit scoring were developed to show a diversity of classification methods applied to distinguish good and bad customers (Louzada et al., 2016). A majority of the studies which developed computation-based credit scoring models had the objective of obtaining better classification accuracy than other models. One of the statistical methods that is non-parametric which can be used for credit scoring is clustering analysis. The essential advantages of this type of analysis is that a specification of distribution is not assumed for data, so it is a suitable method where sufficient prior hypotheses does not exist (Lim and Sohn, 2007).

The researchers seek to solve the uncertainty and imprecisions in data and the future performance of the customers based on fuzzy clustering methods. Credit scoring databases in particular are often large and contain a lot of redundant and irrelevant features, so it is more demanding in terms of computation cost to classify such data. However, this difficulty can be overcome by using a feature selection method. Thus the selection of features is one of the important issue and most challenging in credit scoring. The models of the following researchers (Hoffmann et al. (2007), Lahsasna et al. (2010), Zhou (2012) and Gholamian et al. (2013)) adopted fuzzy clustering methods for handling uncertainty to evaluate credit risk problem without using feature subset selection techniques.

As result, many researchers have used fuzzy clustering with features subset selection in the credit scoring problem (Mehdizadeh (2009), Sadatrasoul et al. (2015) and Zhou and Li (2016)). In addition, there are two main problems in clustering analysis which include the number of clusters and the initial centres of clusters. In clustering analysis, one of the most challenging hard problems to solve is the number of clusters. Usually this number is fixed before clustering data by researchers. However, having a predetermined number of clusters is not realistic for a lot of data analysis in the real world. Also, the above researches used only features selection without integrating the two problems of feature selection and the number of clusters as pre-processing before clustering the data. After having gained the most relevant features and the optimal number of clusters, the initial centres of clusters were selected by the optimization method. Therefore, this study proposes a general and comprehensive algorithm to classify the credit scoring problem based on clustering analysis given fuzzy environment features selection and the two problems of clustering algorithms.

1.4 Motivation

The systems of credit risk evaluation have a main role in decision-making of the financial sector. These systems diminish possible risks to enable faster decisions of credit and to reduce the cost of credit analysis. The need to control and effectively to manage credit risk has led financial institutions to strive to improve the designed techniques for credit (Louzada et al., 2016). The discovery or extracting knowledge from datasets is one among the most desirable tasks in credit scoring. The data
mining can deal with this problem in the form of data clustering which is one of the most common data mining tasks which can be used for credit scoring (Lim and Sohn, 2007). The difficulties of handling uncertainties while designing a successful credit scoring model is becoming more complex. Furthermore, the fuzzy approach is more efficient than the others to handle imprecisions and uncertainties (Safaei et al., 2008). The credit scoring databases in particular are often large and contain a lot of redundant and irrelevant features so it is more demanding in terms of computation cost to classify such data. However, this difficulty can be overcome by using a features selection method. Moreover, the particle swarm optimization algorithm (PSO) is a type of global optimisation algorithm that is used for solving the problems wherein the optimal solution lies as a point in the parameters of multi-dimensional space. The PSO shows a better-structured neighbourhood which helps it display a better recombination method as compared to the GA. Also, it contains a velocity term which helps it attain a faster convergence to a good solution. The binary and the continuous PSOs are used for the filter and the wrapper methods and for the single and multiple objective feature selection (Poli et al., 2007). One main problem in traditional BPSO is that the new position of a particle is solely decided by its velocity while the particles current position hardly has any in influence in updating its next position (Zhang and Liu, 2008). So a modification of the updating function of the (BPSO) is needed by using velocity and position to update the next position for every particle in the swarm in contrast to the original binary particle swarm optimization which used only velocity. In addition, there are two main problems in clustering analysis which include the number of clusters and the initial centres of clusters. This thesis intends to develop a new BPSO to solve the feature subset selection and the number of clusters simultaneously.

Also the currently used techniques for initialising the cluster centres are classified into random sampling techniques density estimation techniques and distance optimisation processes. Out of these, the random sampling is very popular owing to its simplicity. The artificial neural network (ANN) has been widely utilized for as long as three decades for clustering and data classification. Also, the competitive or the winner takes all types of artificial neural network (ANN) are used for clustering the input data. In the competitive learning, the similar data patterns are often grouped by a network and can be represented by one unit neuron. In the comparative information designs are frequently gathered by a system and can be spoken to by one unit neuron. This type of collection is carried out consequently utilizing information relationship. Some well known illustrations where the (ANN) technique was used for clustering is the Kohonen algorithm self organising maps (SOM) (Khan et al., 2001). Since the Kohonen algorithm (KA) used random selecting technique to gain initial weights we seek to modify this method by using density selecting technique. After this the optimisation weights were selected as initial centres of the clustering model to overcome the drawbacks of the existing method. This new method has higher performance than the existing method in accuracy of classification of the credit scoring problem. The multi-criteria decision-making (MCDM) models were utilized to assess a finite set of alternatives (fuzzy clustering methods) with respect to multiple criteria (external and internal) at the same time.
1.5 Preliminaries

Many fuzzy methods, like the Fuzzy Clustering methods, were developed to help in making decisions, based on the fuzzy set theory (Zadeh, 1965). Clustering plays an important role in decision making and also helps in data mining, pattern recognition, and data modelling. The clustering technique classifies the data set based on the measure of similarity of the objects present in the data. The clustering of the data into many partitions helps in exploring the characteristic relationships between the objects in the data which already contains many samples of any decision procedure. The principle objective of the clustering process is classifying the data into more similarity-based symmetric clusters.

Unsupervised learning refers to extracting knowledge from any data set using the clustering technique if this technique used the similarity measure instead of the corrective acts that are supervised by known relations. The similarity between the objects in the data set is vital for clustering. One type of similarity measure involved in fuzzy clustering includes the distance measurement and many types of distance measures are used in the existing fuzzy clustering algorithms. The clustering algorithms classify the data objects (including entities, patterns, observances, instances, units) into a particular cluster (categories, groups or subsets). The cluster is defined as an aggregate of many points within a test space and the distance between any two points within the cluster is seen to be lesser than the distance present between a point inside and that present outside the cluster (Kaufman and Rousseeuw, 2009). Also a cluster is defined as a continuous region of space (D-dimensional feature space) which contains a higher density of points. Two such high-density regions are separated from each other by a space which contains lower density points. In the above definitions, the cluster has been defined based on the internal homogeneity and an external separation i.e., the objects within one cluster are similar to one another, while the objects in the different clusters are dissimilar from each other.

1.6 The objectives of research

This research aims to propose a model under the fuzzy clustering method for credit scoring with the following objectives:

1. To modify the binary particle swarm optimization algorithm to identify the number of clusters and to choose the most relevant features subset.
2. To modify the Kohonen algorithm to select the initial centres of clusters for the Gustafson-Kessel algorithm.
3. To develop a mathematical model based on fuzzy clustering analysis (Gustafson-Kessel method) for credit risk assessment.
4. To develop the fuzzy analytic hierarchy process (MFAHP) with ranking function to compute the weights for criteria (external and internal).
(5) To extend the multi criteria decision-making (MCDM) model (TOPSIS method) by hybridizing with the modified fuzzy analytic hierarchy process (MFAHP) to evaluate a finite set of fuzzy clustering methods.

(6) To validate and verify the improvement of Gustafson -Kessel method on simulation data, standard real data and data from Gulf Commercial Bank in Iraq.

1.7 Scope of the present work

The scope of this study involves a proposed modified binary particle swarm optimization (MBPSO) which is a hybrid model between the fuzzy clustering method (kernel fuzzy c-means method) and the binary particle swarm optimization method to predefine the number of clusters and to choose the most relevant features subset. Also proposed a modified Kohonen network algorithm (MK) for the selection of the initial centres for the Gustafson-Kessel algorithm (GK). After that the two modification methods (MBPSO) and (MK) with (GK) used as a new classification method. Multi-criteria decision making (MCDM) models were used to evaluate a finite set of alternatives (fuzzy clustering methods) with respect to multiple criteria (external and internal). Simulation study was conducted in this study. In addition, the credit scoring problem was studied. Here two datasets Australian and German were investigated to test the methods comprehensively. A set of real data from Gulf Commercial Bank in Iraq was also used for analysis purposes.

In this thesis, we are mainly concerned with determining a numerical solution for classifying the credit score data with the help of fuzzy clustering analysis. This thesis has been organised as per the general structure of the dissertations and thesis submitted to the Universiti Putra, Malaysia. This thesis consists of 5 chapters.

Chapter 1: provides a general introduction about credit scoring, problem statement, objectives and contributions of this thesis.

Chapter 2: provides a literature review and it investigates the studies that are related to the research topic. These studies include some which have been published earlier and some which are still in progress with regards to the credit scoring processes including the various clustering methods.

Chapter 3: this chapter contains the proposed methods, including the modified binary particle swarm optimisation for predefining the cluster number and selecting the most relevant feature subset simultaneously. And the modified Kohonen network algorithm which helps in selecting the initial centres of the Gustafson-Kessel algorithm. Also, the modified form of TOPSIS method was proposed for ranking the fuzzy clustering techniques using multiple criteria.

Chapter 4: describes the results and the discussion when the real and the predicted data was studied using various inference methods. All the results
obtained were evaluated, tabulated and discussed in further detail. In this chapter, the major findings have been highlighted and these findings were then compared to the objectives performance metrics that were used in this study for method evaluation.

Chapter 5: presents the conclusions of this study. Also future work and recommendations for some further research have been described in the chapter. This chapter also includes some remarks regarding the fulfilment of the objectives of this study. Also some suggestions are given for improving this work in future.
BIBLIOGRAPHY


