CLASSIFICATION OF STOCK MARKET INDEX BASED ON PREDICTIVE FUZZY DECISION TREE

RASHID HAFEEZ KHOKHAR

A thesis submitted in fulfilment of the requirement for the award of the degree of Master of Science (Computer Science)

Faculty of Computer Science and Information System
Universiti Teknologi Malaysia

JUNE 2005
Dedicated to my beloved Parents and Teachers
ACKNOLEDGEMENT

First of all I offer my humble thanks to Almighty Allah SWT, who made the things possible, which were impossible for me. Many Darood on our Holy prophet (peace be upon him) being a trivial ummati.

I would like to thank my supervisor, Professor Madya Dr Mohd Noor Md Sap for introducing me in data mining research. I want to express my deep gratitude to my supervisor for his support and guidance through my research as a master student. I am very grateful for the inspiring discussions that led to this thesis. Without their continued support and interest, this thesis would not have been the same as presented here.

I would like to thanks my dear parents Mohd Hafeez Khokhar and Zaineb Perveen, and my brothers for their continuous love, encouraging, support and confidence. In addition, I want to express my profound gratitude to my brothers, Assoc Prof Amjid Hafeez Khokhar and Eng Zahid Hafeez Khokhar for their valuable suggestions during my whole carrier. Without them, this thesis could not have been written.

My fellow postgraduate students Hafiz Faisal Zafar, Anjum Iqbal, Abdul Majid Awan, Mohd Azam Rana, Ahmad Hussain, Abdul Rab, and Adnan Younas should also be recognized for their valuable suggestions and support.

I must express my sincerest and heartiest thanks to my wife Faiqa Rashid and my other family members for their encouraging attitude unfailing patience and excellent co-operation through out my research.
Over the past decade many attempts have been made to predict stock market data using statistical and data mining models. However, most methods suffer from serious drawback due to requiring long training times, results are often hard to understand, and producing inaccurate predictions. In addition, the trader’s expectations to predict stock markets are seriously affected by some uncertain factors including political situation, oil price, overall world situation, local stock markets etc. Therefore, predicting stock price movements is quite difficult. Data mining techniques are able to uncover hidden patterns and predict future trends and behaviors in financial markets. In this research, another modification of Fuzzy Decision Tree (FDT) classification techniques called predictive FDT is presented that aims to combine symbolic decision trees in data classification with approximate reasoning offered by fuzzy representation. The intent is to exploit complementary advantages of both: ability to learn from examples, high knowledge comprehensibility of decision trees, and the ability to deal with uncertain information of fuzzy representation. In particular, predictive FDT algorithm is based on the concept of degree of importance of attribute contributing to the classification. After constructing predictive FDT, Weighted Fuzzy Production Rules (WFPRs) are extracted from predictive FDT, and then more significant WFPR’s are mined by using similarity-based fuzzy reasoning method. In fuzzy reasoning method the weights are assigned to each proposition in the antecedent part and the Certainty Factor (CF) is computed for the consequent part of each Fuzzy Production Rule (FPR). Finally, these rules are used to predict time series stock market in different periods to time. The predictive FDT’s are tested using three data sets including Kuala Lumpur Stock Exchange (KLSE), New York Stock Exchange (NYSE) and London Stock Exchange (LSE). The experimental results show that the predictive FDT algorithm and fuzzy reasoning method provides the reasonable performance for comprehensibility (no of rules), complexity (no of nodes) and predictive accuracy of WFPRs for stock market time series data.
ABSTRAK

Setelah berdekad lamanya, banyak percubaan telah dilakukan untuk meramalkan data pasaran saham masa menggunakan model statistik dan perlombongan data. Walau bagaimanapun, kebanyakan kaedah memerlukan masa latihan yang lama, biasanya keputusan yang diperolehi sukar untuk difahami, dan menghasilkan ramalan yang tidak tepat. Tambahan pula, jangkaan pedagang terhadap pasaran saham adalah dipengaruhi oleh faktor-faktor seperti situasi politik, harga minyak, situasi dunia keseluruhan, pasaran saham tempatan dan lain-lain. Oleh itu, peramalan aliran harga saham adalah amat sukar. Perlombongan data berkebolehan untuk mencari pola tersembunyi dan meramalkan arah aliran masa hadapan dan trend dalam pasaran kewangan. Di dalam penyelidikan ini, teknik klasifikasi lain bagi pepohon keputusan kabur (FDT) yang telah diubah suai, dipanggil predictive FDT digunakan untuk menggabungkan pepohon keputusan simbolik di dalam pengelasan data dengan anggaran hujah yang sesuai menggunakan perwakilan kabur. Tujuannya adalah untuk saling melengkapi kebaikan kedua-duanya : keupayaan untuk belajar daripada contoh, kefahaman pengetahuan yang tinggi dalam pepohon keputusan dan keupayaan untuk berhubung dengan maklumat tertentu berkenaan perwakilan kabur. Paling utama, peramalan bagi pepohon keputusan kabur adalah berdasarkan konsep tahap kepentingan yang menyumbang atribut kepada pengelasan. Selepas membina predictive FDT, peraturan pengeluaran pemberat (WFPRs) diesktrakkan daripada predictive FDT, dan seterusnya WFPRs yang lebih baik boleh diperolehi dengan menggunakan kaedah pemikiran kabur berasaskan persamaan. Dalam kaedah yang dicadangkan parameter pemberat boleh dinyatakan kepada setiap penyataan dalam peraturan pengeluaran kabur (FPR) dan faktor kemungkinan (CF) bagi setiap peraturan. Faktor kemungkinan dikira menggunakan beberapa pembolehubah yang penting dari pasaran saham. Kaedah pemikiran peramalan telah diuji dengan menggunakan tiga set data termasuklah KLSE, NYSE dan LSE. Hasil eksperimen menunjukkan bahawa peramalan bagi algoritma FDT dan kaedah pemikiran kabur menyediakan pencapaian yang agak baik untuk mudah difahami (bilangan peraturan), kekompleksan (bilangan nod) dan mempunyai ketepatan ramalan bagi peraturan WFPR untuk pasaran saham bersirikan data.
TABLE OF CONTENTS

CHAPTER    TITLE    PAGE
1  INTRODUCTION    1
  1.1 Overview    1
  1.2 Problem Background    2
    1.2.1 Time Series    2
    1.2.2 Support Vector Machine    3
    1.2.3 Linear and Non-linear Statistical Models    4
    1.2.4 Neural Networks    5
    1.2.5 Association Rules    5
    1.2.6 Classification    6
  1.3 Problem Statement    7
  1.4 Objectives of Research    8
  1.5 Research Scope    9
  1.6 Research Contributions    9
  1.7 Thesis Organization    10

2  LITERATURE REVIEW    12
  2.1 Introduction    12
  2.2 Time Series    12
  2.3 Stock Market Prediction    14
    2.3.1 Support Vector Machine (SVM)    15
    2.3.2 Linear and Non-Linear Statistical Models    16
    2.3.3 Artificial Neural Networks (ANNs)    17
    2.3.4 Association Rules    19
    2.3.5 Fuzzy Sets    21
2.3.6 Classification 23
2.4 Decision Tree Classification 25
  2.4.1 Crisp Decision Tree 25
  2.4.2 Fuzzy Decision Tree 26
2.5 Fuzzy Reasoning Methods 32
  2.5.1 Turkser et al.’s Approximate Analogical Reasoning Schema (AARS) 34
  2.5.2 Chen’s Function T (FT) Method 35
  2.5.3 Yeung et al.’s Equality and Cardinality (EC) Method 36
2.6 Summary 37

3 RESEARCH METHODOLOGY 39
  3.1 Introduction 39
  3.2 Operational Framework 40
    3.2.1 Problem Formulation (phase 1) 41
    3.2.2 System Development (phase 2) 41
      3.2.2.1 Data Collection and Extraction 42
      3.2.2.2 Data Cleaning and Exploration 43
      3.2.2.3 Data Engineering 44
      3.2.2.4 Algorithm Engineering 45
      3.2.2.5 Running the Data Mining Algorithm 49
    3.2.3 Implementation and Integration (phase 3) 50
      3.2.3.1 Integrated System Components 51
      3.2.3.2 System Performance Testing 51
  3.3 Summary 52

4 INDUCTIVE LEARNING OF PREDICTIVE FUZZY DECISION TREE 54
  4.1 Introduction 54
  4.2 Decision Trees Induction 54
  4.3 Fuzzy Decision Tree Induction 55
4.4 Stock Market Data Collection and Extraction

4.5 Data Cleaning using ESTEEM (Elimination of Suspicious Training Examples with Error on the Model) Method

4.6 Predictive Fuzzy Decision Tree (FDT)
   4.6.1 Centroids of Fuzzy Sets (K-Means Algorithm)
   4.6.2 Fuzzification of Numerical Number
   4.6.3 Triangular Membership Function
   4.6.4 Predictive Fuzzy Decision Tree Algorithm

4.7 Summary

5 FUZZY REASONING METHOD BASED ON SIMILARITY TECHNIQUE

5.1 Introduction

5.2 Weighted Fuzzy Production Rules (WFPRs)
   5.2.1 Weighted Fuzzy Production Rules with Single Antecedent
   5.2.2 Weighted Fuzzy Production Rules with Multiple Antecedents
   5.2.3 Transformation of WFPRs from FDT

5.3 Knowledge Parameters

5.4 Similarity-Based Fuzzy Reasoning Methods
   5.4.1 Approximate Analogical Reasoning Schema (AARS)
   5.4.2 Function T (FT) Method
   5.4.3 Degree of Subsethood (DS) Method
   5.4.4 Equality and Cardinality (EC) Method

5.5 Fuzzy Reasoning Method
   5.5.1 Similarity Measure
   5.5.2 Aggregated Weighted Average
   5.5.3 Modification Functions (MF’s)
   5.5.4 Rules Propositions
   5.5.5 Fuzzy Reasoning algorithm

5.6 Summary
6 EXPERIMENTS AND RESULTS 100
6.1 Introduction 100
6.2 Applying the Data Mining Process 101
6.3 Experimental Design 102
6.4 Testing Results and Analysis of FDT 104
6.5 Rules Extraction and Experiments 107
6.6 Analyzing Prediction Results 112
6.7 Summary 114

7 CONCLUSIONS AND RECOMMENDATIONS 115
7.1 Introduction 115
7.2 Discussion 115
7.3 Future Research 117
  7.3.1 Similarity Measures 118
  7.3.2 Predictions 118
  7.3.3 Robustness 119
7.4 Conclusions 119

BIBLIOGRAPHY 121

APPENDICES 138
APPENDIX A 138
APPENDIX B 140
APPENDIX C 142
APPENDIX D 145
APPENDIX E 149
APPENDIX F 150
<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>APPENDIX G</td>
<td>151</td>
</tr>
<tr>
<td>APPENDIX H1</td>
<td>152</td>
</tr>
<tr>
<td>APPENDIX H2</td>
<td>152</td>
</tr>
<tr>
<td>APPENDIX H3</td>
<td>153</td>
</tr>
<tr>
<td>APPENDIX H4</td>
<td>153</td>
</tr>
<tr>
<td>APPENDIX H5</td>
<td>154</td>
</tr>
<tr>
<td>APPENDIX H6</td>
<td>154</td>
</tr>
<tr>
<td>APPENDIX H7</td>
<td>155</td>
</tr>
<tr>
<td>APPENDIX H8</td>
<td>155</td>
</tr>
</tbody>
</table>

**PUBLICATIONS**

<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>156</td>
</tr>
</tbody>
</table>
### LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Summary of analytical comparison of fuzzy decision trees</td>
<td>31</td>
</tr>
<tr>
<td>2.2</td>
<td>FPR’s with single antecedent</td>
<td>34</td>
</tr>
<tr>
<td>2.3</td>
<td>FPR’s with a multiple antecedents</td>
<td>35</td>
</tr>
<tr>
<td>2.4</td>
<td>Analysis of existing similarity-based methods</td>
<td>37</td>
</tr>
<tr>
<td>4.1</td>
<td>A comparison between the fuzzy decision tree and the crisp</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>decision tree</td>
<td></td>
</tr>
<tr>
<td>4.2</td>
<td>20 real time (5-Minute Bars) examples of stock market data</td>
<td>62</td>
</tr>
<tr>
<td>4.3</td>
<td>Change in stock trading prices for every 5 minutes of time</td>
<td>63</td>
</tr>
<tr>
<td></td>
<td>series stock market</td>
<td></td>
</tr>
<tr>
<td>4.4</td>
<td>After training real time examples of stock market with fuzzy</td>
<td>73</td>
</tr>
<tr>
<td></td>
<td>representation</td>
<td></td>
</tr>
<tr>
<td>6.1</td>
<td>Summary of the experiments for I = FuzzyID3, II =</td>
<td>105</td>
</tr>
<tr>
<td></td>
<td>Wang et al., and III = Predictive FDT</td>
<td></td>
</tr>
<tr>
<td>6.2</td>
<td>An example of 13 extracted WFPRs from FDT in single</td>
<td>108</td>
</tr>
<tr>
<td></td>
<td>experiment.</td>
<td></td>
</tr>
<tr>
<td>6.3</td>
<td>Testing results for significant rules using prediction in</td>
<td>111</td>
</tr>
<tr>
<td></td>
<td>change method</td>
<td></td>
</tr>
<tr>
<td>6.4</td>
<td>The mean squared error between the actual values and the</td>
<td>112</td>
</tr>
<tr>
<td></td>
<td>predicted results for 3-case study stock price indexes is</td>
<td></td>
</tr>
<tr>
<td></td>
<td>up to 3 months</td>
<td></td>
</tr>
</tbody>
</table>
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>FIGURE NO.</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Research operational frame work</td>
<td>41</td>
</tr>
<tr>
<td>3.2</td>
<td>System development process</td>
<td>47</td>
</tr>
<tr>
<td>4.1(a)</td>
<td>Fuzzy decision tree using Fuzzy ID3 heuristic (Umanol et al., 1994)</td>
<td>60</td>
</tr>
<tr>
<td>4.1(b)</td>
<td>Fuzzy decision tree by Yuan and Shaw’s (1995) heuristic</td>
<td>60</td>
</tr>
<tr>
<td>4.1(c)</td>
<td>Fuzzy decision tree by using Wang et al., (2001) heuristic</td>
<td>60</td>
</tr>
<tr>
<td>4.2</td>
<td>The Esteem method of outlier removal</td>
<td>64</td>
</tr>
<tr>
<td>4.3</td>
<td>Procedure for predictive fuzzy decision tree classification method</td>
<td>66</td>
</tr>
<tr>
<td>4.4</td>
<td>K-Means algorithms to find centers of the equal patterns</td>
<td>67</td>
</tr>
<tr>
<td>4.5</td>
<td>Centroids between two clusters</td>
<td>68</td>
</tr>
<tr>
<td>4.6(a)</td>
<td>Linguistic terms for oil price</td>
<td>71</td>
</tr>
<tr>
<td>4.6(b)</td>
<td>Linguistic terms for change in close price</td>
<td>71</td>
</tr>
<tr>
<td>4.6(c)</td>
<td>Linguistic terms for local stock market</td>
<td>71</td>
</tr>
<tr>
<td>4.6(d)</td>
<td>Linguistic terms for primary signal</td>
<td>71</td>
</tr>
<tr>
<td>4.7</td>
<td>Fuzzy Decision Tree by using predictive FDT algorithm to train table 4.2</td>
<td>77</td>
</tr>
<tr>
<td>5.1</td>
<td>Disconsistency measure</td>
<td>85</td>
</tr>
<tr>
<td>5.2</td>
<td>More or Less form</td>
<td>96</td>
</tr>
<tr>
<td>5.3</td>
<td>Membership value reduction form</td>
<td>96</td>
</tr>
<tr>
<td>6.1</td>
<td>Experimental design for stock market prediction by using 5 years information (Jan 1, 1999 to Dec 30, 2004) from KLSE, NYSE and LSE.</td>
<td>103</td>
</tr>
</tbody>
</table>
6.2(a) Experimental results of FuzzyID3, Wang FDT, and Predictive FDT for KLSE

6.2(b) Experimental results of FuzzyID3, Wang FDT, and Predictive FDT for NYSE

6.2(c) Experimental results of FuzzyID3, Wang FDT, and Predictive FDT for LSE

6.3 The prediction process for Jan 1 to March 30, 2005 using the historical stock market data from (Jan 1, 1999 to Dec 30, 2004).

6.4 (a) Core Points of the rules for three months prediction

6.4 (b) Results using core points of rules for three months predictions from (Jan 1, 1999 to Dec 30, 2004)

6.5(a) Three Months (Jan 1 to March 31, 2005) prediction of Telekom Index from KLSE using random walk and predictive FDT

6.5(b) Three Months (Jan 1 to March 31, 2005) prediction of DowChemical Index from NYSE using random walk and predictive FDT

6.5(c) Three Months (Jan 1 to March 31, 2005) prediction of ABN-Amro Index from LSE using random walk and predictive FDT
LIST OF SYMBOLS AND ABBREVIATIONS

$P_o$ - Open price

$P_c$ - Close price

$V$ - Volume

$C_{P_o}$ - Change in open = {Low, Average, High},

$C_{P_c}$ - Change in close = {Low, Average, High},

$C_V$ - Change in volume = {Low, Average, High},

$Entr_i^{(k)}$ - Minimum classification information-entropy for each attribute ($k$)

$Ambig_i^{(k)}$ - Minimum classification ambiguity for each attribute ($k$)

$Th = \lambda_o$ - Threshold value

$CF = \mu$ - Certainty factor

$W$ - Weight for single and multiple antecedent

$Gw$ - Weighted average

$AG_w$ - Aggregated weighted average

$x_i^{(j)}$ - Data point

$\mathbf{c}_j$ - Cluster centre

$k$ - Linguistic terms

$S$ - Nonleaf node

$\nabla$ - Symmetrical difference of $A$ and $A'$
GR - Gain ratio
A - Antecedent (pattern)
A' - An observation (fact) of A
C - Consequent of a rule
C' - Result of consequent
μₐ(x) - Degree of membership x by linguistic term A
FDT - Fuzzy Decision Tree
WFPR’s - Weighted Fuzzy Production Rules
FPR’s - Fuzzy Production Rules
KDD - Knowledge Discovery in Databases
AARS - Approximate Analogical Reasoning Schema
SMP - System Marginal Price
FT - Function T
DS - Degree of Subsethood
IC - Inclusion Cardinality
EC - Equality and Cardinality
PR - Predictive Reasoning
KLSE - Kuala Lumpur Stock Exchange
NYSE - New York Stock Exchange
LSE - London Stock Exchange
CF - Certainty Factor
DM - Degree of Measure
SVM - Support Vector Machine
SVR - Support Vector Regression
PCD - Prediction of change in direction
1.1 Overview

With the increase of economic globalization and evolution of information technology, financial time series data are being generated and accumulated at an unprecedented pace. As a result, there has been a critical need for automated approaches to effective and efficient utilization of massive amount of financial data to support companies and individuals in strategic planning and investment for decision-making. In this chapter, some problems are discussed that investors are facing during crucial part of their decision process for the selection of real time stock market to invest in. These problems are highlighted by analyzing the existing data mining and statistical techniques for stock market predictions. Among the variety of existing techniques for stock market prediction, Support Vector Machine (SVM), Linear and Non-linear models, Neural Networks, Association Rules, and Classification have found to be good techniques for stock market predictions. However, instead of incredible results for time series stock market prediction, these techniques still have some serious drawbacks to handle time series stock market data. These problems are highlighted in the following section.
1.2 Problem Background

The stock market is a rather complicated system, and good predictions for its developments are the key to successful trading. Traders must predict stock price movements in order to sell at top range and to buy at bottom range. As stock trading is a very risky business (Torben and Lund, 1997), it is necessary to evaluate the risks and benefits before entering into any trading. The key to realize high profits in stock trading is to determine the suitable trading time when the risk of trading should be minimum. Many attempts have been made for meaningful prediction of stock market by using data mining and statistical techniques like Time Series (Agrawak R. et al., 1995a; Breidt, 2002), Support Vector Machine (Alan Fan et al., 2001; Haiqin, 2002), Neural Networks (Xiaohua et al., 2003; Raymond, 2004), Linear and Non-linear models (Weiss, E. 2000; Chinn et al., 2001), Association Rules (Ke and Yu 2000; Sarjon and Noor, 2002a) and Classification (Agrawal R. et al., 2000; Han, J and Pei, 2000). However, these techniques to predict stock market real time data are yet to be achieved good classifiers (model). The following subsections present some general problem background regarding these techniques to predict stock market.

1.2.1 Time Series

Investors are facing with an enormous amount of stocks in the market. The meaningful prediction of time series data is one of the fast growing research areas in field of financial engineering. Investors in the market want to maximize their stock return by buying or selling their investments at an appropriate time. Since stock market data are highly time-variant and are normally in a nonlinear pattern, predicting the future trend (i.e., rise, decrease, or remain steady) of a stock is a challenging problem. Agrawak R. et al., (1995a; 1995b) introduces different similarity queries on time series data to find the behaviors of selling or buying patterns of different stocks. In queries of this type, approximate rather than exact matching is required. Recently some researchers
use data mining techniques for attempting to index, cluster, classify and mine association rules from increasing massive sources of time series data (Iztok et al., 2002; Yimin and Dit-Yan, 2004).

For example Keogh and Pazzani (1998) introduce a new scalable time series classification algorithm. Das et al., (1998) attempt to show, how association rules can be learned from time series. Han et al. (1999) investigate time series databases for periodic segments and partial periodic patterns using data mining methods. All these algorithms that operate on time series data need to compute the similarity between patterns by using Euclidean distance and Euclidean distance sensitive to the absolute offsets of time sequences.

Iztok et al., (2002) present an algorithm for matching sequences with the set of time series. The matching algorithm proposed by them is not enough to be effectively used for other domains including stock market data, sensor data in engineering environments, and medical measurements. In another attempt, the model-based clustering method (Yimin and Dit-Yan, 2004) can incorporate prior knowledge more naturally in finding the correct number of clusters. However, the clustering performances of model-based clustering method degrade significantly when the underlying clusters are very close to each other. Another problem is that it is not designed for modeling the differences in trend of the time series.

1.2.2 Support Vector Machine

Alan Fan et al., (2001) use Support Vector Machine (SVM) to predict stock market. The SVM is a training algorithm for learning classification and regression rules from data (Osuna, 2001). However the predictive accuracy of SVM achieved by Alan Fan et al., (2001) in stock market is relatively lower than other classification applications (Jingtao, 1997; Haiqin, 2002). Also the existing relationship between the future stock
returns and its accounting information, one would expect it to be a weak relationship. Support Vector Regression (SVR) is the extended form of SVM that can be applied for the prediction of financial time series data (Jingtao, 1997; Haiqin, 2002). In financial data, due to the embedded noise, one must set a suitable margin in order to obtain a good prediction (Haiqin, 2002). Haiqin et al., (2002) has extended the standard SVR with adaptive margin and classified into four cases. The model proposed by them requires the adaptation of the margin width and the degree of asymmetry and no exact algorithm for such margin setting has been introduced. Some researchers (Mukherjee, 1997; Chih-chung and Chih-jen, 2001) try to set this margin with 0 or a very small value, but SVR is still insensitive and non-adaptive to the input data. This may result in less than optimal performance in the testing data while it obtains a good result on the training data.

1.2.3 Linear and Non-linear Statistical Models

The Autoregressive Conditional Heteroskedasticity (ARCH) class of models (Chinn et al., 2001; Pierre and Sébastien, 2004) has become a core part of empirical finance. ARCH is a nonlinear stochastic process, where the variance is time-varying, and a function of the past variance. ARCH processes have frequency distributions which have high peaks at the mean and fat-tails, much like fractal distributions. The issue of forecasting with ARCH models has been discussed by (Koenker and Zhao, 1996). However, the procedures developed are restricted to a limited class of ARCH models, often do not take account of parameter uncertainty, and often have questionable finite sample properties. Another statistical model is Auto Regressive Integrated Moving Average (ARIMA) models have been already applied to forecast commodity prices (Weiss, E., 2000; Chinn et al., 2001). However, since the ARIMA models are linear and most real world applications involve non-linear problems, this introduces a limitation in the accuracy of the predictions generated (Ferreira et al., 2004).
1.2.4 Neural Networks

Artificial Neural Networks (ANN) techniques that have been widely used for load forecasting are now used for price prediction (Nicolaisen, 2000; Hippert, 2001). Using neural networks to model and predict stock market returns has been the subjects of recent empirical and theoretical investigations by academics and practitioners alike (Xiaohua et al., 2003; Raymond, 2004). Xiaohua et al., (2003) investigate whether trading volume can significantly improve the forecasting performance of neural networks, or whether neural networks can adequately model such nonlinearity. Such types of neural networks cannot handle the nonlinearity between stock return and trading volume. Raymond, (2004) introduce a feasible and efficient solution (iJADE Stock Advisor) for automatic intelligent agent-based stock prediction. However, the model appeared highly sensitive to the training parameters and also these method have various numbers of hidden neurons are tested in the experiments and no significant improvements appear, it may be due to not finding the optimal architecture and available training methods.

1.2.5 Association Rules

Association rule mining uncovers interesting correlation patterns among a large set of data items by showing attribute-value conditions that occur together frequently. It has been applied successfully in a wide range of business predicting problems (Ke et al., 2000; Sarjon and Noor, 2002). Pasquier and Bastide (1997) proposed an algorithm, called a close, for finding frequent closed itemsets and their support in a stock market database. However, a close method is costly when mining long patterns or with low minimum support threshold in large database like stock market and also cannot generate association rules at higher levels. In another attempt, typical algorithms for discovering frequent itemsets in stock market by using association rules operate in a bottom-up, breadth-first search direction (Dao-I and Kedem, 2002). The computation starts from
frequent 1-itemsets (the minimum length frequent itemsets) and continuous until all maximal (length) frequent itemsets are found. During the execution, every frequent itemset is explicitly considered. Such algorithms perform well when all maximal frequent itemsets are short. However, performance drastically depreciates when some of the maximal frequent itemsets are long.

1.2.6 Classification

Among discovering different kinds of knowledge from large databases, classification has been recognized as an important problem in data mining (Agrawal R. et al., 2000; Hung-Ju and Chun 2002). Classification is one kind of data mining technique to identify essential features of different classes based on a set of training data and then classify unseen instances into the appropriate classes. Many popular classification techniques, Apriori-like approach (Agrawal R. et al., 2000; Xi Ma., 2004), FP-growth (Han, J. and Pei, 2000; Yabo et al., 2002), Naive Bayes Classifier (Hung-Ju and Chun, 2002), Decision Theoretic Model (Elovici and Braha, 2003) Information Network (Last and Maimon, 2004) have been used as data mining problems. These approaches are still not suitable where the real databases contains all records, huge space is required to serve the mining, and large applications need more scalability.

Most trading practice adopted by financial analysts relies on accurate classification prediction of the price levels of financial instruments. However, some recent studies have suggested that trading strategies guided by forecasts on the direction of the change in price level are more effective and may generate higher profits. Wu and Zhang (1997) investigate the predictability of the direction of change in the future spot exchange rate. In another study Aggarwal R. and Demaskey (1997) find that the classification performance of cross hedging improves significantly rates can be predicted. O’Connor et al., (1997) conduct a laboratory-based experiment and conclude that individuals show different tendencies and behaviors for upward and downward series. The findings in
these studies are reasonable because accurate point estimation, as judged by its deviation from the actual observation may not be a good predictor of the direction of change in the instruments price level.

1.3 Problem Statement

In recent years, there have been a growing number of studies looking at the direction or trend of movements of various kinds of statistical and data mining techniques (such as Mark et al., 2000; Elovici and Braha, 2003; Zhang and Zhou, 2004). However, as discussed in problem background these techniques have some limitations to handle time series and huge stock market data. Also none of these studies provide a comparative evaluation of different classification techniques regarding their ability to predict the sign of the index return. Among the variety of data mining techniques, classification has found scale well, run fast, and produce highly interpretable results.

In this research, our main focus on the development of classification data mining model to identifying a better decision making classifiers for stock market prediction. In classification, decision tree with fuzzy sets have been used to building classification models for predicting classes for unseen records. The predictive FDT algorithm and similarity-based fuzzy reasoning method are used to answer the following research questions

- How classification data mining techniques are able to uncover hidden patterns and predict future trends and behaviors in financial markets?
- How the trader’s expectations can be fulfilled to handle uncertain factors, including political situation, oil price, overall world situation, local stock markets etc?
How the optimal predictive FDT can be constructed that has better comprehensibility (no of rules), less complexity (no of nodes) and better learning accuracy?

How similarity-based fuzzy reasoning method can improve the learning accuracy of generated weighted fuzzy production rules for the prediction of stock market data?

1.4 Objectives of Research

The main objectives of this research is to build accurate model by using classification and fuzzy sets in order to make the methods accessible, “user-friendly”, and prepared for the broader population of economists, analysts, and financial professionals. To deal with these tasks first predictive FDT are constructed and then WFPR,s are extracted from predictive FDT for predicting stock market. The main objectives of this research are as follows:

- To develop and enhance data mining algorithms for extracting the pattern of knowledge from stock market database.
- To construct predictive FDT for the classification of stock market index that has better performance of comprehensibility (no of rules), lower complexity (no of nodes) and improves the learning accuracy.
- To mine the more significant WFPR’s among the extracted rules from predictive FDT by using similarity-based fuzzy reasoning method.
- To evaluate and compare the performance of predictive FDT and similarity-based fuzzy reasoning method with the existing standard methods in order to find the strength and weaknesses of proposed algorithm.
1.5 Research Scope

This research focuses on the enhancement of data mining algorithms in financial application. In data mining algorithms, classification model has been developed to extract the weighted fuzzy production rules for time series stock market prediction. The scope of this research covers the following points:

- Reviews and comparisons of the existing data mining methods to predict future trends and behaviors in financial markets.
- Analyze and evaluate the performance of the predictive FDT algorithm by using three stock exchanges, i.e. KLSE, LSE, and NYSE.
- Test and analyze the predictive FDT algorithms on the chaotic and complex nature of intraday stock market data from 100 different associate stocks.
- Analyze and evaluate the results of predictive FDT and similarity-based fuzzy reasoning method for maximum three months prediction of real time stock market.

1.6 Research Contributions

In this research, the data mining and artificial intelligence technique are used to build classification model for stock market prediction. First, the predictive FDT classifier is constructed and then WFPR’s are extracted from predictive FDT for stock market prediction. Some points of major contributions of this research are described as follows:

- The enhancement of classification based data mining algorithms to predict trends and behaviors in financial market.
The construction of predictive fuzzy decision tree classification, that provides the reasonable performance of the parameters comprehensibility (no of rules), complexity (no of nodes) and predictive accuracy.

Careful assignment of some uncertain factors including oil price, local stock market, natural disasters etc, for the classification of stock market index.

In predictive reasoning method, the conventional fuzzy production rules are enhanced to assign a weighting factor to each proposition in the antecedent parts to have various degree of relative importance with respect to the same consequent.

1.7 Thesis Organization

The general research background is given in this chapter. Next chapter presents the brief overview of financial application with description of previous works used on data mining and statistical techniques. Chapter 3 discuss, when we have together defined the problem to solve, in the remaining steps we must collect the relevant data if it does not already exist, clean the data, engineer the data to be maximally useful with the problem at hand, engineer a mining algorithm, run the mining algorithm and explain the results to the expert. In chapter 4 predictive fuzzy decision trees is constructed. In this chapter, first 3 steps are important to implement before applying the predictive FDT algorithm. In first step, K-means clustering method is used to find the center for every fuzzy set and calculate the degree of membership by applying triangular membership function. Finally construct predictive FDT by using fuzzy decision tree algorithm. In chapter 5, similarity-based fuzzy reasoning method is used for mining weighted fuzzy production rules. In this chapter, decision has been made about the more suitable WFPR’s for prediction of stock market. In Chapter 6, the experiments and results are presented that are produced from predictive FDT and similarity-based fuzzy reasoning method. The accuracy of extracted WFPR’s are also compared with the standard method of stock market
prediction. Finally in chapter 7, the conclusions of predictive fuzzy decision tree and extracted weighted fuzzy production rules have been presented. Some challenges and emerging trends are identified for future research also suggested in this chapter.
BIBLIOGRAPHY


