Abstract

Portfolio optimization is the process of making investment decisions on holding a set of financial assets to meet various criteria. A variety of investment assets around the world make this multi-faceted decision problem very complicated. Econometric and statistical models as well as machine learning and data mining techniques have been used by many researchers and analysts to propose heuristic solutions for portfolio optimization. However, a literature review shows that the existing models are still not practical as they do not always perform better than even the naïve strategy of investing in all available assets in the market. The methodology proposed in this thesis is an alternative heuristic solution to help investors make stock investment decisions through a semi-automated process. The proposed solution is based on the fact that the investment decision cannot be fully automated because investors’ preferences that are the key factors in making investment decision, vary among different people. For this purpose, a semi-automated framework called SMPOpt (Stock Market Portfolio Optimizer) has been designed and implemented. In the proposed framework, the goal is to learn from the historical fundamental analysis of companies to discover the optimum portfolio by considering investors’ preferences. The Portfolio optimization problem is formulated and broken down into steps to be able to apply data mining techniques such as Clustering and Ranking, and Social Network Analysis. Some of these techniques are customized based on the temporal behaviour of financial datasets. For instance, the ranking algorithm based on Support Vector Machine (SVMRank) is modified and a new algorithm called Time-Series SVMRank is proposed. A comprehensive experimental study has been conducted using the real stock exchange market datasets from the past recent decades to evaluate the
proposed portfolio optimization solution. The obtained results confirmed the strength of the proposed methodology.
Acknowledgements

Working on my PhD thesis has been a wonderful learning experience in my life and I am indebted to many people in this unforgettable experience. First of all, I am deeply grateful to my advisor Dr. Reda Alhajj for his guidance, patience, support and providing me with an excellent atmosphere for doing research. I would like to thank my co-advisor Dr. Victoria Mitchell for her excellent guidance on financial aspects of my PhD career. She has been a steady influence throughout this work.

I am also grateful to my other supervisory committee member, Dr. Jon Rokne, as well as examiners, Dr. Alfred Lehar and Dr. Alex Thomo. My sincere thanks also goes to Dr. Johannes Fürnkranz for giving me the summer visiting opportunity in his group in Darmstadt university.

The Natural Sciences and Engineering Research Council of Canada (NSERC) are gratefully acknowledged for their support in the form of graduate scholarship.

I am grateful to my counselor Dr. J.J. Miles who has influence on my spiritual attitude.

I would also like to thank my friend Keivan for all his help, valuable advice and support in this career.

Last but not the least, I would like to thank my parents Haideh and Sadegh, for giving birth to me at the first place and supporting me spiritually throughout my entire life. My deepest gratefulness goes to my elder sister Masi and her husband Mehrdad for all their support throughout my PhD.
To Tranquility and Peace
# Table of Contents

Abstract ...................................................................................................................... ii
Acknowledgements ....................................................................................................... iv
Table of Contents ......................................................................................................... vi
List of Tables ................................................................................................................ viii
List of Figures and Illustrations .................................................................................... ix
List of Abbreviations ................................................................................................... xii

CHAPTER ONE: INTRODUCTION .............................................................................. 1
  1.1 Problem Definition ............................................................................................. 1
  1.2 Motivation .......................................................................................................... 3
  1.3 Contribution ...................................................................................................... 7
  1.4 Thesis Organization ......................................................................................... 12

CHAPTER TWO: BACKGROUND AND RELATED WORKS ................................... 13
  2.1 Related Financial Theories ............................................................................... 13
  2.2 Related Computer Science based Heuristic Approaches .................................. 16
  2.3 Similar tools to SMPOpt .................................................................................. 18
  2.4 Approaches used in SMPOpt .......................................................................... 19

CHAPTER THREE: STOCK PORTFOLIO OPTIMIZATION (PROBLEM DEFINITION) ...................................................................................................... 22
  3.1 Identifying the Effective Factors on Stock Portfolio Optimization ..... 22
  3.2 Definition of Stock Portfolio Optimization ...................................................... 25

CHAPTER FOUR: FIRST SOLUTION IN SMPOPT: A SEMI-AUTOMATED PORTFOLIO OPTIMIZER ................................................................................... 26
  4.1 Overview of the Proposed Semi-Automated Process ........................................... 26
  4.2 Four Phase Portfolio Optimization Approach ....................................................... 29
    4.2.1 Stock Social Network Construction ............................................................ 29
      4.2.1.1 Other Stock Performance Measurements ............................................ 35
      4.2.1.2 Time Series Correlation ...................................................................... 37
    4.2.2 Stock Clustering ....................................................................................... 38
      4.2.2.1 Mapping Propensity for Risk to Number of Clusters ....................... 40
      4.2.2.2 Clustering Algorithms ......................................................................... 42
    4.2.3 Stock Ranking .......................................................................................... 45
      4.2.3.1 Stock Ranking Problem Formulation .................................................. 47
      4.2.3.2 Stock Ranking using SVMRank .............................................................. 50
      4.2.3.3 Time Series SVMRank ....................................................................... 52
    4.2.4 Portfolio Construction ............................................................................... 54
  4.3 Financial Data Pre-Processing ............................................................................ 56
    4.3.1 Stock Feature Selection ............................................................................ 57

CHAPTER FIVE: SECOND SOLUTION IN SMPOPT: MANAGED PORTFOLIO RECOMMENDATION .......................................................... 61
  5.1 Investor’s Feature Extraction .............................................................................. 61
5.2 Social Network Construction ................................................................. 62
5.3 Investment Recommendation System ..................................................... 63
  5.3.1 Experts Clustering ............................................................................. 64
  5.3.2 Non-Professional Investors Classification .......................................... 64
  5.3.3 Finding the Representative Expert in a Community ......................... 65

CHAPTER SIX: SMPOPT EVALUATION THROUGH EXPERIMENTAL STUDIES ............................................................... 69
6.1 Performance Metrics .............................................................................. 71
  6.1.1 Sharpe Ratio .................................................................................... 71
  6.1.2 Future Return .................................................................................. 71
  6.1.3 NDCG ............................................................................................ 72
6.2 Benchmarks .......................................................................................... 72
  6.2.1 All Market Portfolio (Maximum Diversification) ............................... 72
  6.2.2 Single Criterion Stock Ranking ....................................................... 73
  6.2.3 Static Diversification ....................................................................... 73
  6.2.4 Managed Portfolio .......................................................................... 73
6.3 Experiment Setups and Results ............................................................... 74
  6.3.1 Experiment 1[67] ............................................................................ 74
  6.3.2 Experiment 2 .................................................................................. 77
  6.3.3 Experiment 3 .................................................................................. 82
  6.3.4 Experiment 4 .................................................................................. 84
  6.3.5 Experiment 5 .................................................................................. 88
  6.3.6 Experiment 6 .................................................................................. 90
  6.3.7 Experiment 7[68] .......................................................................... 93
  6.3.8 Experiment 8 .................................................................................. 96
6.4 Threats to Validity ................................................................................ 99

CHAPTER SEVEN: SMPOPT IMPLEMENTATION ................................................................. 102
7.1 SMPOpt Architecture .......................................................................... 102
7.2 Proposed Process in SMPOpt ............................................................... 105

CHAPTER EIGHT: SUMMARY, CONCLUSION AND FUTURE DIRECTIONS .... 109
8.1 Summary and Conclusion ................................................................... 109
8.2 Future Directions ............................................................................... 110

REFERENCES ............................................................................................. 113
### List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>List of Experimental Studies</td>
<td>70</td>
</tr>
<tr>
<td>2</td>
<td>Synthetic Eigenvector Centrality and Expected Return of Dow Jones stock</td>
<td>76</td>
</tr>
<tr>
<td>3</td>
<td>p-values of Paired T-Test for Sharpe Ratios in Exp 1</td>
<td>77</td>
</tr>
<tr>
<td>4</td>
<td>p-value of paired t-tests in Exp 2</td>
<td>80</td>
</tr>
<tr>
<td>5</td>
<td>Internal Factors used in Exp 3, 4, 6 and 8</td>
<td>83</td>
</tr>
<tr>
<td>6</td>
<td>p-Values of t-test hypothesis testing in Exp 4</td>
<td>87</td>
</tr>
<tr>
<td>7</td>
<td>NDCG improvement by feature selection</td>
<td>89</td>
</tr>
<tr>
<td>8</td>
<td>External Factors used in Exp 8 [69]</td>
<td>97</td>
</tr>
</tbody>
</table>
List of Figures and Illustrations

Figure 1 - Block Diagram of Thesis Organization .........................................................12
Figure 2 – Effective Factors on Stock Portfolio Optimization ...........................................24
Figure 3 - Component Diagram of Semi-Automated Portfolio Optimizer in SMPOpt ....27
Figure 4 - Inputs and Outputs of Network Construction Phase ......................................30
Figure 5 - List of Stocks selection GUI in SMPOpt..........................................................30
Figure 6 - Data Import Setting GUI in SMPOpt .............................................................31
Figure 7 - Network Setting GUI in SMPOpt .................................................................31
Figure 8 - Social Network of the Dow Jones Index .......................................................35
Figure 9 - Weights function in TS_SVMRank ...............................................................38
Figure 10 - Inputs and Outputs of Stock Clustering Phase .............................................39
Figure 11 - Stock Clustering Setting GUI in SMPOpt ....................................................39
Figure 12 - Example of mapping between propensity for risk and optimum number of clusters .....................................................................................................................42
Figure 13 - K-means clustering results for sample Dow Jones Index ..............................43
Figure 14 - Louvain clustering results for sample Dow Jones Index ...............................45
Figure 15 - Inputs and Outputs of Stock Ranking Phase ...............................................46
Figure 16 - Data Import Menu in SMPOpt ...................................................................46
Figure 17 - List of Internal and External Factors Selection GUI in SMPOpt ...................47
Figure 18 - Example of how weight vector \( \tilde{w} \) rank 5 stocks ...................................50
Figure 19 - Stock Ranking and Return Setting GUI in SMPOpt ......................................53
Figure 20 - Within Cluster Ranking of the Dow Jones Index .......................................54
Figure 21 - Portfolio Setting GUI in SMPOpt ...............................................................55
Figure 22 - Portfolio of Weighted Dow Jones Stocks ....................................................56
Figure 23 - Example ranking training and test sets .........................................................58
Figure 46 - Sample User Profile File in SMPOpt .............................................................. 104
Figure 47 - Sample Code in SMPOpt .............................................................................. 105
Figure 48 - Main Screen of SMPOpt .............................................................................. 106
Figure 49 - Semi-automated Portfolio Constructed in SMPOpt ................................. 107
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AI</td>
<td>Artificial Intelligence</td>
</tr>
<tr>
<td>APT</td>
<td>Arbitrage Pricing Theory</td>
</tr>
<tr>
<td>ARIMA</td>
<td>Autoregressive Integrated Moving Average</td>
</tr>
<tr>
<td>CAPM</td>
<td>Capital Asset Pricing Model</td>
</tr>
<tr>
<td>DCG</td>
<td>Discounted Cumulative Gain</td>
</tr>
<tr>
<td>EMH</td>
<td>Efficient Market Hypothesis</td>
</tr>
<tr>
<td>GA</td>
<td>Genetic Algorithm</td>
</tr>
<tr>
<td>GPD</td>
<td>Gross Domestic Product</td>
</tr>
<tr>
<td>GICS</td>
<td>Global Industry Classification Standard</td>
</tr>
<tr>
<td>GUI</td>
<td>Graphical User Interface</td>
</tr>
<tr>
<td>IDCG</td>
<td>Ideal Discounted Cumulative Gain</td>
</tr>
<tr>
<td>NDCG</td>
<td>Normalized Discounted Cumulative Gain</td>
</tr>
<tr>
<td>NN</td>
<td>Neural Network</td>
</tr>
<tr>
<td>NYSE</td>
<td>New York Stock Exchange</td>
</tr>
<tr>
<td>SMPOpt</td>
<td>Stock Market Portfolio Optimizer</td>
</tr>
<tr>
<td>SNA</td>
<td>Social Network Analysis</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machine</td>
</tr>
<tr>
<td>SVMRank</td>
<td>Support Vector Machines for Ranking</td>
</tr>
<tr>
<td>S&amp;P</td>
<td>Standard &amp; Poor’s</td>
</tr>
<tr>
<td>TS_SVMRank</td>
<td>Time Series Support Vector Machines for Ranking</td>
</tr>
</tbody>
</table>
Chapter One: Introduction

The realm of the stock market has always been appealing to individuals because of its beneficial potential. Finding an appropriate set of stocks for investment to ultimately gain more return and face less risk, compared to other selections, attracts many people, whether domain experts or not. In stock markets, companies from various economic sectors (including Energy, Materials, Industry, Consumer Discretionary, Consumer Staples, Health Care, Financial, Information Technology, Telecommunication Services and Utilities) sell some portion of their institutions’ stocks. Individuals can have a comprehensive aggregate overview of all available stocks to buy, and consequently, can reach a reliable decision on how much money to invest in which stocks. The potential benefit of the stock market is attracting more people to invest.

1.1 Problem Definition

A variety of investment instruments around the world make the investment decision process very complicated and challenging in which different factors need to be considered. In other words, making investment decisions on various available instruments is a multi-faceted decision problem and thus a challenging task. As a solution, more automation is required through the process to replace tedious manual investigations. Among various investment instruments such as company bonds, stocks and derivatives, this thesis is limited to stock market analysis.

In finance, a collection of assets held by an investor is called portfolio. Portfolio optimization is the process of making investment decisions on holding a set of financial assets to meet various criteria. In other words, portfolio optimization is the decision process of asset selection and weighting, such that the collection of assets satisfies an investor’s objectives. The stock portfolio optimization process involves forecasting the performance and the volatility of
stocks as well as models for using these predictions in order to obtain a portfolio that suits the investor’s preference profile.

In financial investments, it is important for investors to manage the level of risk to which they subject themselves while searching for portfolio return [1]. In general, investment opportunities that offer higher returns also entail higher risk [1]. Therefore, there is always a trade-off between risk and return in the investment decision process. Financial theories define the investment risk in a way that it can be measured and then relate the measurable risk to the level of return that can be expected from the investment [1]. The trade-off between risk and return, introduced in modern portfolio theory, is an important subject in making investment decisions and thus should be considered by any automated analysis.

In addition, several researchers have used the criterion of a firm’s health based on financial indicators, a.k.a., fundamental analysis in the investment analysis. Fundamental analysis is the study of the sector and company financial indicators to determine the value of a stock. It aims to determine the financial health of the company, by useful ratios, based on the company’s financial statements [2, 3]. Each stock instance can be described by a set of features which represent important financial information regarding the company that the stock represents.

To deal with the portfolio optimization problem, several researchers concentrate on the development of models which, in addition to the two basic criteria of return and risk, also consider other equally important criteria derived from fundamental analysis, as well as the investor’s preferences and policies [2]. Subsequently, the portfolio optimization problem has been defined as finding the optimum portfolio of assets based on multi objectives such as return, risk and other financial criteria.
The risk/return tradeoff underlies the diversification concept, such that portfolio risk can be reduced by combining assets whose returns are weakly correlated. The idea of diversification is to reduce risk by investing in a range of stocks. If the selected stocks have no or little similarity according to their trends, a portfolio consisting of these stocks will face less risk than the average risk.

In general, a solution for stock portfolio optimization involves building models from past available data to forecast the performance and the volatility of assets in future, in order to obtain the optimum portfolio for an individual's preferences. Financial theorists and investors have been dealing with this issue for many years. For their support, investors have access to a wide variety of historical financial indicators about available stocks.

1.2 Motivation

In the traditional investment methodology, each instrument was investigated thoroughly as an individual. The existence of a variety of investment instruments around the world makes the traditional methodology not efficiently applicable today. Further, investigating each instrument individually, without considering their correlations, is not sufficient for making investment decisions [4, 5]. In traditional investment decision approaches, each available stock is investigated individually to maximize the expected return of the portfolio as much as possible. There were two main shortcomings to the traditional approaches: (1) they only take individual instruments into consideration (independent of their correlation) in order to construct the portfolio; and (2) their main objective is to maximize the expected return without considering the concept of investment risk [4, 5].

In the 1950s Harry Markowitz proposed the modern portfolio theory, and since then studying the correlation between instruments and the trade-off between return and risk of the
investment have been taken into consideration by researchers [4]. In his seminal article on portfolio selection, Markowitz (1952) outlined the effects of correlation on asset risk and return concluding that the portfolio risk associated with asset covariance could be reduced through diversification across industries [4]. He demonstrated that the portfolio risk comes from the covariance of the assets making up the portfolio. Therefore, an optimum portfolio could be built from individual stocks that have good performance and do not correlate positively to each other.

Since Markowitz’s revelation, financial theorists and investment strategists have struggled with how best to enact a diversification strategy amid unpredictable market behaviour. They attempt to maximize the expected return from a collection of assets, for a given amount of portfolio risk. Conversely, investors seek to minimize portfolio risk for a given level of expected return. Derived from the tenets of modern portfolio theory, several other financial models have been proposed that assume the investment decision problem is simple and solvable. These assumptions make the solutions impractical [5]. As an example one unrealistic assumption is strong Efficient Market Hypothesis (EMH). Based on this hypothesis, all the required information for investment decision is available in the market itself, and thus all the investors make the same forecasts concerning the assets.

Although there exist several financial theories and approaches that deal with the issue of return and risk, a significant obstacle, which still remains, is to apply those theories in the real world since it is sometimes an unattainable task to complete. Several approaches suggested for portfolio management by financial advisors are based on inapplicable assumptions such as EMH and normal data distribution. Therefore, a main impediment is to apply those approaches in the real world, especially when non-experts want to utilize them to come to a conclusion.
A great deal of efforts has been devoted to developing systems for predicting stock returns. However, limited success has been achieved. It is believed that the main reason is that the structural relationship between an asset price and its determinants change abruptly over time. This phenomenon of unstable structural parameters in asset price models is a special case of a general fundamental critique of econometric and statistical models [6].

In order to cope with the aforementioned shortcomings, a more comprehensive solution is needed. In practice, financial experts apply their intelligence (human-thinking) to solve the problem based on their knowledge of theories and existing strategies. In this case, investment decision results constitute what is called a managed portfolio.

However, integrating domain expert knowledge into the process of making investment decisions is very costly and experts need to concentrate on a limited number of available instruments as studying a huge number of them through this process is not feasible. Although professional analysts and fund managers make subjective judgments based on objective technical indicators, it is too complicated for non-professionals to do so.

To cope with this impediment, machine learning and data mining techniques have been utilized, and their notable power has thoroughly been proven. In many applications, Artificial Intelligence (AI) can partly replace human intelligence to reduce the cost of and bias in decision making. A well designed AI based model integrates one or more perspectives acquired from experts into a systematic process capable of reporting predictions with high accuracy. In order to deal with problems needing heuristic solutions, AI based approaches which include machine learning and data mining techniques are capable of addressing the problems.

Since the 1990s computer scientists have started to apply AI techniques in the financial domain by proposing heuristic solutions for investment decision making. These solutions were
later referred to as Business Intelligent techniques. Various techniques have been proposed for this purpose such as Neural Network (NN), Genetic Algorithm (GA) and Support Vector Machine (SVM). Recent research presents encouraging results on stock selection using data mining techniques [7-11]. There has been some efforts where researchers have tried to combine different techniques to propose hybrid methodologies in order to improve the effectiveness of the solutions [12-17].

Heuristic solutions based on econometric and statistical models as well as machine learning and data mining techniques have a main shortcoming. They lay in the exclusion of investor preferences or the propensity for risk that shapes a successful scenario. Unfortunately, these models have had mixed results in practice. Portfolio performance in the short term holds up well, but long term forecasts are outperformed by a “random walk down wall street” [18]. Debating the merits of the Efficient Market Hypothesis, a common criticism of portfolio optimization tools is that they fail to fully incorporate relevant information, market complexities, and investor preferences that underly stock market variability. In an attempt to predict future returns, optimization models typically specify a functional form that reflects historical financial data, from which expected future returns are extrapolated, a form that may not hold over time. Current events suggest this approach is not conducive to capturing the uncertainty and behavior that leads to high market variability. Variability that gave way to the market crash of 2008 - i.e., Lehman Brothers defaulting on its creditors, the U.S. government refusing to bail them out, and the illiquidity of mortgage-backed securities [19].

Among existing investment systems two basic categories of systems can be identified: systems that are proposed by the research sectors which are based on scientific methods and are clearly worded but often not applied in the real-world, and systems developed by the industrial
sector, which are completely operational while their internal methodologies are not publicly revealed.

The lack of a unified system that can combine the advantages of these two types of systems is obvious. In this thesis a semi-automated system called *Stock Market Portfolio Optimizer* (SMPOpt), is introduced. The ultimate goal of this system is to guide investors’ decisions. Following several efforts by researchers in which they have tried to combine different machine learning and data mining techniques to propose hybrid methodologies in order to improve the effectiveness of the solutions, a new hybrid solution is proposed in SMPOpt. In this sense, combining *Social Network Analysis* (SNA), data mining techniques such as Clustering and machine learning techniques such as Support Vector Machine (SVM) could lead to an effective integrated solution that combines the advantages of strong mathematical foundations and prediction. SMPOpt is a comprehensive framework by integrating these techniques in the form of a decision support tool. The proposed heuristic solution combines financial analysis with data mining and artificial intelligence in a decision support environment. While the DSS is fully automated, it permits user interaction in a semi-automated mode for customized optimization decisions.

1.3 Contribution

Two disjoint strategies have been proposed and implemented in SMPOpt: 1) collecting information about the history of stocks in the market, applying diversification strategy based on investors preference (propensity for risk) and learning relationships between effective factors with future of stocks, with the goal of constructing and suggesting a particular portfolio to the investor, 2) collecting information about history of expert investors as well as non-professional
investors, with the goal of suggesting portfolio of an expert to the non-professional investor. These strategies are briefly explained next.

**The first** approach is a semi-automated technique for multi-objective portfolio optimization. Investors with different levels of expertise can benefit from the partial or final result of the process. Different categories of effective factors are generalized as: 1) investors preferences, 2) financial indicators about companies (called internal factors), and 3) other factors from the world outside of the companies (called external factors). Quantitative models for stock selection and portfolio management face the challenge of determining the most efficacious factors. In SMPOpt, investors are asked to specify their preferences in terms of which financial indicators are to be used in the training set of the learning system. In addition, investors communicate with SMPOpt to specify their propensity for investment risk in order to derive the optimum solution based on multi criteria. Therefore the multi-objective optimization problem is defined based on the specified factors. Using historical data to identify relationships among entities, an assessment of portfolio risk and return associated with a specific set of market conditions can be exploited. This approach permits more flexibility in multi-objective portfolio optimization, while accounting for historical stock trends and correlations crucial to deriving a successful diversification strategy. The proposed approach is accomplished in a four step process: (1) social network construction, (2) stock clustering, (3) stock ranking, and (4) portfolio construction.

Social Network Analysis (SNA) is a technique first used in sociology and anthropology. Recently computer scientists have realized that this model is general enough to be applied to any domain where the entities and their inter-connections can be separated into actors and their links,
respectively [20]. In this case, SNA can be applied as a powerful data mining technique that can extract valuable knowledge from the raw data [21].

Utilizing social network theory and analysis, a social network of companies in the stock exchange market is first constructed where the companies are the actors of the network and the links between them represent their performance historical correlations. The rationale behind this network is that there is an implicit relationship between the human management team of different companies. The historical correlation between the performances of a pair of companies can represent their similarity. Another reason of using performance historical correlation to measure the link between companies is to engage in a diversification strategy which is explained later.

Recently there have been a few studies which suggest the network representation of the investment instruments [22-25]. Their main goal is to study the relationship between companies and try to predict their future relationship. However, in this thesis the constructed social network is used for further steps in the proposed optimization solution. On the other hand, time series behavior of financial data is not evitable, and this might influence the effectiveness of the basic general algorithm in the optimization process. Therefore, more customization is required to build a social network of companies with time series correlation which is explain later in this thesis.

After network construction, an unsupervised data mining technique, namely clustering, is applied to the network in order to cluster the companies into homogenous groups. An unsupervised learning method seeks to determine how the data can be organized without knowing any pre-defined classification. Clustering is a method to assign a set of objects into groups (called clusters), so that objects in the same cluster are similar and objects in different clusters are dissimilar in some sense. The idea behind applying this method in investment management problem is to employ a diversification strategy. In this strategy, the goal is to invest
in various types of instruments in order to reduce the investment risk [5]. Using k-means clustering, highly correlated stocks are grouped together, a necessary insight for portfolio diversification.

The rationale behind discovering clusters of companies is to cover all of them in the final portfolio in order to have diversified investments. More diversification is achieved by greater number of clusters which reduces the investment risk. Therefore the investor’s propensity for risk is reflected in the number of clusters which will be explained later in this thesis.

After stock clustering, support vector machine learning, is applied to each stock cluster to rank order stocks within each group according to expected returns. With this information teased out of the market, a diversification strategy can be developed that optimizes the multi-objective portfolio. In other words, in this step the portfolio optimization problem is formulated as an asset ranking problem based on multiple objectives.

In stock investment, the result of the ranking process plays a fundamental role. The asset ranking process is usually performed by applying some objective criteria to measure the company’s performance. In traditional portfolio management, the usual approach to select winners and losers has been applied to evaluate individual asset’s past returns over the ranking period. The realized return as a selection criterion is a simple measure, which does not include the risk component of the stock behavior in the ranking period. Although, in many cases, approaches based on a single selection criterion provide good results, in general multiple criteria have to be considered for stock ranking [3].

In this thesis the capability of Support Vector Machines for Ranking (SVMRank) to rank stocks is evaluated. SVMRank was first proposed in [26] for ranking search results. Here it is used for stock ranking based on indicators that capture the past and future performance of the
respective companies. Although SVM has been shown to be useful for the problem of stock selection in some earlier works such as those described in [9, 17], to the best of my knowledge, there is no work that has applied SVM to the problem of stock ranking.

In addition, the fact that in non-stationary financial time series the dependency between input variables and output variables gradually changes over time is not evitable. To customize this technique a modified version of SVMRank, called Time Series SVMRank (TS_SVMRank) is proposed in this thesis, to model non-stationary financial time series. The TS_SVMRank is obtained by a modification in the error function in support vector machines. This procedure is based on the prior knowledge that the recent past data could provide more important information than the distant past data.

The last step in the first strategy in SMPOpt is portfolio construction based on the result of previous steps. For this purpose, top stocks from the ranked list of each cluster are selected to be added to the portfolio. After stock selection, their weights (number of shares to buy from each) should be identified. This phase is called stock weighting which is not the main focus of the proposed heuristic solution. Simple weighting function such as equal weights and market capitalization weight is used which are explained later.

In the second strategy in SMPOpt, social network of financial experts is built based on their publicly available portfolios. This social network is then used for further analysis to recommend an appropriate managed portfolio to non-professional investors based on their behavioral similarities to the expert investors. Each expert is a node in this network and the weighted link between two nodes shows how similar two experts are based on their portfolios. In the next step, various communities of experts are detected from this social network. On the other hand, a non-professional investor is assigned to an appropriate community of experts based on
the similarity that he/she has to the experts in each of the existing communities. One expert is then selected as the representative of each community whose portfolio is suggested to the non-professional investor.

As discussed above, the contribution of this thesis to the problem of portfolio optimization is the application and customization of appropriate data mining and SNA techniques to help investors in the decision making process. The proposed heuristic approach is implemented in SMPOpt and evaluated through a series of experimental studies. In each experimental setup a specific phase is focused leading to evaluate the whole proposed process.

1.4 Thesis Organization

This thesis proceeds as follows. Background on relevant theories and concepts, and related studies both by finance researchers and computer researchers are reported in Chapter 2. The problem of multi-objective portfolio optimization is defined formally in Chapter 3. Chapters 4 and 5, respectively, present the details of two proposed strategies in SMPOpt. Next, all experimental studies are discussed in Chapter 6 followed by SMPOpt implementation details in Chapter 7. Finally, Chapter 8 presents the summary and conclusions drawn from this thesis and directions for future work. Figure 1 shows the block diagram of chapters in this thesis.

![Figure 1 - Block Diagram of Thesis Organization](image-url)
Chapter Two: **Background and Related Works**

Stock portfolio optimization has been focused by many financial researchers as well as computer scientists. Several theories have been proposed and several software tools have been created to support investment decision making process. This chapter reviews financial theories and computer science based heuristic approaches followed by brief introduction to some of the existing software tools for this purpose. Next, the background on approaches used in the proposed solution in this thesis are discussed.

### 2.1 Related Financial Theories

Prior to 1950, investment decisions focused on individual stock returns with little consideration of portfolio risk. In 1952, Markowitz was the first to quantify the link that exists between portfolio risk and return through which he founded the *Modern Portfolio Theory* [5]. He demonstrated that portfolio risk came from the covariance of assets making up the portfolio. He empirically demonstrated a link between portfolio risk and cumulative asset returns, measuring portfolio risk relative to the covariance among assets. His theory targeted to maximize the use of the investor’s terminal wealth [4, 27]. In other words, this theory attempts to mathematically identify the portfolio with the highest return at each level of risk.

The optimization problem proposed by Markowitz consists of finding the stock weights in the portfolio which minimizes the variance of the portfolio for a given value of the average return of stocks in the portfolio. The solution of this optimization problem has been found by Markowitz at 1959. This solution relies upon a series of assumptions that are rarely observed in practice [5]. For instance, the asset returns are assumed to be Gaussian variables where fat tails in price return distribution are observed, and the parameters used in the optimization are assumed constant [28].
Portfolio risk is defined as the likelihood that a group of assets (e.g., stock portfolio) will not earn an expected rate of return. It is comprised of two subcategories of risk – systematic risk and unsystematic risk. Systematic risk (a.k.a., market risk or undiversifiable risk) is the level of exposure to adverse environmental, social or geopolitical factors that affect the entire market. Consequently, systematic risk cannot be reduced through asset diversification. In contrast, unsystematic risk (a.k.a., business specific or idiosyncratic risk) refers to company-specific or industry-specific exposure to adversity that is uncorrelated with aggregate market returns. Thus, unsystematic risk can be reduced through asset diversification [1]. In other words, return on investment instruments is affected in response to various events and conditions that affect the environment. Some events influence all companies, such as war, global disasters and economic events, while others affect only specific companies, such as weather changes and demand conditions [1]. The changes in return that occur in response to events affecting all the market are known as Market Risk and the remaining changes are called Business-Specific Risk [1].

During 1950s and 1960s, Sharpe and Lintner introduced the notion of risk into the valuation of assets. Their model is called Capital Asset Pricing Model (CAPM) [5]. It evaluates the asset return in relation to the market return and the sensitivity of the asset to the market [5]. In other words, by measuring the market risk of each asset a risk-adjusted expected return is measured by CAPM. Building on the insights of modern portfolio theory, the CAPM provided investors with a formulation for determining an asset’s required rate of return taking into account the asset’s sensitivity to systematic (undiversifiable) risk [5]. Armed with measures for both constituents of portfolio risk, theoretically portfolio optimization should be readily achieved. However, portfolio optimization is often elusive or short-term at best [19].
CAPM is based on a strong set of theoretical assumptions which are not entirely reflected on markets in practice [5]. A strong criticism of CAPM is that it fails to account for information asymmetries that exist in the market. A related theory, the Efficient Market Hypothesis (EMH), addresses this criticism. According to the EMH, there are three levels of information efficiency that can characterize a market - strong, semi-strong, and weak. At the strong level no information asymmetries exist – even hidden and insider information are reflected in the price of an asset. At the semi-strong level all publicly available information is reflected in the price. At the weak level only historical information is reflected in an asset’s price [29, 30].

Given the assumption of strong market efficiencies, portfolio optimization models based on CAPM tend to invest more heavily in high risk assets than is optimal [30]. Consequently, multi-factor models have been developed that relax the contentious assumptions underlying CAPM. One such model is Arbitrage Pricing Theory (APT). Where CAPM focuses on market risk in asset valuation, APT takes additional factors into consideration under the assumption of semi-strong market efficiency. The number and nature of factors considered in APT are left up to the individual modeller. Due to the wide range of factors, a portfolio investor can incorporate in APT models, APT model comparisons are limited [5].

Unrealistic assumptions and time complexity of the required calculation in financial theories are also issues that reduce their applicability to real world optimization problems [5]. In [31] it is mentioned that despite the sophisticated theoretical models developed in the last 50 years and the advances in methods for estimating the parameters on these models, investors continue to use simple allocation rules such as 1/N strategy for allocating their wealth across assets. The 1/N strategy is a naïve rule in which a fraction 1/N of wealth is allocated to each of
the N assets available for investment. In [31] it was shown that out of the 14 models evaluated, none is consistently better than the naïve 1/N benchmark.

In practice, a more comprehensive optimization solution is often needed. To satisfy this need, professional analysts and fund managers make subjective portfolio decisions, guided by technical indicators (e.g., profit margin or revenue per share of the companies). Computer scientists apply Artificial Intelligence (AI) and data mining techniques to bring greater objectivity to portfolio decision making. AI is coupled with soft computing techniques – such as, Support Vector Machines – that are widely accepted by portfolio investors and diversification strategist evaluating market behavior [32].

2.2 Related Computer Science based Heuristic Approaches

The portfolio optimization problem has been studied in computer science since 1990s. Past AI research on portfolio optimization formulated the investment management problem as a classification ([33]) or regression ([34]) problem.

In classification, various techniques applied to label different classes defined in the formulated classification problem. For instance, a simple binary classification is defined in [7] based on the sign of the return which can be positive or negative. Another example of class labeling is described in [35]; it ranks the stocks based on expected return and assigns top stocks in the ranked list to the successful class and the rest to non-successful class.

In regression, the main goal is to predict the value of a variable such as stock's return or market's return by studying its historical trend. For this purpose, various AI techniques have been used by researchers. For instance, SVM and Neural Network (NN) are applied in [17] to solve regression estimation.
Researchers in [33] applied machine learning techniques such as *Minimax Probability Machine* (MPM) and SVM as a solution of the formulated classification problem. Their results in terms of accuracy of classifier showed that the MPM and SVM methods are good candidates for identifying which stock should be traded. The comparison of the two techniques in their study provided similar results [33].

Non-linearity in the financial markets is recognized by several works such as [7, 14, 17]. SVM and Neural Networks (NN) are two techniques successfully applied in solving nonlinear regression estimation problems in this domain [17]. Recent studies found that SVM outperforms neural networks, decision trees, and ordered choice models in predicting outcomes [36, 37]. It was empirically found that SVM is very effective in forecasting financial outcomes using times series data [10]. SVM is a promising method for this purpose because it uses a risk function consisting of the empirical error and a regularized term which is derived from the structural risk minimization principle [10].

Related research employing hybrid approaches that couple machine learning with data mining techniques have achieved some success. For instance, researchers in [34] used SVM in combination with other classification methods to forecast the direction of short-term movements in the market. Another hybrid model based on *Autoregressive Integrated Moving Average* (ARIMA) and SVM was employed in [17]. Widely used in time series analysis to forecast outcomes, ARIMA works best with data trends that are linear or quadratic. It is not suitable for nonlinear patterns, where SVM is most helpful [17]. Additional hybrid approaches are based on *Artificial Neural Networks* (ANNs) and Genetic Algorithm (GA), as proposed in [14].

Other studies of portfolio management techniques have offered partial solutions to the portfolio optimization problem. One promising approach used a fuzzy SVM to predict stock
market outcomes [38]. In [39] a fuzzy rule based expert system was developed for stock price analysis by modeling rule uncertainties. Authors of [40] proposed a basic portfolio selection model in which rate of asset return and future risk were represented by triangular fuzzy membership. After clustering the data, a fuzzy optimization model was used to determine the optimal proportion of each cluster to investment in.

2.3 Similar tools to SMPOpt

There are several tools available to the public, at different costs, in order to help the investor with making investment decisions and achieving a satisfactory return compared to the risk taken. A variety of services are offered by them such as only getting access to required financial data, or far more applying different types of analyses to recommend some investment guidance trying to bridge the transition from analysis to portfolio construction.

Investing online has become the norm for individual investors over the past decade. They provide a trading platform which acts as the hub, allowing investors to purchase and sell stocks. Due to their lower fees, as opposed to full service brokers who also give advice to clients, many people are interested in these online services. Thus many, if not all, brokers are now offering online services with unique trading platforms. As an example, Questrade ([41]) provides online brokerage services by offering different investment user accounts such as individual and joint accounts, informal trust accounts, tax free savings accounts, and registered retirement savings plans. Other examples of online brokerage services are Stocktale ([42]), Plus500 ([43]), MadScan ([44]), and Spectrum Live ([45]).

Included with the trading platform, provided by an online brokerage service, are tools to track and monitor stocks, as well as research tools, real-time streaming quotes and up-to-date news releases. For instance, stock screener such as Yahoo! Finance and Google Finance ([46]),
is a tool that evaluates stocks based on criteria and generates a list of potential trading ideas for the investors.

In addition to online services, there are several software packages available to be purchased with the goal of helping the investors in their decision making process. Value Stock Selector ([47]), Manage Invest ([48]), Checklist Investor ([49]), and Stock Assault ([50]) are some examples of investment decision softwares. As they are all commercial products their methodologies are not open to public and thus a brief description about their analyses can be found. For instance, Stock Assault developer claims that some AI techniques are used in the tool without explaining about these process clearly.

2.4 Approaches used in SMPOpt

Recall there are two main objectives in portfolio optimization; maximize the portfolio’s return while minimizing portfolio risk. Riskier stocks offer higher returns to compensate for the risk differential relative to other stocks. To create a portfolio that minimizes risk, investors diversify their stock holdings, acquiring stocks that are weakly or negatively correlated to offset any individual stock’s unsystematic risk [1, 5]. These are the basic tenents of modern portfolio theory. In this thesis, a comprehensive solution is proposed to the portfolio optimization problem that employs data mining in the form of Social Network Analysis (SNA) and Clustering, with AI in the form of SVM ranking to structure the stocks for portfolio selection.

SNA was introduced in 1934 as a subarea of sociology and anthropology to study the connectedness of people in groups [20]. Based in graph theory, a social network consists of nodes on a graph and their relations. After realizing that this model is general enough to be applied to any domain where the entities and their inter-connections can be separated into actors and their links, there are some efforts to use the network approach in the financial domain, e.g.,
Researchers in [22] built a network representing the stock market prices over time. Vertices of the graph represent stocks and the weight of an edge connecting two vertices is determined by calculating correlation coefficient of their returns over the time considered period. Then some dependencies among stocks were analyzed using this representation. The analysis led to the conclusion that market data has a power-law structure [22]. Therefore, they concluded that the concept of self-organized networks is applicable in finance. Researchers in [23] followed an approach similar to the one presented in [22] and built network of the Chinese stock market. By conducting the same analysis, they also found that power-law model is valid in financial networks like many other real-life networks. A refinement to the approach researchers in [25] suggested using partial correlations, in effect removing the influence of market risk. They claim that in stock markets, the stock price of a given company is not only affected by its own fundamentals and other associated listed companies, it is also influenced by the fluctuation of the stock market index. They constructed minimum spanning tree based on simple correlation coefficient, and partial correlation coefficient. Finally, they analyzed distribution features of stocks in different manners to conclude that partial correlation coefficient by considering market index was a better choice. The studies described in [22, 23, 25] focused on network’s structural properties, topological stability and the evolution of the market. In this thesis the goal is to use the constructed network for further analysis to suggest appropriate investment guidance.

After stock network construction, clustering analysis is applied with the purpose of obtaining meaningful partitions of stocks according to their historical similarities. Based on diversification strategy, in portfolio formation it is often advantageous to diversify the portfolio by selecting stocks from several groups. Each stock sold on a Stock Exchange is classified by industry type. These categories of stocks can be used for diversification among various
industries. In addition, some techniques use factors such as industry, size and growth to form stock clusters. Researchers in [51] applied data mining techniques to determine a stock’s industrial category given a historical record of that stock’s prices. In this thesis, the correlation between each pair of stocks measured in network construction step are used to discover the clusters of stocks which are then used to apply diversification strategy.

The majority of AI work pertaining to financial markets focuses on predicting stock price movement. However, a few studies have applied AI techniques to stock valuation and ranking [6, 11]. The authors of [6] examined the use of neural networks as a substitute for statistical forecasting techniques to rank order stocks. Their experimental results show that neural networks outperform statistical methods.

In [11] a genetic programming technique using multi-objective fitness functions was proposed for selecting factors and constructing multi-factor models for ranking stocks. To evaluate their technique, the authors built a portfolio by buying the top portion and selling short selling stocks from the bottom portion. Their benchmark was the equal weighted S&P 500 stock returns. The study’s results indicate their technique provided an efficient and effective means of rank ordering stocks.

In [3] and [2] the interaction between the decision maker and the system was taken into consideration to evaluate stocks based on investor’s preferences, in order to rank them and single out eligible ones to be included in the portfolio. To the best of my knowledge, no research exists that evaluates the appropriateness of SVMRank for rank-ordering stocks based on a set of financial criteria. The use of support vector machines for stock ranking is particularly promising given its ability to sort and categorize data drawn from non-linear distributions.
Chapter Three: **Stock Portfolio Optimization (Problem Definition)**

Identifying effective factors in a problem is a significant step of problem definition. On the other hand, as some general purpose solutions might be able to solve the current problem, formulating the problem in a way to be solved by such a solution is helpful. In some cases, there might be a need to modify the general solution to fit with the current problem. This step is called solution customization in which specific characteristics of the problem should be considered in order to modify the solution to match the problem.

The above steps form the journey that has been passed in this thesis to define the stock portfolio optimization problem and proposed a heuristic solution for that. Identifying the effective factors, and defining the problem are discussed in this chapter and the rest of steps of formulating the problem, finding general purpose solution as a solution, and customizing the general purpose solution are discussed in Chapters 4 and 5.

### 3.1 Identifying the Effective Factors on Stock Portfolio Optimization

In stock markets, companies from all economic sectors can register to sell a percentage of their organization to the public in order to raise capital. To facilitate this type of transactions, a company must be scheduled to be listed on a stock exchange that brings buyers and sellers together. Exchanges provide clearing and settling services guaranteeing shares to buyers and payment to sellers. The exchange thus eliminates the risk of default by either buyer or seller.

The largest stock exchanges in the USA and Canada are the New York Stock Exchange and the Toronto Stock Exchange. Some examples of key European exchanges are the London Stock Exchange and Paris Bourse. Buyers and sellers in the stock market range from small retail investors to large hedge fund traders, who work with substantial pools of capital.
Stock orders enter the exchange and bids to buy and sell are matched. If a spread exists, a difference between bid and asking price, no trade will take place. After a trade has been made details are recorded and a notification is sent back to the investor.

Although there is a significant amount of human contact in this process, computers play an important role, especially for so-called "program trading". Online financial websites like Yahoo! Finance [52] offer investors a comprehensive view of available stocks to buy. This information contains details required to construct an order as well as financial statistics that can be used to decide on which stocks to purchase or sell. This free service is provided in near real time.

Placing an order is easy, but making a profitable investment decision requires due diligence and calculated timing. Financial theorists and investors are continually researching new ways to improve their investment strategies. Many analytical methods have been proposed to help investors build their stock portfolio, but the selection can be overwhelming with over 63,000 publicly traded companies worldwide to choose from.

Knowing what to buy and when to buy or sell a stock requires understanding the factors that control market volatility. Essentially this is predicting reasons for the rise and fall of a particular stock’s share price. There are many factors that can affect the price of each stock. Factors and their importance differ depending on a stock’s characteristics. For instance, fluctuations in crude oil prices can impact energy companies share prices positively and transportation companies share prices negatively. Being aware of how stocks fluctuation is correlated can also help in timing profitable buying and selling activity. Wise investment decisions cannot be made without understanding the relationship between a stock’s share price and key internal and external factors [4, 5].
So making an investment decision is a challenging task in which different factors need to be considered. The factors can be grouped into internal and external factors. Internal factors are those that illustrate the managerial and overall status of a company. There are many financial ratios designed to illustrate some aspect of how a company is performing [1]. These ratios are measured that can be used as internal factors.

External factors are influences that exist outside the company’s immediate control. Examples of these factors include changes in government policies, energy prices, and weather patterns.

In addition to effective factors on the performance of companies, both from inside and outside of the companies, the investor’s preferences form an important and critical parameter that needs to be considered in the investment decision making process. For instance, propensities for risk of various people are different, which could lead to different investment decisions. To summarize the effective factors, three main categories are defined as investor’s preference, internal and external factors. Figure 2 shows these three categories as the inputs of a stock portfolio optimizer.

Figure 2 – Effective Factors on Stock Portfolio Optimization
3.2 Definition of Stock Portfolio Optimization

Stock portfolio optimization is the process of making investment decision for holding a set of stocks to meet various criteria.

**Definition:** Given \( n \) stocks, their historical price and \( m_1 \) predictor features from fundamental analysis of the companies and \( m_1 \) predictor features from outside of the world, portfolio optimization is finding \( x_i \) for stock \( i \) as the proportion of that stock in the final portfolio with the goal of finding an appropriate cut for the trade-off between various criteria with respect to the investor’s preferences.

Based on the above definition, the current problem is a multi-objective optimization problem. Several artificial intelligence techniques such as Neural Network, Genetic Algorithm, Support Vector Machine can be used as a solution of such a problem. The proposed solution in this thesis is a hybrid technique containing data mining and machine learning algorithms. For this purpose, this problem is first broken down into steps of stock selection and weighting. The proposed solutions are implemented in a system called Stock Market Portfolio Optimizer (SMPOpt).

Two main strategies are followed in SMPOpt; **first** a semi-automated process is defined through which the investor is guided to build the portfolio based on his/her preferences and the state of financial factors. **Second**, existing available managed portfolios (built by financial experts) are analysed to find the most appropriate managed portfolio for a non-professional investor based on his/her similarities to experts. These two strategies are, respectively, explained in Chapters 4 and 5.
Chapter Four: **First Solution in SMPOpt: A Semi-Automated Portfolio Optimizer**

4.1 **Overview of the Proposed Semi-Automated Process**

The main goal in the proposed methodology is to apply appropriate level of diversification in order to find the optimum cut for the trade-off between risk and return based on the investor’s objective. For this purpose, as explained in Chapter 3, three categories of effective factors are considered; investor’s preference, internal factors from inside of companies, and external factors from the world. These are three sources of input to the proposed process. Different levels of automation can be achieved from this process. Investors can be involved in the process of portfolio construction in a semi-automated or fully-automated process. In the semi-automated process, investors can benefit from the partial result from each step and then continue the rest manually, while in the fully-automated process all the steps are performed by the system and the final portfolio is suggested to the investor without any user collaboration. Therefore, different types of output are provided by SMPOpt; partial and complete results.

In this chapter, it is explained how existing data mining and social network analysis techniques are utilized to proposed a heuristic based solution for portfolio optimization. As explained in Chapter 3, this problem can be divided into two phases; stock selection and stock weighting. In stock selection phase, a subset of stocks is picked to be added to the portfolio. As considering all possible subsets of stocks and evaluate the risk and return of the portfolio containing each of these subsets are very time consuming, a heuristic method is proposed for subset selection. For all the stocks that are not selected in the first phase their weights in the portfolio would be zero and the rest construct the portfolio based on specified stock weights. A pre-processing step might be required to make financial dataset ready for further analysis. Figure 3 illustrates the component diagram of this process including inputs and outputs.
The proposed portfolio optimization solution is enacted in two phases of (1) stock selection and (2) stock weighting. Stock selection includes a three-steps heuristic: (1-1) network construction, (1-2) stock clustering, (1-3) stock ranking.

The network construction step involves building a social network where actors represent companies (stocks) in the market, and the weighted link between each pair of stocks is the correlation of their historical performance. The period of time to be considered as the history can be either specified by a default value or provided by the user. Stock performance, also known as stock value, can be simply measured by the return value [5]. Subsequently, the correlation between historical performances of two stocks can be measured by Correlation Coefficient metric of their historical returns [5]. Building such a network overlaps with what have been proposed in [22, 23, 25]. Another metric for measuring the stock performance available in SMPOpt is based on CAPM theory [5]. In this theory, expected return of each asset can be
measured based on the market risk of that asset. This value is called *risk-adjusted return*. In SMPOpt, *return* and *risk-adjusted return* are two examples of general performance metric proposed to be used for social network construction. Investors can provide any performance metric to SMPOpt to be used by the system in this step. Furthermore, to customized correlation measurement for time series dataset, a new metric is proposed in this thesis. Considering historical data to find the current correlation between the stocks assumes that the future will reproduce the past without any modification [5]. However, this is not true. To improve the results and to take the evolution into account, the recent period in the history should be considered more important. This customization is based on the prior knowledge that the recent past data could provide more important information than the distant past data. Therefore, in calculating the correlation, recent data points should have higher weights compare to the distant past points. The output of this phase can be either shown to the user as a partial result or used for further analysis steps. Section 4.2.1 covers details about the network construction phase.

For **stock clustering**, a clustering technique is applied to the network to combine company stocks into homogeneous groups to be used for reducing risk through diversification. Any existing clustering algorithm can be used for this purpose. K-means and Louvain methods are two available clustering algorithms in SMPOpt. Number of clusters can be explicitly provided by the user. However, SMPOpt has the power to apply multi-objective genetic algorithm optimization process to find the optimum number of clusters based on user’s propensity for risk in terms of a number between 0 and 1 provided to the system. Section 4.2.2 covers details about the stock clustering phase.

In **stock ranking**, stocks within each of the discovered clusters are ordered. For this purpose, a learning process discovers the relationship between historical values of effective
factors (from both internal and external categories) and the target rank score based on stocks future performance value. This learning process is achieved by applying a ranking algorithm based on SVMRank. To customize the learning process for time series dataset, a modified version of SVMRank is proposed in this thesis which is called Time Series SVMRank. Section 4.2.3 covers details about the stock ranking phase.

Once rank orderings are revealed, SMPOpt moves to step four, *portfolio construction*, where the highest performing stocks are selected and weighted. The selected assets form diversified portfolio of weakly or negatively correlated high performing stocks that satisfies user’s objectives. Section 4.2.4 covers details about the portfolio construction phase.

As shown in Figure 3, investor’s preference is provided in terms of propensity for risk and some other required parameters. As briefly explained above, each phase needs some parameters, such as period of time and performance metric in network construction phase, which can be set by user. However there are some default settings for each of them which allow the user to get benefit of the fully automated process. Details on all components of the proposed framework are explained in the sequel in the rest of this chapter.

4.2 Four Phase Portfolio Optimization Approach

4.2.1 Stock Social Network Construction

Based on the fact that the performance of stocks simulates human behavior of the management team, it is proposed to build a social network of stocks dynamically over time to be able to discover dynamic clusters of stocks. Actors of the proposed network are $n$ available stocks in the portfolio optimization problem. The link between actors is defined in a way to make the network as the basic data structure of the proposed heuristic solution to this problem.
Weighted link between each pair of stocks is the correlation between their historical performances.

Figure 4 reports inputs and outputs of this component. List of the stocks to be used is required to be specified by user. For each stock in this list a node is created in the network. In addition, there are three other optional inputs from the investor; period of history, stock performance metric, and correlation formula. Investor can set these parameters or leave it to the system to use the default values. The historical performance of all selected stocks need to be retrieved from the source of internal factors. The output of this component is the constructed social network, which is shown to the user as the adjacency matrix or the visual network.

Figure 4 - Inputs and Outputs of Network Construction Phase

Figure 5 illustrates the Graphical User Interface (GUI) in SMPOpt through which the user can select a list of stocks for further analysis starting with social network construction.

Figure 5 - List of Stocks selection GUI in SMPOpt
In SMPOpt, historical performance values are imported from financial data sources such as Yahoo! Finance and Bloomberg [53]. Figure 6 shows import setting GUI in which the user can specify the period of time and list of stocks to be retrieved from Yahoo Finance!. These data are first downloaded and stored in csv file and then transferred to the local SMPOpt database.

![Figure 6 - Data Import Setting GUI in SMPOpt](image)

Other parameters for network construction in Figure 4 can be set in Network Setting GUI of SMPOpt (Figure 7).

![Figure 7 - Network Setting GUI in SMPOpt](image)

Performance measurement of the stocks is the first stage in portfolio performance analysis. Return on an asset is the basic element employed to determine stock performance, also known as stock value [5]. For measuring the return, a period is assumed as an interval of time during which an asset is held without being modified, and companies make payments to its
shareholders in terms of dividends at the end of the period. When a company earns a profit, it can be put for two uses: (1) re-invested in the business, and (2) paid to the shareholders as dividend. Many companies retain a portion of their earnings and pay the remainder as a dividend [5]. Available raw data contains stocks’ prices over the time and dividend paid by each of them at the end of each period. Return of stock \( i \) at time \( T \) is calculated by Equation 1 [5]. Mean of historical returns of stock \( i \) in a period of time is considered as the Expected Return of stock \( i \) at the end of that period \((ER_{i,T})[5]\).

Equation 1 \[ R_{i,T} = \ln \left( \frac{P_{i,T} + D_{i,T}}{P_{i,T-\Delta t}} \right) \]

where, \( P_{i,T-\Delta t} \) is the price of the stock at time \( T-\Delta t \); \( P_{i,T} \) is the price of the stock at time \( T \); and \( D_{i,T} \) is the dividend paid at time \( T \).

As discussed in modern portfolio theories, the concept of return is not sufficient on its own to analyze the results of a portfolio [5]. To analyze portfolio performance more precisely, a quantitative measurement of risk is required. Investment risk can be defined as earning a return that is less than what has been expected [1]. To measure this concept, in portfolio theory, return is considered as a random variable. The reason is that it is influenced by a significant number of uncertainties in both the future price and the future dividend [1]. By considering the mean of historical returns of an asset as its expected return, the probability of less than the mean can be measured as the risk of that asset. Theorists have noticed that return distributions are usually relatively symmetrical. In other words, return is distributed normally, and thus a large left side always implies a large right side as well. Based in this assumption, in portfolio theory, risk is variability of the return. In other words, risk of an asset is defined as the standard deviation of the
probability distribution of its return [1]. Equation 2 is the \textbf{Risk of stock} \( i \) during the period ending at \( T \) [5].

\textbf{Equation 2} \[ \sigma_{i,T} = \sqrt{\frac{\sum_{t=1}^{T}(R_{i,t}-\bar{R}_i)^2}{T}} \]

where, \( \bar{R}_i \) is the average return of stock \( i \) during the period ending at \( T \).

\textit{Return of the portfolio} of \( n \) stocks at time \( T \) based on the assumption that the portfolio has a fixed composition throughout the evaluation period is measured by Equation 3 [5].

\textbf{Equation 3} \[ R_{p,T} = \sum_{i=1}^{n} x_{i,T} \cdot R_{i,T} \]

where, \( R_{i,T} \) is the return of stock \( i \) at time \( T \); \( x_{i,T} \) is the weight of stock \( i \) in the portfolio at time \( T \).

The relative risks of stocks are entirely changed in portfolio compared to individual stocks [1]. The complete characteristic of portfolio risk requires that the behaviour of the stocks return be compared with that of other stocks. For this purpose, covariance of the returns of each pair of the stocks should be considered. The classic mean-variance approach that underlies Modern Portfolio Theory assumes stock returns are normally distributed. Consequently, the variance-covariance matrix of selected stocks is used to quantify risk as variability of the portfolio’s return [1]. Equation 4 shows how Markowitz measures the \textbf{Risk of the portfolio} during the period ending at \( T \) [4].

\textbf{Equation 4} \[ \sigma_{p,T} = \sum_{i=1}^{n} \sum_{j=1}^{n} x_{i,T} \cdot x_{j,T} \cdot \rho_{ij,T} \]

where, \( \rho_{ij,T} = \frac{\sigma_{ij,T}}{\sigma_{i,T} \cdot \sigma_{j,T}} \) is correlation coefficient; \( \sigma_{ij,T} = \frac{\sqrt{n}}{T}(R_{i} - \bar{R}_i).{(R_{j} - \bar{R}_j)} \) is the covariance of the return on stock \( i \) and \( j \) during the period ending at \( T \); and \( \sigma_{i,T} \) is the standard deviation of stock \( i \) during the period ending at \( T \).
Based on Equation 4, the correlation between stocks has direct influence on the portfolio risk. In other words, the weaker correlations between stocks in the portfolio lead to the greater reduction in portfolio risk. This is the rationale behind diversification strategy which is adding diverse stocks (by having weaker correlations) to the portfolio [1].

In the proposed approach, to solve the multi-objective portfolio optimization problem (higher return - lower risk), the network is built in a way to be able to apply diversification strategy (to reduce the risk) and in the other hand to increase the return (detail explained later in stock ranking). For this purpose the weights of the links in the basic form of the proposed network are measured by correlation coefficient (Equation 5) [4].

\[
\rho_{ij,T} = \frac{\sum_{t=1}^{T}(R_{i,t} - \bar{R}_i)(R_{j,t} - \bar{R}_j)}{\sqrt{\sum_{t=1}^{T}(R_{i,t} - \bar{R}_i)^2} \cdot \sqrt{\sum_{t=1}^{T}(R_{j,t} - \bar{R}_j)^2}}
\]

The value of the correlation coefficient of two series is always between -1 and 1. The greater the absolute value is, the more correlated are the series. If the series have the same fluctuations over time, they will have a positive correlation. If they fluctuate contrary to each other, the value would be negative. But if there is no correlation between stocks’ fluctuations, the value would be close to zero.

Figure 8 illustrates an example of network construction for the Dow Jones Index in Dec. 2004 based on historical data for Jan. 2004 (created and visualized by SMPOpt).
Building the network based on Equation 5 is the basic form of the proposed network. However, more general social network construction is proposed in this thesis in which any kind of performance metric can be used by applying any correlation formula.

4.2.1.1 Other Stock Performance Measurements

In addition to the basic return as the performance metric of stock, other measurements such as different financial indicators can be used. In SMPOpt it is possible for the investor to introduce a new performance measurement, while is it not a required input and the system can apply the default metric. For this purpose, in addition to return, risk-adjusted return is also available in SMPOpt which is explained next in detail.

As explained before the concept of risk can be broken into two known categories of Market and Business-Specific risks. In a well-diversified portfolio containing companies in fundamentally different industries, there is no Business-Specific risk. Because at the aggregate level this type of risk is washed out statistically [1]. However, Market risk is a share of investment risk that is not eliminated by diversification [5]. The only way to reduce the market risk through diversification is adding stocks to the portfolio which moves countercyclically with
the market. The classic example of such a stock is successful gold mining companies. When returns of most stocks are down, people invest their money in tangible stocks such as gold, which cause increase in their price (increase in the return). The reverse happens when stock returns are generally high in the market [1]. Therefore, such a gold mining stock has the opposite trend of the whole market and thus having such a stock in the portfolio will decrease the market risk of the portfolio. However, available shares of such stocks are few and thus not feasible to always be added to the portfolio [1].

Based on the above discussion, the goal is to reduce the Business-Specific risk of the portfolio, through applying the diversification strategy. However, to define more precise stock value (instead of return), CAPM theory is used to measure Market Risk-Adjusted return. Background about CAPM is reviewed next followed by the formula of risk-adjusted return.

Return values of all the stocks have been affected by market risk. Markowitz and Sharpe proposed CAPM in which market risk of each stock is measured [1]. This metric called stock’s Beta Coefficient (β). β is developed by plotting the historical relationship between Stock Return and Market Return. Market return is the return of a portfolio containing all the stocks in the market. A regression line fitted to data points (Stock Return as horizontal axis and Market Return as vertical axis) is known as the characteristic line for the stock [1]. The slope of this line tells on average how much of a change in Stock Return has come about with a given change in Market Return, which is known as Beta Coefficient of the stock [1].

CAPM theory led to use Beta Coefficient of stock to measure the expected return of the stock by removing the influence of market risk on that stock (Equation 6) [5]. It is proposed that, this performance measure leads to calculate market risk adjusted correlation between stocks by first deleting the impact of the market on each stock.
Equation 6 \[ E(R_{LT}) = R_{RF,T} + (R_{MT} - R_{RF,T}).\beta_{i,T} \]

where, \( R_{RF,T} \) is the risk-free rate at time \( T \) (the current rate of interest paid on three-month Treasury bills is generally taken as the risk-free rate [1]); \( R_{MT} \) is the market return at time \( T \); and \( \beta_{i,T} \) is the Beta Coefficient of stock \( i \) at time \( T \) (Beta is usually calculated based on recent 60 months).

4.2.1.2 Time Series Correlation

Based on the prior knowledge that in the financial time series the recent past data could provide more important information than the distant past data, in calculating the correlation between stock performances based on Equation 5, it is preferred to make the recent period more important. This is achieved by time series analysis. The autoregressive moving average (ARMA) family model can be used for this purpose [5, 17]. The model that is used in this work is AR (autoregressive), in which the performance metric can be expressed as the weighted sum of its past values and a random term (Equation 7) [5].

Equation 7 \[ R(T) = e(T) + \sum_{t=1}^{T} a_t. R(T - t) \]

where, \( R(T) \) is the performance metric at time \( T \); \( e(T) \) is a random term; and \( a_t \) is the weight of importance for each time stamp.

Based on the above model, by considering \( e(T) \) as zero and an appropriate time weight function \( W(t) \) as \( a_t \), a modified version of Correlation Coefficient (Equation 5) called Weighted Correlation Coefficient (Equation 8) is adapted to measure time series correlation [54].

Equation 8 \[ \rho_{ij,T} = \frac{\sum_{t=1}^{T} W(t)(R_{it,T} - \bar{R}_i)(R_{jt,T} - \bar{R}_j)}{\sqrt{\sum_{t=1}^{T} W(t)(R_{it,T} - \bar{R}_i)^2} \cdot \sqrt{\sum_{t=1}^{T} W(t)(R_{jt,T} - \bar{R}_j)^2}} \]

where, \( \bar{R}_i \) is weighted mean based on Equation 9 [54].
The appropriate time weight function need to be ascending to satisfy \( W(i) > W(i-1) \), \( i=1,2,...,k \), so that the weights incline from the distant training data points to the recent points. Exponential weight functions (Equation 10) shown in Figure 9 satisfies the requirement [55]. In this equation \( a \) is the parameter to control the ascending rate and \( k \) is the total number of time instances.

**Equation 10**  
\[
W(i) = \frac{1}{1 + \exp(a - \frac{2ai}{k})}
\]

![Figure 9 - Weights function in TS_SVMRank](image_url)

**4.2.2 Stock Clustering**

After network construction, an unsupervised data mining technique, namely clustering, is applied to the network in order to cluster the companies into homogenous groups. Figure 10 reports inputs and outputs of this component. An unsupervised learning method seeks to determine how the data can be organized without knowing any pre-defined classification. The idea behind applying this method in investment management problem is to employ the diversification strategy. In this strategy, the goal is investing on various types of instruments in order to reduce the investment risk [5]. Figure 11 shows the GUI in SMPOpt through which the user can set stock clustering parameters.
Recall that portfolio risk is reduced through asset diversification; portfolio diversification requires weak correlation among stock returns. The network clustering process groups together highly correlated stocks, while minimizing the between-group correlations. Assignment of stocks to clusters is based on the correlation coefficients (weighted links) generated during social network analysis in stage one.

Detected clusters can show different categories of the companies in the real world based on their business and/or size which change dynamically over time based on the period considered in building the network. The reason behind this belief is what Markowitz mentioned in [4]: “Firms in different industries, especially industries with different economic characteristics, have lower covariance than firms within an industry”. He also mentioned that “It is necessary to avoid
investing in assets with high covariance among themselves. We should diversify across industries”.

The greater the number of distinct clusters found in a portfolio, the more diversified it is - lowering portfolio risk. In this regard a user’s propensity for risk is reflected in the number of clusters they choose to diversify on. In the case of user's lack of sufficient knowledge to specify a particular number of clusters, SMPOpt uses a multi-objective optimization to determine an optimum number of clusters in the dataset based on the specified propensity for risk in terms of a real number in [0,1]. This optimization process is discussed in Section 4.2.2.1.

4.2.2.1 Mapping Propensity for Risk to Number of Clusters

In this phase, SMPOpt applies optimization process which helps the system find a more appropriate number of clusters based on cluster validity analysis as well as investor's propensity for risk. As discussed before, the larger the number of distinct clusters found in a portfolio, the more diversified it is; this is a case which has less risk. Therefore, number of clusters should be set based on required level of diversification which is set based on the investor's propensity for risk.

In addition to the objective about the number of clusters (based on propensity for risk), the goal is maximizing cluster quality which combines two sub-objectives, namely maximizing within cluster similarity and maximizing between clusters dissimilarity. For this purpose, two metrics of cluster homogeneity and cluster separateness are measured [56]. These metrics evaluate the clustering solution by examining how well the clusters are separated and how compact the clusters are. Equation 11 and Equation 12, respectively, show the homogeneity of cluster $C$ and the separateness of clusters $C_i$ and $C_2$. 

40
Equation 11 \[ H(C) = \frac{\sum_{O_i, O_j \in C, i \neq j} \text{Similarity}(O_i, O_j)}{|C|(|C|-1)} \]

Equation 12 \[ S(C_1, C_2) = \frac{\sum_{O_i \in C_1, O_j \in C_2} \text{Distance}(O_i, O_j)}{|C_1||C_2|} \]

The proposed optimization process is looking for a clustering solution with best quality subject to the investor's risk preference. As the number of clusters should have inverse relationship with propensity for risk, Equation 13 shows multi-objectives of the optimization process in SMPOpt.

Equation 13 \textbf{Maximize Quality of Clusters}

\[ \text{Subject to } \left| \frac{k}{n} - (1 - r) \right| < \epsilon \]

where, \( n \) is the maximum number of clusters, \( k \) is the optimum number of clusters, \( r \) is a real number in the range \([0,1]\) showing the user's propensity for risk which is set by the user using a slider in Figure 11, and \( \epsilon \) is small boundary number.

Starting with clustering solution with \( k \) number of clusters, the algorithm is flexible to find a solution in range of \( \epsilon \) from \( k \) within maximum range of 2 to \( n \). \( n \) is the maximum number of clusters which for example can be equal to the number of available stocks (the extreme case of having each single stock as a cluster). The clustering validity analysis measures the quality of each solution by the average homogeneity of all clusters and average separateness of all pair of clusters. In case of finding a solution with acceptable quality, the algorithm stops.

Figure 12 shows an example of clustering solution. In this figure black dots are showing clustering solutions with different normalized number of clusters (number of clusters divided by the maximum number of clusters \( n \)). The highlighted area shows the narrow down solutions.
based on the value of $r$ and $\varepsilon$. The final decision on the best clustering solution is made by considering the quality of clustering solutions in this area. Homogeneity and separateness are both measured and considered. By defining acceptable threshold for one of them, the highest value of the other is picked as the final solution. The number of clusters in the selected solution ($k$) indicates the optimum number of clusters.

![Figure 12 - Example of mapping between propensity for risk and optimum number of clusters](image)

4.2.2.2 Clustering Algorithms

After determining the required number of clusters, it is proposed to apply an existing clustering technique to find appropriate clusters. Two algorithms implemented in SMPOpt are briefly explained next.

**K-Means Clustering:** K-means algorithm ([57]) is applied on the feature vector of companies extracted from the constructed social network. The connectivity among stocks in the network defines a feature vector for each stock. For example, in the case of three stocks $\{S_1, S_2,$
$S_3$} where $S_1$ is connected to $S_2$ and $S_3$ with weights 0.4 and 0.3, respectively, and $S_2$ and $S_3$ are unconnected - the feature vectors for these stocks would be as follows:

\[
S_1 = (1, 0.4, 0.3) \\
S_2 = (0.4, 1, 0) \\
S_3 = (0.3, 0, 1)
\]

These feature vectors are utilized by the K-means algorithm to assign stocks to clusters. The algorithm first selects $k$ stocks randomly to serve as initial group centroids. The distance (e.g., Euclidean distance) between pairs of stocks provides a measure for assigning each stock to the cluster with the closest centroid. When all stocks have been assigned, the centroid of each cluster is recalculated. This iterative process continues until cluster centroids have stabilized [57].

Figure 13 illustrates the k-means clustering results by SMPOpt to discover four clusters on the Dow Jones network in 2004.

![Figure 13 - K-means clustering results for sample Dow Jones Index](image)

**Louvain Clustering:** This is a network based clustering algorithm proposed by Blondel et al. [58]. This method uses a heuristic for finding communities to optimize the modularity of the
whole network. The method measures the quality of the partitions using their modularity. The modularity of a partition is a scalar value between -1 and 1; it is calculated by Equation 14 [58].

**Equation 14**  
\[
Q = \frac{1}{2m} \sum_i \sum_j [A_{ij} - \frac{k_i k_j}{2m}] \delta (C(i), C(j))
\]

where, \(A_{ij}\) is the weight of the edge connecting nodes \(i\) and \(j\); \(k_i\) is the sum of the weights of edges having \(i\) as an end; \(C(i)\) is the community of node \(i\); \(m = \frac{1}{2} \sum_i \sum_j A_{ij}\) and \(\delta(u, v)\) is 1 if \(u = v\) and 0 otherwise [58].

The Louvain method has two phases which are repeated iteratively. At the beginning of the algorithm, each node will have its own cluster. So the number of initial partitions is equal to the number of nodes. Then, for each node \(i\), it is assumed that it is removed from its community and added to the community of one of its neighbours. For each neighbour, the modularity gain of this transfer is computed and the node will be placed in the cluster of the neighbour with the highest positive modularity gain. This process is repeated until no improvement can be achieved [58]. In the next phase, a new network will be built from the original one. Nodes of the new network will be the communities found in the previous phase. The weights of the links between new nodes will be the sum of the weights of the edges between nodes of corresponding communities. Links between nodes of the same community will create self loops. These two phases will be repeated until no further changes occur [58].

As explained above, Louvain algorithm has its own optimization process to find the optimum number of clusters without considering the user’s propensity for risk. However, as explained before the specific level of diversification is determined based on users risk level objective and thus SMPOpt might not be satisfied by the number of clusters from Louvain.
algorithm. To deal with this issue, this algorithm is applied in a hierarchical manner to be able to reach a desirable level of diversification.

Figure 14 illustrates the results of running Louvain method for two levels on the network of Dow Jones stocks from 1990 to 1999. In this example, three clusters are created in the first level. The third cluster was identified as a large cluster, and thus it was divided into three sub-clusters in the next level.

<table>
<thead>
<tr>
<th>Level</th>
<th>Clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>C1</td>
</tr>
<tr>
<td></td>
<td>C2</td>
</tr>
<tr>
<td></td>
<td>C3</td>
</tr>
<tr>
<td>L2</td>
<td>C1</td>
</tr>
<tr>
<td></td>
<td>C2</td>
</tr>
<tr>
<td></td>
<td>C3</td>
</tr>
<tr>
<td></td>
<td>C4</td>
</tr>
<tr>
<td></td>
<td>C5</td>
</tr>
</tbody>
</table>

**Figure 14 - Louvain clustering results for sample Dow Jones Index**

### 4.2.3 Stock Ranking

By constructing the social network of companies and discover appropriate number of clusters, based on the investor’s propensity of risk, SMPOpt tries to reduce the risk of investment (first objective) by investing in stocks from various clusters. To select appropriate stocks from each cluster, SMPOpt tries to increase the expected future return of investment (second objective). For this purpose, the stocks ultimately assigned to each cluster are then rank ordered in stage three of the proposed process called stock ranking.
Figure 15 reports inputs and outputs of this component. List of the stocks to be ranked is required to be specified by user. The user can select any list of stocks such as stocks in each cluster or all available stocks. In addition, there are three other optional inputs from the investor; list of factors, target rank score, and learning algorithm. Investors can set these parameters or may leave it to the system to use the default values. The historical value of internal and external factors for all selected stocks need to be retrieved from financial data sources. The output of this component is the ranked list of stocks, which is shown to the user in text or visual colour hue on the network.

![Figure 15 - Inputs and Outputs of Stock Ranking Phase](image)

Required internal and external factors are imported from financial data sources such as *Yahoo! Finance* and *Bloomberg*. Datasets from these sources are first stored in *csv* files and then imported in the local SMPOpt database. Figure 16 shows import menu in SMPOpt through which *csv* files can be transferred to local database.

![Figure 16 - Data Import Menu in SMPOpt](image)
Available factors which have been imported by the user to the system are listed as shown in Figure 17 (see Internal and External Factors tabs). Names of the features shown to the user are the same as the imported names. In Figure 17 these names are the mnemonics imported from Bloomberg. Here is the place that the user can select which features to be used for further analysis.

![Figure 17 - List of Internal and External Factors Selection GUI in SMPOpt](image)

4.2.3.1 Stock Ranking Problem Formulation

Given $n$ available stocks in the market, the stock ranking problem is to sort these stocks at the current date according to their expected quality in the future. This task can be perform by evaluating the performance of each stock and sort them based on the value of their performance metrics.
As shown in the literature, valuating the stocks is a multi-criteria problem and various indicators from the companies can be considered for this purpose. Given \( m \) features for each of the \( n \) stocks, and the target ranking score of stocks at each point of time in the history starting from \( t_0 \), the learning task can be defined to find a relationship between the values of these features \( X_{ij} \) and the target rank scores \( r_{ij} \). Thus, the training set \( E_{train} \) for the ranker can be formed as shown in Equation 15.

**Equation 15** \( E_{train} = \{(t_b, X_{ij}, r_{ij}) : 0 \leq i \leq c, 1 \leq j \leq n | X_{ij} \in R^m, r_{ij} \in \{1, 2, \ldots, n\}\} \)

This training set is used to train a model for ranking the stocks at the next time point \( t_c \). The model can be evaluated with the test set \( E_{test} \) containing the ranking of the stocks at this time point (Equation 16).

**Equation 16** \( E_{test} = \{(t_c, X_{cj}, r_{cj}) : 1 \leq j \leq n | X_{cj} \in R^m, r_{cj} \in \{1, 2, \ldots, n\}\} \)

In the above equations, \( r_{ij} \) is the target rank score of stock \( j \) at time \( i \) (time point \( t_i \)), which means for this time stamp, \( r_{ia} < r_{ib} \) if stock \( a \) has higher performance than stock \( b \). The target rank score can be defined by the user. The default value of target rank score in SMPOpt is the future return value. Equation 17 ([5]) shows the formula to calculate the future return of stock \( j \) at time \( i \), based on the stock price at time \( i \) and time \( i+\Delta t \). In this equation \( P_{j,i} \) is the price of stock \( j \) at time \( i \) and \( D_{j,i} \) is the dividend paid by stock \( j \) at time \( i \). \( \Delta t \) is the future return term which is usually equal to 1, 3 or 6 months.

**Equation 17** \( R_{j,i} = \ln \left( \frac{P_{j,i+\Delta t} + D_{j,i+\Delta t}}{P_{j,i}} \right) \)

Given a future return term \( \Delta t \) as the target ranking criteria, the rank score of each stock is calculated by comparing future return of the available stocks. The stock with the highest return
value is assigned to 1, and lower values obtain higher rank scores until the last one whose rank score can be \( n \) at most. Two stocks with the same values of return are assigned equal rank score.

**Definition:** The target stock ranking \( r^* \) is a set of pairs of stock feature vectors such that 
\[
(X_{ia},X_{ib}) \in r^*_i \text{ if } r_{ia} < r_{ib}.
\]

To build the training set, the target ranking of \( n \) stocks at each point of time \( i \), from \( t_0 \) to \( t_c \), should be first constructed. However, it is not possible to calculate the rank score of the stocks in points of time \( i \) when \( i+\Delta t_r > t_c \), i.e., \( P_{i+\Delta t} \) and \( D_{i+\Delta t} \) are unknown for all the stocks. Therefore, the training samples are those in the interval \([t_0, t_c-\Delta t] \), which is called the valid target stock ranking period. Constructing the training set from all non-holiday dates in this period may result in a very large database. As the financial indicators from the companies are reported in the basis of regular period such as 6 months or one year, we can sample that database with a fixed gap interval. Given a sampling gap \( \Delta t_s \), the training set contains time samples of \( t_0+k\Delta t_s \), where the time sample appears in the valid target stock ranking period.

**Definition:** The stock ranking training sample set is a set of pairs of time samples and their corresponding target stock rankings. Given \( t_0 \) as the start date of the history, \( t_c \) as the current date, \( \Delta t_r \) as the future return term, and \( \Delta t_s \) as the sampling gap interval, the stock ranking training set is constructed based on Equation 18.

**Equation 18** \[
\{(t_0 + k\Delta t_s, r^*_k\Delta t_s): 0 \leq k \leq \left\lfloor \frac{(t_c-t_0)-\Delta t_r}{\Delta t_s} \right\rfloor \}
\]

Given a stock ranking training sample set of size \( K \), in the stock ranking problem, the learning algorithm will select a ranking function \( f \) from a family of ranking functions \( F \) that maximizes empirical \( \tau \) (Equation 19) on the training sample. In this equation, \( \tau \) is Kendall’s tau coefficient which is a statistic that measures the association between two ranked lists [59].
Equation 19  \( \tau_{\text{Stock Ranking Training Set}}(f) = \frac{1}{k} \sum_{i=1}^{k} \tau(r_f(t_i), r_i^\ast) \)

Thus, \( f \) will be chosen to maximize the number of correct pairwise rank comparisons. This problem is identical to the problem for optimizing a search engine based on click-through data described in [26] where SVMRank was proposed as a solution. This approach is applied to stock ranking problem.

4.2.3.2 Stock Ranking using SVMRank

Considering the class of linear ranking functions (Equation 20), \( \hat{w} \) is a weight vector that is adjusted by learning [26].

Equation 20  \((X_{ta}, X_{tb}) \in f_{\hat{w}}(t_i) \iff \hat{w} \cdot X_{ta} > \hat{w} \cdot X_{tb}\)

Figure 18 illustrates how the weight vector \( \hat{w} \) determines the ordering of five stocks in a two dimensional example (having two features). For any weight vector \( \hat{w} \), the points are ordered by their projection onto \( \hat{w} \) (or, equivalently, by their signed distance to a hyperplane with normal vector \( \hat{w} \)). In the example shown in Figure 18, based on \( \hat{w} \) adjusted by training samples, the five stocks in testing samples are ordered as \( (msft, wmt, ibm, jnj \text{ and } mcd) \).

![Figure 18 - Example of how weight vector \( \hat{w} \) rank 5 stocks](image-url)
Adjusting weight vector \( \bar{w} \) to maximize the empirical \( \tau \) (Equation 19) is equivalent to finding this weight vector so that the maximum number of the following inequalities with different time samples \( i \) in the training set is fulfilled [26].

**Equation 21** \( \forall (X_{ia}, X_{ib}) \in r_i^*: \bar{w}.X_{ia} > \bar{w}.X_{ib} \)

Unfortunately, a direct generalization of this optimization problem is shown to be NP-hard [26]. However, similar to classification SVMs, it is possible to approximate the solution by introducing (non-negative) slack variables \( \xi_{iab} \) and minimizing the upper bound \( \sum \xi_{iab} \) [26]. Adding SVM regularization for margin maximization to the objective leads to the optimization problem in Equation 22. In this equation \( C \) is a parameter that allows trading-off between the empirical error (\( \sum \xi_{iab} \)) and the regularized term (\( \frac{1}{2} \bar{w} \cdot \bar{w} \)).

**Equation 22**\[ \text{Minimize: } \frac{1}{2} \bar{w} \cdot \bar{w} + C \sum_{i=1}^{k} \xi_{iab} \]

**Subject to:** \( \forall i \in \{1, 2, ..., k\} \)

\[ (\forall (X_{ia}, X_{ib}) \in r_i^*: \bar{w}.X_{ia} \geq \bar{w}.X_{ib} + 1 - \xi_{iab}) \]

The above optimization problem is equivalent to the SVM classification problem. This is solved by transforming into a Lagrangian (more details can be found in [60]).

In addition to linear ranking functions, the kernel trick [60] can also be applied to deal with the non-linearity in stock ranking problems. For this purpose, the input vectors are first mapped to a hidden, high dimensional feature space before the construction of the optimal hyperplane [60]. Various types of kernels such as **Polynomial** and **Gaussian Radial Basis Function** can be applied for this purpose.
4.2.3.3 Time Series SVMRank

As shown in Equation 22, the slack variable $\xi_{iab}$ has equal weight $C$ for all time values $i$ as the error between the predicted and actual rank score values. As the value $C$ is a parameter that determines the trade-off between the empirical error and regularized term, by increasing this value the relative importance of the empirical error with respect to the regularized term grows [55].

Based on the prior knowledge that in the financial time series the recent past data could provide more important information than the distant past data, it is proposed that the empirical error from the recent data points should have higher weights in Equation 22. As discussed in Section 4.2.1.2 appropriate ascending function such as Equation 10 can be used as the time weighting function. After deciding on the appropriate weight function, $W(i)$, the optimization problem can be modified from Equation 22 to Equation 23. The modified ranking algorithm based on Equation 23 is called TS_SVMRank.

**Equation 23**

\[
\text{Minimize: } \frac{1}{2} \overrightarrow{w}. \overrightarrow{w} + \sum_{i=1}^{k} W(i). \xi_{iab} \\
\text{Subject to: } \forall i \in \{1, 2, ..., k\} \\
\quad (\forall (X_{ia}, X_{ib}) \in r_i^*: \overrightarrow{w}. X_{ia} \geq \overrightarrow{w}. X_{ib} + 1 - \xi_{iab})
\]

Figure 19 shows stock ranking setting GUI in SMPOpt. Stocks in each clusters can be selected to be ranked independent to other clusters. As Figure 19 illustrates, two options are available for stock ranking. The first one is the proposed approach based on the original and modified version of SVMRank, and the second one is the basic ranking based on single feature. The second approach is used as the benchmark which explained in more detail in Chapter Six.
As explained in this section, ranking process based on SVMRrank needs several parameters. In SMPOpt, after choosing the appropriate stock and feature sets, the user is asked to provide parameters such as $t_0$, $t_c$, $\Delta t_s$ and $\Delta t_r$. These parameters are shown in Figure 19, respectively, by Start of History, Current Date, Training Gap (in the Ranking Setting window) and Term (in the Return Setting window). Data Weighting is the place where user can pick to apply SVMRank or TS_SVMRank. In this process, the user may require specifying other parameters as well. For example, the Rank Score in the Ranking Setting window, and the Method option and the Feature combo list in the Return Setting window make it possible for the user to apply Equation 17 to construct the target rank score based on the future return.

![Figure 19 - Stock Ranking and Return Setting GUI in SMPOpt](image)

Available financial data stored in the database, in addition to the above input settings from the user are then used to build the training and testing sets. The result of the stock ranking is visualized to the user. For this purpose a mapping between the predicted rank scores to the colour of the stock nodes is performed. The more high-lighted stocks show the ones with the
lower rank scores (the top stocks in the ranked list). These stocks are more suggested to the users to be added to their portfolio.

Figure 20 shows the ranking result visualization for Dow Jones example in SMPOpt. The results are based on two financial indicators, Revenue Per Share and Profit Margin, from Jan. 2004 to Dec. 2004 (detail description of these features are explained in Chapter Six). The hue of colours in each cluster represents within cluster ranking of that particular stock.

4.2.4 Portfolio Construction

As discussed before portfolio construction has two steps of stock selection and stock weighting. Up to this point, the proposed heuristic solution for the multi-objective portfolio optimization problem has located the highest performing stocks among weakly correlated clusters – a requisite condition for portfolio diversification. Once pairwise stock correlations are generated through social network analysis, the clustering algorithm is employed to reveal weakly correlated clusters of stocks (addressing the objective to minimize portfolio risk). Application of SVMRank yields a selection of stocks within each cluster (addressing the other objective to
maximize expected future returns). SMPOpt presents the highest performing stocks in each cluster for user consideration in formulating the portfolio.

Time complexity of the proposed stock selection in SMPOpt by selecting top stocks from each cluster is quadratic of the number of stocks; \( O(tn^2) + O(kn^2) + O(tfn + tn \log(tn) + tn^2) \), respectively, for social network construction, k-means clustering, and SVMRank ranking ([61]), where \( n \) is number of stocks, \( t \) is number of time stamps in the historical period of time, \( k \) is number of clusters, and \( f \) is the number of internal and external factors in ranking training process. This time complexity is much more efficient than constructing and evaluating all the \( 2^n \) possible subsets of stocks.

Figure 21 provides a snapshot of the Graphical User Interface in SMPOpt for the selection of stocks in the portfolio setting. Investors can select the stocks presented by SMPOpt or choose other stocks with an understanding of relative ranking within the group.

![Figure 21 - Portfolio Setting GUI in SMPOpt](image)

The final step in portfolio construction is stock weighting to ascertain how many shares of each stock should be purchased (or traded). Typically stock weightings correspond to an individual stock’s Market Capitalization, Return or key financial ratios such as Price/Book or
Price/Earning [5]. Market capitalization (market cap) represents the total value of a company’s stock and is considered a proxy for the company’s net worth. It is measured as the product of the number of outstanding shares and the share price. The default method of weighting in SMPOpt is market capitalization, although users can specify alternative weighting, as seen in Figure 21.

Figure 22 presents the portfolio constructed for the Dow Jones example, where the highest performing stocks in each of four clusters is weighted according to its market capitalization. As an example, assume that we have two clusters. The selected stock from the first cluster has 12% of the whole market and the selected stock from the second cluster has 5% of the whole market. The user can invest $12 / (12+5) \%$ of the funds on the first stock and $5 / (12+5) \%$ of the funds on the second stock.

**Figure 22 - Portfolio of Weighted Dow Jones Stocks**

### 4.3 Financial Data Pre-Processing

Data pre-processing is an often neglected but important step in the data mining process. It includes cleaning, normalization, transformation, feature extraction and selection, etc [56]. Two basic types of pre-processing techniques are implemented in SMPOpt to deal with 1) the historical *missing values* of financial features and 2) *feature selection*.

The historical financial dataset imported from sources such as Yahoo Finance! and Bloomberg might have some missing values. To deal with these missing values, a basic method is applied in SMPOpt to replace the missing value of feature $i$, of stock $j$, at time $t$, with the value
of the same feature of the same stock in the closest feature vector. The closest feature vector can be defined based on time closeness or feature vector similarities. In the case of time closeness, starting from time $t$ the missing value is replaced by the first available value of feature $i$ in either past or future direction in the period of $[t-\Delta t, t+\Delta t]$ to time $t$, where $\Delta t$ can be set as a parameter.

In the case of feature vector similarity, the most similar feature vector considering other available features than $i$ (the missed one) is chosen and missing value is replaced by the corresponding value of feature $i$.

As there are a lot of indicators defined by financial experts about companies, all quantitative models for stock selection and portfolio management face the challenge of determining the most efficacious indicators from the large amount of available financial data. Similar to other models, the result of the proposed portfolio optimization technique (in the ranking step) closely depends on the list of features used in the training set. Next it is explained how SMPOpt can guide investors in selecting an appropriate list of features based on the historical ranking accuracies gain from the proposed technique.

### 4.3.1 Stock Feature Selection

Although in the implemented system SMPOpt, it is very user friendly to select appropriate features according to the investor’s preferences, it is hard to decide on the most appropriate set. Therefore, feature selection techniques can be useful for this purpose.

The general wrapper feature selection methods use a predetermined classification algorithm to search for an optimal subset of features, whereby the quality of a particular subset is evaluated by measuring the accuracy of the classifier model [62]. In this work it is proposed to apply a wrapper technique based on the *Normalized Discounted Cumulative Gain* (NDCG) of the ranking model to select the best subset of available financial features.
NDCG is typically used for evaluating search engines by comparing the predicted ranking with the ideal ranking defined by the user who is searching for specific information [63]. The target ranking in the form of user-provided relevance scores $rel_i$ is used for computing the Discounted Cumulative Gain (DCG), which measures the usefulness or gain of documents based on their position $i$ in the result list (Equation 24).

**Equation 24** \[ DCG = \sum_{i=1}^{n} \frac{2^{rel_i}-1}{\log_2(1+i)} \]

By sorting the documents in the result list based on their relevancy (user's feedback), the Ideal DCG (IDCG) can be measured, which is then used to normalize DCG and calculate NDCG (Equation 25).

**Equation 25** \[ NDCG = \frac{DCG}{IDCG} \]

In the stock ranking evaluation, *Future return* is used as the relevancy of each stock. To avoid negative relevancies, *Normalized return* (between 0 and 1) is used instead. DCG of the stock ranking is calculated based on Equation 24 by replacing $rel_i$ by the *Normalized future return* of stock $i$. To clarify the proposed technique for constructing the stock ranking training and test sets and to measure the performance of the predicted ranking, a simple example is shown in Figure 23.

![Figure 23 - Example ranking training and test sets](image-url)
In this example, there are three stocks ($S_1$, $S_2$ and $S_3$) with two features ($f_1$ and $f_2$). The training set contains the values of these features for all three stocks from $t_0$ to $t_k$ with the gap interval equal to $\Delta t_s$. In this training set, the target rank scores are calculated based on future return of the stocks after $\Delta t_r$. Therefore, the last point of time in the training set is $t_k$ which is $\Delta t_r$ before the current date ($t_c$). The SVMRank learner then uses this training set to build the ranking model. Then the SVMRank classifier uses the built model to predict the ranking score of the stocks in the testing set. The target rank score of the testing set is calculated based on the future return of the stocks at $t_c$ which for $S_1$, $S_2$ and $S_3$ are, respectively, 0.4, -0.7 and 0.6. Thus normalized return for these stocks are, respectively, 0.7, 0.15 and 0.8, and the NDCG of the predicted ranking is calculated as follows:

$$NDCG = \frac{z^{0.7}_{-1} + z^{0.8}_{-1} + z^{0.15}_{-1}}{\log 2 + \log 2 + \log 2} = 0.98$$

As calculating NDCG for the testing set at the current date ($t_c$) requires the target rank score which is obtained based on the price of the stocks after $\Delta t_r$, for the current time it is not possible to calculate the accuracy. However, this required information would be available after passing $\Delta t_r$. If the proposed technique is dynamically applied over time to rank the stocks, the accuracy of past ranking experiences can be calculated and used for the purpose of selecting a better feature set.

Given $t_c$ with known target rank score, the proposed technique tries to find a subset from $m$ available features which results in the highest NDCG. The brute-force way would be building one training set for each of the $2^m$ possible feature subsets, calculating the NDCG of each of them, and picking the one with the highest NDCG. Instead, the feature weights calculated by
SVMRANK are used in the process of building the ranking model. This feature subset, called \textit{MaxNDCGFeaturesSet}, is found based on Algorithm 1 [64].

\begin{table}[h]
\centering
\begin{tabular}{|l|}
\hline
1: Put all $m$ available features in the set $InitialFeatureSet$  \\
2: Build training set containing all features in $InitialFeatureSet$ and run SVMRank to calculate the weights of the features  \\
3: While $InitialFeatureSet$ is not empty  \\
4: Remove the least weighted feature from $InitialFeatureSet$  \\
5: Build training set containing all features in $InitialFeatureSet$ and run SVMRank to calculate the weights of the features  \\
6: Run SVMRank classifier to predict the ranking scores  \\
7: Calculate NDCG  \\
8: Return the subset with maximum NDCG  \\
\hline
\end{tabular}
\caption{MaxNDCGFeatureSet selection at $t_c$.}
\end{table}

Essentially, \textit{MaxNDCGFeaturesSet} is a wrapper algorithm that selects an optimal set at $t_c$. Given $k$ past ranking experiences ($t_{c1}, t_{c2}, \ldots, t_{ck}$) with known target rank scores, the final selected feature set is obtained by Equation 26.

\textbf{Equation 26} \hspace{1em} \textit{FeatureSet} ($t_c$) = $\cup$ \textit{FeatureSet} ($\{t_{c1}, t_{c2}, \ldots, t_{ck}\}$)

In this Equation, \textit{FrequentSet}($\{t_{c1}, t_{c2}, \ldots, t_{ck}\}$) returns features that frequently appear in \textit{MaxNDCGFeaturesSet} of $t_{c1}$ to $t_{ck}$. A feature is frequent if it satisfies a minimum frequency threshold called support ($s$). In other words, \textit{FrequentSet} finds frequent features that appear in at least $\%s$ of all \textit{MaxNDCGFeaturesSet} from the past ranking experiences. Finally, the feature set is the union of all frequent features with respect to the value of support $s$. 

60
Chapter Five: Second Solution in SMPOpt: Managed Portfolio Recommendation

In the second strategy in SMPOpt it is proposed that constructing a social network of financial experts, based on their investment behavior, can be useful for further analysis with the purpose of investment recommendation to non-professional investors. As discussed before, various financial experts have different investment behavior and preferences and thus similarity between them can be defined based on how their preferences are similar to each other. In this strategy, a virtual social network is constructed containing financial experts as the actors and their investment preference similarities as the link between them.

An investor’s portfolio is an important source of information that can be used to extract human investment behavior. For instance, characteristics such as investor’s propensity for risk can be measured from his/her available portfolio. In this chapter, feature extraction about investors from their publicly available portfolios and the social network construction process based on these features are discussed. This social network is then used to categorize the experts into various clusters and to apply further analysis in the proposed recommendation process.

5.1 Investor’s Feature Extraction

In order to extract the features about professional investors, it is proposed to measure the performance of their portfolio. As discussed before, in portfolio performance analysis, return on an asset is the basic element employed to determine its performance [5]. For measuring the return, a period is assumed as an interval of time during which an asset is held without being modified, and companies make payments to its shareholders in terms of dividends at the end of the period. Return of stock i is calculated based on Equation 1.

Expected return on the stocks comprising the portfolio is then used to measure the expected return of the portfolio. Expected return of the portfolio of n stocks at time T based on
the assumption that the portfolio has a fixed composition throughout the evaluation period is measured by Equation 3 [5].

As discussed in modern portfolio theories, the concept of return is not sufficient on its own to analyze the performance of a portfolio [5]. To analyze portfolio performance more precisely, we need a quantitative measurement of risk. Equation 4 shows how Markowitz measures the **Risk of the portfolio**.

### 5.2 Social Network Construction

After extracting characteristic features about investors from their portfolios, the social network of investors is constructed based on one of the network construction approaches proposed and implemented in the SNA tool called NetDriller developed by our research group and to which I have contributed [65]. Next the approach used in this work which is based on actors clustering result is briefly explained.

In many applications, there are several items identified by a feature vector (a set of features). To apply SNA and obtain useful knowledge from the relationship that exists between the items, we need to construct a network of these items. To measure the similarity between each pair of items, it is proposed to apply data mining technique of clustering. The items are clustered by K-means algorithm ([56]) with different values of K (number of clusters). The number of common clusters of two items is then used to measure the similarity between them.

In the case of having \( n \) items, \( n-2 \) clustering solutions are constructed by applying the K-means algorithm with K set to 2, 3, ..., \( n-1 \), respectively. The similarity of two items \( a_i \) and \( a_j \) is calculated based on Equation 27. In this formula, \( \text{CommonCluster}(a_i, a_j, k) \) is 1 if in the clustering solution with \( k \) number of clusters, \( a_i \) and \( a_j \) appear in the same cluster, and is 0 otherwise. It is
also possible to receive specific range of number of clusters from the user instead of applying all possible $n$-2 clustering solutions.

**Equation 27**  
\[
\text{Similarity} \left( a_i, a_j \right) = \frac{\sum_{k=2}^{n-1} \text{CommonCluster}(a_i, a_j, k)}{n-2}
\]

To demonstrate how the above methodology works, it is applied on a simple example. Figure 24 shows four items with three features and the network constructed by NetDriller based on clustering solutions with 2 and 3 as numbers of clusters. 2 clusters are $\{a_1, a_3\}$ and $\{a_2, a_4\}$ and 3 clusters are $\{a_1, a_3\}$, $\{a_2\}$ and $\{a_4\}$. Therefore, $a_1$ and $a_3$ are in the same clusters in both clustering solutions and thus the weight between them is 1 (2 out of 2). $a_2$ and $a_4$ are in the same cluster in one of the clustering solutions and thus the weight between them is 0.5 (1 out of 2).

![Figure 24 - Network Construction based on Clustering](image_url)

Based on modern portfolio theory, higher return comes with higher risk. Thus by considering two features of return and risk (extracted from investor’s portfolio), the link’s weight in the social network (constructed as explained above) shows how similar two investors are in terms of user’s propensity for risk.

### 5.3 Investment Recommendation System

The constructed social network of expert investors is then used in the investment recommendation system. In this system the goal is selecting the best appropriate managed portfolio (built by an expert) and recommending it to the non-professional investor. The propensity for risk is the key characteristic to find the most similar expert to the investor. Next,
experts clustering and classification of the non-professional investors to assign him/her to one of the discovered clusters of experts are explained. After finding the most appropriate cluster for the investor, one of the experts in this cluster needs to be picked as the representative of the cluster whose portfolio is to be suggested to the investor.

5.3.1 Experts Clustering

As discussed in the social network construction section, the link between two experts is weighted based on various clustering solutions with various numbers of clusters. In this step one of these clustering solutions is considered as communities of experts discovered based on K-means algorithm.

The goal is detecting various clusters containing people with different propensity of risk to be able to label the clusters based on how risk takers are the people in each cluster. For this purpose the five Likert scale of very high, high, medium, low and very low is used. Distributions of the risk value in each of these clusters are then used to label each cluster. By accepting the assumption from modern portfolio theory on the trade-off between risk and return, it is believed that the risk value should be uniformly distributed among the five discovered clusters. This assumption is tested in the case study explained in the next section.

5.3.2 Non-Professional Investors Classification

The result of experts’ clustering is then used as the training set to learn a classification model with the purpose of assigning non-professionals to an appropriate expert community. Figure 25 illustrates a sample Training and testing sets containing 12 expert investors and 7 non-professionals. In this example two features of risk and return of both experts and non-professionals are extracted from their available portfolios.
Different classification techniques such as Support Vector Machine, Naïve bayes, or Decision Tree ([56]) can be used to build the classifier from experts’ communities.

5.3.3 Finding the Representative Expert in a Community

After assigning the non-professional investor to one of experts’ communities, a representative expert in that community is suggested to the investor. To find the representative person in a social network, SNA is applied by measuring centrality metrics. Within the scope of graph theory and network analysis, there are various types of measures of the centrality of a vertex within a graph that determine its relative importance. Degree centrality and closeness centrality are two measures that are widely used in network analysis [20].

Degree centrality is defined as the number of ties that a node has in a graph. It can be interpreted in terms of the immediate risk of a node for catching whatever is flowing through the network [20]. Closeness centrality is a measure to show how long it will take to spread information from a node to all other nodes in the graph sequentially. In graphs the distance metric between pairs of nodes is defined by the length of their shortest paths. The closeness of a
node is the inverse of how far the node is to all other nodes (sum of its distance to all other nodes) [20].

In each community of investors, the one with the highest centrality is selected as the representative of that community. To make the proposed recommendation system clear, the whole process is illustrated in an example. Figure 26 and Figure 27 show social network construction and further steps on investors shown in Figure 25.

**Figure 26 - Different Clustering solutions for sample experts in Figure 25**

Figure 26 shows 8 clustering solutions with $k$ from 2 to 9, on 12 expert investors ($a, b, c, ..., l$) in the above example. These clustering solutions are the result of k-means algorithm on a dataset containing two features of Risk and Return for the expert investors. More similar experts stay in the same cluster by increasing number of clusters. For instance, $k$ and $j$ stay in the same
cluster in all of the 8 solutions, while \( h \) and \( g \) are in the same cluster only in one solution out of 8.

The social network of experts is then constructed based on these clustering solutions (Figure 27 (a)). As an example, clusters with \( k=3 \) show people with high, mid and low risk (respectively, shown by red, yellow and green in Figure 27 (a)). These clusters are considered as different communities of experts used in further analysis in the recommendation system.

Whenever a non-professional investor looks for the most appropriate expert to mimic his/her investment behavior, it is very important to consider their preference similarities. Propensity for risk is an important aspect about the investor, which was the main focus to discover various communities of experts in the constructed social network.

![Figure 27 - (a) Experts’ Social Network and Communities (b) Assigning non-professionals to Experts’ communities](image)

Each non-professional is first assigned to one of the experts’ communities. For this purpose, required features about the investor (risk and return) are extracted from their available
portfolios, which are then used to construct the classification testing set. Figure 27 (b) shows the classification result of non-professional investors by the color that illustrates the assigned cluster.

The degree centralities of each expert investor are shown adjacent to the nodes in Figure 27 (a). The expert with the highest centralities is the one whose portfolio is recommended to the non-professional investor. For instance, in low risk community (green cluster) experts $c$ and $d$ have the highest centrality value (0.583), and thus the portfolio of either of them would be appropriate for a non-professional investor who has been assigned to this community (for example $A$ and $B$).
Chapter Six: SMPOpt Evaluation through Experimental Studies

The proposed techniques explained in Chapter 4 and 5 are evaluated through a series of experimental studies. In all of these studies, the idea of back testing are applied in which past points of time are simulated as the time of making investment decision and SMPOpt recommends investment guideline based on the data available up to that point. Using SMPOpt, it is possible to set the experimental points as the current time.

Table 1 lists all of the experimental studies performed in this thesis with brief explanation about their objectives. Each of these 8 studies focuses on a specific portion of the proposed methodology. For instance, the first study targets to evaluate social network construction and clustering with known number of clusters, while for ranking it applies simple traditional single criterion ranking. As the first strategy in SMPOpt (Chapter 4) has more components to be evaluated, most of these studies (all except experiment 7) target the first strategy and the only one which is evaluating the second strategy (Chapter 5) is experiment 7.

In all of them, real financial data on Stock Exchange Markets in US are used which are extracted from two sources of information; Yahoo! Finance and Bloomberg. As Table 1 shows, S&P 500 (Standard & Poor’s Composite Index of 500 Stocks) and Dow Jones are two market indices used in most of the setups.

Based on the objective of each of these studies, they create different kinds of results which lead to have different benchmarks in each study. Appropriate performance metrics are used in each to compare the performance of the proposed result with the benchmark(s).
<table>
<thead>
<tr>
<th>Experiment Setup</th>
<th>Evaluated Strategy</th>
<th>Applied Proposed Components</th>
<th>Proposed Result</th>
<th>Benchmark(s)</th>
<th>Performance Metric</th>
<th>Stock Market Exchange</th>
<th>Time Period</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1</td>
<td>Ranking: SVMRank</td>
<td>Ranked List of Stocks</td>
<td>Ranked List of Stocks based on CAPM</td>
<td>NDCG</td>
<td>Dow Jones</td>
<td>2004-2006</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>Ranking: SVMRank, TS-SVMRank</td>
<td>n-Top Equal Weight Portfolio</td>
<td>All Market Equal Weight Portfolio</td>
<td>Sharpe Ratio</td>
<td>S&amp;P 100</td>
<td>2002-2010</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>Ranking: SVMRank Feature Selection</td>
<td>Ranked List of Stocks based on all features</td>
<td>Ranked List of Stocks based on all features</td>
<td>NDCG</td>
<td>Dow Jones</td>
<td>2004-2009</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>Network: Clustering based Clustering: K-Means (n=5)</td>
<td>Suggested Managed Portfolio</td>
<td>Other Managed Portfolio</td>
<td>Sharpe Ratio</td>
<td>StockPickr Stocks</td>
<td>2012</td>
</tr>
</tbody>
</table>
6.1 Performance Metrics

Different performance metrics used in the experimental studies are explained here.

6.1.1 Sharpe Ratio

As discussed before, the expected return value on its own is not a sufficient criterion for assessing the performance and it is necessary to associate a measure of the risk as well [5]. In other words, expected return alone only enables comparison between portfolios with the same level of risk, while we need a risk-adjusted performance value [5].

For measuring the performance of a portfolio, Sharpe (1966) has defined a reward-to-variability ratio which is known as the Sharpe Ratio [5]. The purpose of this metric is to associate a measure of the portfolio risk with the portfolio return. Equation 28 illustrates the formula to calculate the Sharpe Ratio of a portfolio at time $T$.

**Equation 28**

$$ S_{p,T} = \frac{E_{R_{p,T}} - R_{RF,T}}{\sigma_{p,T}} $$

where, $E_{R_{p,T}}$ is the expected return of the portfolio at time $T$; $R_{RF,T}$ is the risk-free rate at time $T$ (US Treasury Bill rate is used as the risk free rate); $\sigma_{p,T}$ is standard deviation of the portfolio at time $T$.

The Sharpe ratio measures the amount of return added to the portfolio per unit of risk [35]. This is a popular performance metric for comparing the managed portfolio with benchmarks [5]. Higher Sharpe ratio shows portfolios with higher amount of return which is expected to be added to the portfolio per unit of risk.

6.1.2 Future Return

The reason of defining future return of the portfolio (Equation 3 based on Equation 17) as a performance metric is to consider the return in future time as the investment objective.
However as discussed before return is not enough and risk concept should not be forgotten. However, in one of the experimental setup (experiment 6), in addition to Sharpe Ratio, Future Return of the constructed portfolio and the benchmark are calculated in order to compare their future status.

6.1.3 NDCG

This performance metric (Equation 25) is used to measure the performance of a ranked list in order to compare the performance of two ranked lists. As can be seen in Equation 24, DCG gives a higher weight to lower-ranked items. This is desirable for the ranking for search results because here only the top-ranked documents are really important for the user. Similarly, for the ranking of stocks, it is much more important to get the ranking of the top-most stocks correct than the ranking of uninteresting stocks near the bottom of the ranking. Thus, it is decided to adopt this performance measure instead of alternative measures like Kendall’s tau or Spearman rank correlation, which give equal importance to all ranks.

6.2 Benchmarks

Different benchmarks used in the experimental studies are explained here.

6.2.1 All Market Portfolio (Maximum Diversification)

The experimental study described in [31] illustrates that despite the sophisticated theoretical models developed in the last 50 years, none is consistently better than the naïve $1/N$ rule in which a fraction $1/N$ of wealth is allocated to each of the $N$ available assets. Their conclusion was based on the comparison of the portfolios’ performance constructed from the $1/N$ rule as well as from 14 financial models across seven empirical datasets. Therefore, in this thesis, the $1/N$ rule is picked as one of the benchmarks to compare the performance of the proposed portfolio with a portfolio containing all stocks. Such a portfolio has maximum possible
diversification. In addition to the portfolio with equal weights on all the stocks, Market-Capitalization is also applied to build Market-Cap all-market benchmark.

6.2.2 Single Criterion Stock Ranking

This benchmark follows the traditional method of stock ranking based on a single criterion without involving any learning process. After ranking the stocks based on single criteria (for instance return or risk-adjusted return based on CAPM), n-Top stocks are picked to construct the portfolio.

6.2.3 Static Diversification

Global Industry Classification Standard (GICS) clusters the companies based on their business type [66]. In this standard, for instance companies of S&P500 market are divided into 10 sectors of Energy, Materials, Industrials, Consumer Discretionary, Consumer Staples, Health Care, Financials, Information Technology, Telecommunication Services, and Utilities. Using these clusters as the static communities of companies, diversified portfolio is constructed as a benchmark.

6.2.4 Managed Portfolio

If finance experts apply their personal capabilities and intelligence to solve the problem based on their knowledge of theories and existing strategies, investment decision results constitute what is called a managed portfolio. Publicly available stocks of financial experts reported in StockPickr website[^1] is used as a benchmark in this thesis.

[^1]: [http://Stockpickr.com](http://Stockpickr.com)
6.3 Experiment Setups and Results

In this section details on each experimental setup as well as the results, in terms of comparison of the performance metric for the proposed result with the benchmark, are reported.

6.3.1 Experiment 1[67]

In this experimental study, the main objective is to evaluate initial phases of the first strategy in SMPOpt. Social network of the companies are built and used to discover the required number of clusters based on Louvain algorithm. However, instead of the learning phase for stock ranking, three simple selection criteria are used; (1) highest eigenvector centrality, (2) highest expected return, and (3) expected return over a threshold. Finally, Market Capitalization is used for stock weighting.

The stock performance metric used for network construction in this study is 6 month return (Equation 1 with $\Delta t = 6$ month) and the link weight is calculated based on correlation coefficient (Equation 5). In this study, the above steps are applied on New York Stock Exchange (NYSE). Each market is represented by an aggregated value of its stocks called Market Index. S&P 500 (Standard & Poor’s Composite Index of 500 Stocks) is a medium-sized index made up of 500 stocks that are traded on NYSE [5]. The experiment is on 404 stocks available from April 1998 to April 2009 in this market. 13 different time intervals are defined using a sliding window with length 5 years at each experimental interval and the next interval is gained by shifting the time window 6 months forward. In other words, the first experimental period is from April 1998 to April 2003, the second is for the period from October 1998 to October 2003, and so on until the last one which is for the period from April 2004 to April 2009. In this study, clustering is applied to detect 10 clusters. The reason for picking 10 as the required number of clusters is that according to the Global Industry Classification Standard (GICS), companies comprising the
S&P500 Index are divided into 10 sectors. Threshold value that is used in the third selection method is set to 0.2.

Eigenvector centrality is a measure of the importance of a node in a social network. To each node of a graph a relative score is assigned, in a manner that nodes with connections to highly-connected nodes get higher rank than those with the same number of connections to low-scoring nodes. In the first proposed subset selection method, the node with the highest eigenvector centrality in each cluster is added to the portfolio. Selected stocks of the second portfolio are the ones with the highest expected return in their own clusters. The expected return of a stock is obtained by calculating the average of its returns over the time period. In the third portfolio, stocks are selected based on a threshold on their expected returns. If for a given cluster, there is no stock with an expected return value greater than the threshold, the stock with the highest expected return value will be selected from that particular cluster. Otherwise, all the stocks with the expected values greater than the threshold will be selected from the cluster. In this case, more than one stock might be selected from each cluster.

To further illustrate the three proposed selection techniques, Table 2 shows an example by listing the stocks in each cluster of Dow Jones Example. In Table 2, synthetic Eigenvector Centrality and Expected Return values are provided (X(a, b); where X is the stock’s name, a is Eigenvector Centrality, and b is Expected Return). The lists are sorted based on Eigenvector Centrality, bold Expected Returns are those higher than the threshold (0.2), and the bold underlined Expected Return is the highest return in each cluster. In this example, the above three methods, respectively, select \( P_1\{MCD, XOM, GE, BAC, HPQ\} \), \( P_2\{WMT, JNJ, AXP, DIS, HPQ\} \) and \( P_3\{MCD, WMT, PG, MSFT, JNJ, PFE, IBM, AXP, CVX, DIS, HPQ, UTX, BA, INTC\} \) as constructed portfolio.
Table 2 - Synthetic Eigenvector Centrality and Expected Return of Dow Jones stock

| C1  | MCD(0.6, 0.31), CSCO(0.31, 0.15), WMT(0.2, 0.33), HD(0.11, 0.13) |
| C2  | XOM(0.79, 0.17), PG(0.66, 0.22), MRK(0.61, 0.14), KO(0.58, 0.09), MSFT(0.53, 0.3), INJ(0.5, 0.43), PFE(0.49, 0.3), IBM(0.35, 0.28) |
| C3  | GE(0.9, 0.18), AXP(0.89, 0.3), CVX(0.34, 0.23), T(0.32, 0.15), VZ(0.28, 0.06) |
| C4  | BAS(0.98, 0.11), JMP(0.93, 0.16), DIS(0.81, 0.18), DD(0.81, 0.17), TRV(0.75, 0.09) |
| C5  | HPS(0.89, 0.36), UTX(0.78, 0.23), AA(0.75, 0.19), MMM(0.74, 0.16), BA(0.67, 0.2), CAT(0.62, 0.13), INTC(0.54, 0.33) |

The benchmark is to invest in a market capitalization-weighted portfolio of all the available stocks in the market. The performances of the three proposed portfolios as well as the benchmark are evaluated in terms of Sharpe Ratio in different time stamps in the history.

Figure 28 - Sharpe Ratio of three approaches and the benchmark in Exp 1

Figure 28 illustrates Sharpe Ratio of benchmark portfolio as well as portfolios from the three proposed selection techniques. This graph shows that, most of the time, portfolios based on the proposed approaches have higher Sharpe Ratio compared to the benchmark.

To statistically evaluate the result, hypothesis testing pair t-test is applied. The hypothesis is that the proposed approaches have higher Sharpe Ratio compared to the benchmark. Table 3 shows p-values of this test on each of the three methods. The result shows that the first and last methods improve the Sharpe Ratio with very high level of confidence, while the confidence of accepting the hypothesis on the second method is less.
Table 3 - p-values of Paired T-Test for Sharpe Ratios in Exp 1

<table>
<thead>
<tr>
<th></th>
<th>Benchmark</th>
</tr>
</thead>
<tbody>
<tr>
<td>Highest Eigenvector</td>
<td>0.002</td>
</tr>
<tr>
<td>Highest Expected Return</td>
<td>0.169</td>
</tr>
<tr>
<td>Expected Return more than thr</td>
<td>0.002</td>
</tr>
</tbody>
</table>

To illustrate the trade-off between portfolios risk and return, risk and expected return values of all the portfolios are shown in Figure 29. Figure 29 (a) shows that, as expected, the risk of the benchmark is less than all of three proposed approaches. The reason is that investing in all available stocks in the market leads to a very well diversified portfolio, while this approach does not necessarily lead to the highest expected return (Figure 29 (b)). By considering both objectives of having lower risk and higher expected return, the Sharpe Ratio of the portfolios are measured and compared as shown in Figure 28.

Figure 29 - (a) Portfolio Risk, (b) Portfolio Return in Exp 1

6.3.2 Experiment 2

The main objective of this experimental study, similar to experiment 1, is to evaluate initial phases of the first strategy in SMPOpt. However, the difference between this experiment
and the previous one is that this one is more focused on stock performance and link weight formulas used in the social network construction phase.

This experiment is performed on 336 stocks available from April 1998 to April 2009 in the S&P500 market. Four different social networks of these 336 stocks are constructed as follows; 1) \textbf{CorrReturn}: stock performance and the weight of links in this network are, respectively, calculated based on Equation 1 and Equation 5, 2) \textbf{CorrRiskAdjReturn}: weight of links is calculated by Equation 5 after replacing \( R_{it} \) by \( E(R_{i,T}) \) from Equation 6; 3) \textbf{WeightedCorrReturn}: weight of links in this network is calculated based on Equation 8 by using Equation 10 as the time weight function (with \( a=10 \) ); and 4) \textbf{WeightedCorrRiskAdjReturn}: weight of links is calculated by Equation 8 after replacing \( R_{it} \) by \( E(R_{i,T}) \) from Equation 6. To measure the original return value of each company in each day, the time interval (\( \Delta t \) in Equation 1) is set to 6 months. The constructed networks are then used to detect 10 clusters by the k-mean clustering algorithm (equal to the number of company sectors in GICS).

Two benchmarks are defined in this work to be compared with the proposed portfolio approach. As explained earlier, in traditional portfolio management, investor’s decision was made based on making more return on the investment without considering the concept of risk. The first benchmark used in this work is to implement the above traditional way by sorting the available companies based on their current return value and picking \( n \) top stocks from the list.

In addition to the above benchmark, the static clustering based on GICS standard is used to build a diversified portfolio by selecting one stock (with the highest return value) from each of the clusters. The difference between this benchmark and the proposed approach is that clusters of the companies are static over time and the dynamic correlations between companies are not
considered to discover the clusters. However, the strategy for picking one stock from each cluster is the same as in the proposed approach.

After network construction and clustering, one stock from each cluster is picked and market-capital-weighted portfolios are built from the selected subset of the stocks. There are four hypotheses as follows, which are evaluated through this experimental study:

**H1.** Considering both objectives of lower risk and higher return in investment decision making leads to have a portfolio with higher Sharpe ratio compared to a portfolio only based on higher return. In other words, Sharpe Ratio of the proposed approach is higher than the first benchmark.

**H2.** Diversification based on dynamic correlation between historical performances of the companies leads to a portfolio with a higher Sharpe ratio compared to a diversified portfolio based on static clustering that shows the business type of the companies. In other words, Sharpe Ratio of the proposed approach is higher than the second benchmark.

**H3.** In the proposed approach, using Risk-Adjusted return in the network construction instead of original return improves the Sharpe ratio of the result portfolio.

**H4.** In the proposed approach, measuring time series correlation in the network construction instead of original correlation improves the Sharpe ratio of the result portfolio.

To evaluate the above hypotheses, in this experiment, 12 different time intervals are defined using a sliding window with length 5 years at each experimental interval and the next interval is gained by shifting the time window 6 months forward. In other words, the first experimental period is from April 1998 to April 2003, the second is for the period from October 1998 to October 2003, and so on until the last one which is for the period from October 2003 to
October 2008. At each experimental point, 6 portfolios are constructed (4 proposed approaches and 2 benchmarks). Sharpe ratios of all these portfolios are then computed.

Figure 30 shows the minimum, maximum and the average of Sharpe ratios of each 6 portfolios over 12 experimental points. The averages of all the four proposed approaches are more than two benchmarks. Table 3 reports the pair t-test p-value of having greater Sharpe ratio from the proposed approaches compared to the benchmarks. These values show that it is not possible to reject the null hypotheses of H1 and H2 with acceptable level of confidence.

![Figure 30 - Max, Min and Average of Sharpe Ratios of proposed approaches and benchmarks in Exp 2](image)

**Table 4 - p-value of paired t-tests in Exp 2**

<table>
<thead>
<tr>
<th></th>
<th>Benchmark1</th>
<th>Benchmark2</th>
</tr>
</thead>
<tbody>
<tr>
<td>CorrReturn</td>
<td>0.078</td>
<td>0.005</td>
</tr>
<tr>
<td>CorrRiskAdjReturn</td>
<td>0.002</td>
<td>0.077</td>
</tr>
<tr>
<td>WeightedCorrReturn</td>
<td>0.037</td>
<td>0.095</td>
</tr>
<tr>
<td>WeightedCorrRiskAdjReturn</td>
<td>0.042</td>
<td>0.05</td>
</tr>
</tbody>
</table>

The highlighted row in Table 4 shows that WeightedCorrRiskAdjReturn has higher Sharpe ratio compared to both benchmarks. Figure 31 shows Sharpe ratio of all the 12 experimental points for this method as well as two benchmarks. As shown in Figure 31, except
for a period of time in 2005 and 2006, the Sharpe ratio of the proposed portfolio is higher than both benchmarks and in 2008 there is huge difference between them.

![Figure 31 - Sharpe Ratios over time in Exp 2](image)

Furthermore, to evaluate H3 and H4 directly paired t-test is applied. The p-value of CorrRiskAdjReturn has a higher Sharpe ratio than CorrReturn at 0.56 and the p-value of WeightedCorrRiskAdjReturn has a higher Sharpe ratio than WeightedCorrReturn at 0.32. For H4, the p-value of WeightedCorrReturn has a higher Sharpe ratio than CorrReturn at 0.17 and the p-value of WeightedCorrRiskAdjReturn has a higher Sharpe ratio than CorrRiskAdjReturn at 0.15. The p-value of t-test should be at most 0.05 (the commonly accepted level of confidence) to be able to reject the null hypothesis. In the above cases, as the p-value for both H3 and H4 are all more than 0.05, the null hypothesis cannot be rejected based on the experimental result of this work. Thus further experiment needs to be performed to be able to decide about these hypotheses.
6.3.3 Experiment 3

The main objectives in this experimental study is to show that using the proposed stock ranking technique based on SVMRank results in a higher NDCG compared to the benchmark.

In this study, the real financial data for 30 stocks from Dow Jones market in the US are used. The time period chosen for this experimental study is from 2003 to 2007, when there was less market shocks in the US. In this study internal factors about the companies are used to create the feature vector of the companies, while external factors from the outside world are not considered in this setup. Among several internal financial indicators available in this period of time for the above 30 stocks in the Bloomberg dataset, 11 of them are selected to be used; Profit margin, Return on asset, Return on capital, ROA to ROE, ROA based on bottom EPS, Revenue per share, Retention ratio, Total debt to total asset, Total debt to total equity, Total debt to total capital, and Total investment to total liability. Table 5 briefly describes these features. Each indicator is designed to illustrate some aspect of how the business is performing [1].

10 experimental points are defined starting from Jan. 2004 to Dec. 2006 with the interval of 3 months. At each point the ranking process is executed and the NDCG values obtained from the proposed approach with the benchmark are compared. The training set of each point is started from one year before the current time. For instance, the training set of April 2006 starts from April 2005. For each experimental point, the sampling gap interval, $\Delta t_s$, is set to three days. Three series of experiments are conducted with different values of the future return term, $\Delta t_r$, 1-month, 3-month and 6-month terms. The SVMRank algorithm with linear ranking functions is used in this study.

Furthermore, in this study, available stocks are ranked based on the expected return calculated by CAPM (Equation 6) as the benchmark method. In SMPOpt, in addition to ranking
stocks based on SVMRank, ranking based on a single performance metric have been implemented. SMPOpt allows a user to select a method of ranking. If a single performance metric is desired, it is possible for the user to select CAPM. The NDCG value of a ranking result is calculated and reported by SMPOpt.

Table 5 - Internal Factors used in Exp 3, 4, 6 and 8

<table>
<thead>
<tr>
<th>Bloomberg Mnemonic</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PROF_MARGIN</td>
<td>Indicates how much out of every dollar of sales, the company actually keeps in earning: Net Income / Revenue</td>
</tr>
<tr>
<td>RETURN_ON_ASSET</td>
<td>Quantifies the company's success of effort to earn a profit with respect to its total asset: Net Income / Total Assets</td>
</tr>
<tr>
<td>RETURN_ON_CAP</td>
<td>Quantifies the company's success of effort to earn a profit with respect to its capital: Net Income / (Total Assets - Current Liabilities)</td>
</tr>
<tr>
<td>ROA_TO_ROE</td>
<td>Quantifies the ratios of Return On Asset to Return On Equity (ROE: Net Income as a percentage of shareholders' equity): Shareholder's Equity / Total Assets</td>
</tr>
<tr>
<td>ROA_BASED_ON_BOTTOM_EPS</td>
<td>Indicates Return On Asset calculated based on the last line of the company's income statement. This reflects the fact that all expenses have already been taken out of revenues, and there is nothing left to subtract.</td>
</tr>
<tr>
<td>REVENUE_PER_SH</td>
<td>Indicates Revenue with respect to each share price. Revenue is the income that a company receives from its normal business activities, usually from the sale of goods and services to customers.</td>
</tr>
<tr>
<td>RETENTION_RATIO</td>
<td>Quantifies the percent of earnings credited to retained earnings: (Net Income - Dividends)/Net Income</td>
</tr>
<tr>
<td>TOT_DEBT_TO_TOT_ASSET</td>
<td>Quantifies company's financial risk by determining how much of the company's assets have been financed by debt: Debt / Total Assets</td>
</tr>
<tr>
<td>TOT_DEBT_TO_TOT_CAP</td>
<td>Quantifies company's financial leverage: Debt / (Shareholder's Equity + Debt)</td>
</tr>
<tr>
<td>TOT_DEBT_TO_TOT_EQY</td>
<td>Indicates what proportion of equity and debt the company is using to finance its assets: Total Liabilities / Shareholder's Equity</td>
</tr>
<tr>
<td>TOT_INVEST_TO_TOT_LIAB</td>
<td>Indicates total Investment of the company to the total liabilities: Total Investment / Total Liabilities</td>
</tr>
</tbody>
</table>

Figure 32(a) illustrates the box plot of the three series of experiments by using 1-month, 3-month and 6-month as the future return terms. The NDCG values of the stock ranking based on CAPM are also calculated for the above return terms. As Figure 32 (a) shows that the NDCG of
ranking based on SVMRank has higher average value compared to CAPM ranking in all the experiments.

Figure 32 – (a) Box Plot of SVMRank and CAPM, (b) Normal Probability Plot of NDCG in Exp 3

NDCG values of the three series of experiments are combined in Figure 32 (b). This figure shows the normal probability plot of all NDCG values obtained from SVMRank and CAPM. This figure shows that, with higher probability, SVMRank has higher NDCG compared to CAPM. For instance, with the probability of 95%, SVMRank leads to NDCG equal to 0.98 while CAPM leads to NDCG equal to 0.93.

The hypothesis testing of pair t-test is also performed to compare the NDCG of two methods. The hypothesis is SVMRank leads to rank the stocks with higher NDCG compared to CAPM. The pair t-test results in a p-value equal to 0.027, which shows that the null hypothesis can strongly be rejected, and thus the alternative hypothesis could not be rejected.

6.3.4 Experiment 4

The main objective of this experimental study is to evaluate the effectiveness of TS_SVMRank. For this purpose, both SVMRank and TS_SVMRank are applied to rank the stocks in S&P100 and then the portfolio is built from the top stocks. Past points of time are
simulated as the time of making investment decision and building the portfolio based on the data available up to that point. This simulation is repeated for 36 points from 2002 to 2010 (4 points from each 9 years). Historical price values as well as other financial metrics of the stocks in S&P 100 for the period from 2000 to 2011 are retrieved from Bloomberg dataset. Among several financial metrics available in this period of time for 100 stocks, the same 11 features in experiment 3 (Table 5) are used in this study.

In the dataset retrieved from Bloomberg, some of these features have missing values for some of the stocks. To deal with them, those stocks are excluded from the initial available set. The result was a narrow down set containing 57 stocks which used as the input set to the optimization process.

The proposed optimization process is applied at each 36 experimental points (for example February 4, 2002). First, available stocks are ranked based on the ranking model built by SVMRank and TS-SVMRank from 11 available features as the predictors and future return as the target ranking. Then, $n$ top stocks with equal weights are added to the portfolio with $n$ equal to 5, 10, or 20. In the ranking learning process at each experimental point one recent past year is considered as the training period. In this process the future return internal ($\Delta t_r$ in Equation 18) and the sampling gap internal are, respectively, set to 6 months and 3 days. The time weighting is based on exponential weighting function (Equation 10) with $a=10$.

This study has two parts; comparing the proposed approach first with All Stocks with equal weights benchmark, and second with Single Criterion portfolios in which a single stock feature is used to rank the available stocks and then, similar to the proposed approach, to select $n$ top stocks from the ranked list.
In the first part of this study, SMPOpt was used to create the benchmark which is a portfolio containing all available stocks (all 57 stocks used in this study). In addition, SMPOpt is used to apply SVMRank and TS_SVMRank to rank the stocks at each point of time. The top stocks in each ranked list are added to the proposed portfolio. Three portfolios are constructed (containing 5, 10, and 20 top stocks) from each of two ranked lists (in total six portfolios). The
Sharpe ratio of 7 portfolios (6 proposed+1 benchmark) at each point is calculated. Figure 33 illustrates the Sharpe ratio of these portfolios in all 36 experimental points.

The hypothesis is that applying the proposed heuristic techniques leads to a managed portfolio and thus is expected to construct portfolios with higher Sharpe ratio. To be able to statistically accept this hypothesis, paired t-test is applied. Table 2 shows the p-values of this test on comparing both SVMRank and TS_SVMRank with the benchmark and comparing TS_SVMRank with SVMRank.

Table 6 - p-Values of t-test hypothesis testing in Exp 4

<table>
<thead>
<tr>
<th></th>
<th>p-Value</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top 5</td>
<td>SVMRank &gt; All Stocks 0.002</td>
<td>TS_SVMRank &gt; SVMRank 0.21</td>
</tr>
<tr>
<td></td>
<td>TS_SVMRank &gt; All Stocks 0.003</td>
<td></td>
</tr>
<tr>
<td>Top 10</td>
<td>SVMRank &gt; All Stocks 0.004</td>
<td>TS_SVMRank &gt; SVMRank 0.08</td>
</tr>
<tr>
<td></td>
<td>TS_SVMRank &gt; All Stocks 0.003</td>
<td></td>
</tr>
<tr>
<td>Top 20</td>
<td>SVMRank &gt; All Stocks 0.022</td>
<td>TS_SVMRank &gt; SVMRank 0.01</td>
</tr>
<tr>
<td></td>
<td>TS_SVMRank &gt; All Stocks 0.002</td>
<td></td>
</tr>
</tbody>
</table>

As the result shows, based on the p-values in the first column of Table 6, the null hypothesis can be strongly rejected. This means both SVMRank and TS_SVMRank on average lead to make portfolios with higher Sharpe ratios. The second column of this table shows the p-value of comparing TS_SVMRank with SVMRank. It can be seen that in the Top 5 selections the hypothesis of improvement in the modified algorithm in comparison to the basic SVMRank algorithm cannot be accepted. However, in the Top 20 selections this hypothesis can strongly be accepted. This shows that the difference between the two algorithms has influence on ranking result but not necessarily the few top ones.

In the second part of this experimental study, the proposed ranking techniques are compared with ranking the available stocks based on a single criterion. For instance, some features (Prof_Margin, Return_on_Asset, Return_On_Cap and ROA_Based_On_Bottom_EPS) are separately used in this study to rank the stocks at all 36 experimental points. The ranked
listed are then used to build portfolios containing \( n \) top stocks (with \( n \) equal to 5, 10, and 20). Figure 34 shows the aggregated result of the Sharpe ratio values in the form of boxplot. As this figure shows, TS_SVMRank and SVMRank have the maximum average Sharpe ratio in all of the experimental runs.

![Figure 34 - Boxplot of Sharpe Ratio of different portfolios in Exp 4](image)

### 6.3.5 Experiment 5

The main objective of this study is to evaluate the proposed stock feature selection technique. For this purpose, the same setup as experiment 3 is used. Figure 35 illustrates the experimental points in experiment 3 (\( tc_1 - tc_{10} \)) as well as new points in this experiment (\( tc_{11} - tc_{13} \)). Past ranking experiences in experiment 3 are used to construct MaxNDCGFeaturesSets and to extract the frequent features (Equation 26). Figure 36 shows the frequency of appearance of each 11 financial features used in this study in MaxNDCGFeaturesSets from \( tc_1 \) to \( tc_{10} \).

The resulting feature set (based on Equation 26) with support of 0.60 is \{\texttt{RETURN\_ON\_CAP}, \texttt{REVENUE\_PER\_SH}, and \texttt{TOT\_DEBT\_TO\_TOT\_ASSET}\}. This set is then used in the last three experimental points (\( tc_{11} \) to \( tc_{13} \)). The ranking process based on SVMRank has been performed in three cases of 1-month, 3-month and 6-month as the future return term,
with two sets of features: first, all 11 original features, and second, 3 features selected by the proposed method.

Figure 35 - Experimental Points in Exp 3 and 5

Figure 36 - Feature appearance frequency

Table 7 shows the NDCG of each run. According to all NDCG values reported in this table the proposed feature selection technique improves the ranking accuracy in terms of NDCG.

Table 7 - NDCG improvement by feature selection

<table>
<thead>
<tr>
<th></th>
<th>tc11</th>
<th>tc22</th>
<th>tc12</th>
</tr>
</thead>
<tbody>
<tr>
<td>11 features</td>
<td>0.9</td>
<td>0.94</td>
<td>0.91</td>
</tr>
<tr>
<td>3 features</td>
<td>0.91</td>
<td>0.97</td>
<td>0.92</td>
</tr>
</tbody>
</table>
6.3.6 Experiment 6

The main objective of this experimental study is to evaluate the combination of social network construction, clustering and SVMRank. In this study the market is confined to available stocks to Standard & Poor’s Composite Index of 500 Stocks (S&P 500), a medium-sized index made up of 500 stocks that are traded on NYSE [5]. The historical price values, market capitalization, as well as other financial metrics of the stocks in S&P 500 for the period from 2000 to 2011 are retrieved from the Bloomberg dataset. Among several financial indicators available for this time period, the same 11 features used in experiment 3 and 4 are used. As the Bloomberg dataset contained missing values for some stocks those stocks are dropped from this study. As a result of culling stocks with missing values, the revised dataset is comprised of 312 stocks which are used as the input set to the optimization process.

Portfolio optimization is simulated at different points in time to generate expected returns beyond each simulation point (or time stamp) – thus building a portfolio based on data available up to that point. This simulation is repeated for 36 time stamps between 2002 and 2010, inclusive (generating new portfolios at 4 points each year at quarterly intervals over the 9 years period). Each optimization run followed the same procedure.

First, a social network of 312 stocks was constructed based on the correlation between their historical return values in the current year. To measure the return value of each company on each “optimization” day, the time interval ($\Delta t$ in Equation 1) was set to 6 months. Second, the constructed network was used to detect 10 clusters by the k-mean clustering algorithm (equal to the number of company sectors in GICS). Third, after identifying 10 clusters, stocks in each cluster are ranked based on the ranking model built by SVMRank from the 11 financial indicators (internal factors) that serve as predictors, and future return as the target ranking. In the
ranking learning process, the future return internal \((\Delta t_r)\) in Equation 18) and the sampling gap internal \((\Delta t_s)\) in Equation 18) were set to 6 months and 1 day, respectively. Fourth, top stocks from each cluster were added to the portfolio, and fifth, stock weights were assigned based on their market capitalization.

Portfolio risk for all optimization runs generated by SMPOpt (36 runs between February 2002 and November 2010) and the S&P 500 benchmark is depicted in Figure 37. Note that portfolio risk for the benchmark is always lower than portfolio risk associated with SMPOpt generated portfolios in keeping with Markowitz’s findings. Since the benchmark contains all the available S&P stocks, it is fully capitalized. A fully capitalized portfolio will always garner the lowest portfolio risk [4].

To compare the proposed portfolios more precisely with the benchmark, their Sharpe ratios over time are measured. Portfolios with high Sharpe ratios are consider the better investment choice. The results presented in Figure 38 shows that the portfolios generated by SMPOpt have higher Sharpe ratios than the fully capitalized S&P benchmark, across all of the experimental runs. The year 2006 is noteworthy as some extreme differences exist between the benchmark and SMPOpt portfolios, while in other years the differences are much less.

Base on the assumption that the primary objective in portfolio optimization is to maximize future returns, the future return of each SMPOpt portfolio is also measured. The future price values (with \(\Delta t\) equal to 6 months) are available for all the experimental points (time stamps). As Figure 39 reports, SMPOpt generated portfolios tend to have higher future returns than the S&P benchmark. A paired t-test reveals that these differences are statistically signification (p-value of 0.047). Consequently the null hypothesis, that most of the time the
portfolio constructed from the proposed optimization approach has higher future return compared to the benchmark of investing in all available stocks, can be rejected.

Figure 37 - Portfolio Risk over time in Exp 6

Figure 38 - Portfolio Sharpe Ratio over time in Exp 6

Figure 39 - Portfolio Future Return over time in Exp 6
6.3.7 Experiment 7[68]

The main objective of this experiment is to evaluate the second strategy in SMPOpt discussed in Chapter 5. For this purpose, publicly available portfolios in StockPickr website [69] are used. StockPickr is a financial services site to incorporate both investment ideas as well as social networking. This community, known as the Stock Idea Network, contains insight from professional investors as well as community members (non-professionals). In this study portfolios of 215 experts and 57 non-professional users in January 2012 have been downloaded. Each portfolio shows the existing stocks and their proportion weights.

For the feature extraction phase, the historical price of the stocks that exist in the portfolios are needed. Daily price of these stocks have been downloaded from Yahoo! Finance for the period of January 2010 to January 2012. 6-month returns of all of the stocks are calculated based on Equation 1 ($\Delta t = 6$ months) for the period of January 2011 to January 2012. The expected return of each of these stocks is then calculated; it is used to calculate the expected return of each portfolio. The risk of the portfolio is also measured based on Equation 4 during the period starting in January 2011.

The extracted risk and return of experts’ portfolios are then used in social network construction. For this purpose the k-means algorithm is invoked with $k=2, 3, \ldots, 9$. Figure 40 shows the constructed social network and distribution of risk value in 5 clusters (the clustering solution with $k=5$). As can be seen in this figure, risk values of all investors are almost uniformly distributed in these 5 clusters, which conforms the risk and return trade-off proposed in the modern portfolio theory. Therefore, these clusters can be labeled based on five point Likert scale of very high, high, medium, low and very low.
These 5 clusters are then used as the training set of the classification step to assign each non-professional investor to one of these 5 communities. SVM classification is applied in this step. In the last step, closeness centralities for all the nodes in each community are measured and the one with the highest value is picked as the representative of that group of people. The portfolio of this node is recommended to the investors assigned to this group. To evaluate whether the recommended portfolio is better than non-professional’s current own portfolio, the performance of those portfolios are compared.

Figure 40 - Social Network of 125 professional investors in StockPickr with the risk value distribution throughout the discovered clustered

Based on the classification process result, 57 non-professional investors used in this study have been assigned to three communities of Very High Risk, Medium Risk and Low Risk. To evaluate whether the recommended portfolio is an appropriate one compared to managed
portfolios from other communities, the Sharpe ratios of representative portfolios in all 5 communities are measured and compared with the Sharpe ratios of the portfolios constructed by non-professionals.

Figure 41 illustrates the aggregated result in the form of boxplot of the Sharpe ratios for different communities. In each community the Sharpe ratio of the representative person, Sharpe ratios of naïve (non-professional) investors assigned to this community, and Sharpe ratio of representative of four other communities are shown. As can be seen, in all three groups, the Sharpe ratio of the representative expert (recommended expert) is higher than the Sharpe ratio of all Naïve investors. This confirms that the portfolio recommended to the non-professional by SMPOpt has higher Sharpe ratio compared to their current portfolio.

![Figure 41 - Boxplot of Sharpe ratio of portfolios in three clusters](image)

However, the Sharpe ratios of representatives from other communities do not necessarily have higher value compared to naïve investors. This shows that there might be some non-professionals with higher performance, in terms of Sharpe ratio, than the professionals and thus not every professional portfolio could be a good replacement for every non-professional
Propensity for risk is an important investment behavioral and preferences that need to be considered for recommending a managed portfolio to a non-professional investor. For instance, naïve investors assigned to Low Risk community, have lower risk compared to representative of other clusters; so as Figure 41 shows, their Sharpe ratio might be higher than the managed portfolios from other clusters (based on Equation 28). Thus those managed portfolios are not necessary better investment options for them compared to their current portfolios.

6.3.8 Experiment 8

The main objective of this experimental study is to evaluate the combination of social network construction based on time series correlation, clustering based on optimum number of clusters calculated from the user's propensity for risk, and TS_SVMRank by using both internal and external factors. In order to simulate real investors with different risk preferences and compare the constructed portfolio suggested to them by SMPOpt with their own built portfolios, the same financial experts downloaded from StockPickr in Experiment 7 are used. Assuming that the level of risk in the current expert's portfolio is consistent with his/her propensity for risk, their publicly available portfolios are used to measure their risk preferences. In this study the market is confined to available stocks in S&P 500. Subset of stocks from S&P 500 which exists in an expert's portfolio is considered as his/her selected subset. Equal weights of these selected stocks are then considered as the portfolio. The historical price values, as well as other financial metrics (internal factors) of the stocks in S&P 500 in January 2012 are retrieved from the Bloomberg dataset.
In this study, the same 11 features used in experiments 3 and 4 are used. As the Bloomberg dataset contained missing values for some stocks, those stocks are dropped from this study. As a result of culling stocks with missing values, the revised dataset is comprised of 465 stocks which are used as the input set to the optimization process. External factors in this study are retrieved from Trading Economics Website [70]. Three external factors in US country are used in this setup. Table 8 reports the brief explanation about these factors.

SMPOpt performs the optimization procedure to build appropriate portfolio for each type of user. First, a social network of 465 stocks was constructed based on the time series correlation (Equation 8) between their historical return values in the last recent year by using Equation 10 as the time weight function (with $a=10$). To measure the return value of each company, the time interval ($\Delta t$ in Equation 1) was set to 6 months. Second, the optimum number of clusters ($k$) is calculated based on Equation 13 with homogeneity threshold as 0.5. In this process $\varepsilon$ starts from 0.1 and grows incrementally to catch at least one clustering solution. The constructed network
was then used to detect $k$ clusters by k-mean clustering algorithm. **Third,** after identifying optimum number of clusters, stocks in each cluster are ranked based on the ranking model built by TS_SVMRank from the 11 internal factors and 3 external factors that serve as predictors, and future return as the target ranking. The time weighting is based on exponential weighting function (Equation 10) with $a=10$. In the ranking learning process, one recent year is used as the training set. The future return interval ($\Delta t_r$ in Equation 18) and the sampling gap interval ($\Delta t_s$ in Equation 18) were set to 6 months and 1 day, respectively. **Forth,** top stocks from each cluster were added to the portfolio, and **fifth,** stock weights were assigned based on equal weights.

The same 5 clusters of experts discovered in experiment 7 (Figure 40) are used in this study to label the risk preference of each expert as *Very Low, Low, Mid, High* and *Very High*. Then SMPOpt applies the optimization process with $r$ set to 0.01, 0.2, 0.5, 0.8, and 0.99, respectively, for investors in each of the above clusters. Optimum number of clusters discovered for each group of investors are, respectively, 19, 21, 25, 27, and 29.

**Figure 42 - Sharpe Ratio of SMPOpt portfolio compared to experts' portfolios**

Figure 42 illustrates the boxplot of Sharpe Ratio of all portfolios built by experts in each investor cluster. Small squares in this figure show the Sharpe Ratio of portfolios built by
SMPOpt for each group. As the result shows, the Sharpe Ratio of SMPOpt portfolio in each group is higher than the Sharpe Ratio of most of the managed portfolios (built by experts).

6.4 Threats to Validity

There are some shortcomings in any experimental study that threaten the validity of its results. In this section the threats of the experiments in this thesis are discussed in terms of Conclusion, Internal, Construct and External validities. Figure 43(a) illustrates experiment principles and the place of each type of these threats [71].

![Diagram](image.png)

**Figure 43 - Threats to Validity of Experimental Studies**

In Figure 43 the experiment objective is shown on the top and the observation in the performed experiment is on the bottom. Conclusion validity is concerned with the statistical relationship between the treatment and the outcome. If this relationship is observed in Internal validity we must make sure that it is a causal relationship not a result of a factor to which we have no control. Furthermore, Construct validity is concerned with the relation between objective and observation. In this validity two things need to be ensured: 1) the treatment reflects the construct of the cause well, and 2) the outcome reflects the construct of the effect well.
Finally the *External* validity is concerned with generalizing the observation by questioning whether the result of the study can be generalized outside the scope of the study [71].

Figure 43(b) illustrated a general objective of this thesis and a sample observation in an experimental setup. There are several threats to validity (discussed next) which can be considered for improvement in future works.

*Conclusion Validity:* As this validity is concerned with the statistical significance to reject the null hypothesis, all of the hypotheses in this thesis tested by pair t-test which leads to p-value less than 0.05 are failed to be rejected. In other words, the probability of rejecting null hypothesis while it is true is less than 0.05. Therefore, there is no conclusion thread about them. However, there is also some hypotheses (for instance, H3 and H4 in Exp 2) which has p-value greater than 0.05, and thus are threatened by conclusion validity.

*Internal Validity:* Although the t-test has a high level of significance, it is not correct to strongly conclude a causal relationship based on limited number of experimental runs with specific settings. A more extensive study with 100 or more runs would provide greater conclusive evidence. Another threat that must be considered is the likelihood of a spurious or confounding relationship embedded in the experimental design. While highly unlikely there is the possibility that illusory correlations are embedded in the results.

*Construct Validity:* Input parameters set in the studies in this thesis are not the only possible values. Variety of settings need to be applied before generalization. In addition, the proposed framework implemented in SMPOpt can be accomplished by other algorithms at each step. For instance, in the ranking step SVMRank has been implemented and evaluated, while other algorithms might affect the results. On the other hand, performance metrics applied such as Sharpe ratio and NDCG are not the only metrics for measuring the performance of the
portfolio or a ranked list of stocks. Other metrics and algorithms need to be used and compared before generalizations.

*External Validity:* Some issues limit the generalization of the results. For instance, one issue is the market representativeness from which stocks are used in the experiment. Other markets with various characteristics should be tested before generalization.
Chapter Seven: **SMPOpt Implementation**

### 7.1 SMPOpt Architecture

The proposed portfolio management system, SMPOpt, has been implemented in Java. In SMPOpt, the user can import financial datasets from different sources such as *Yahoo! Finance*, *Bloomberg*, and *Trading Economics*. The imported datasets are stored in a local database managed by *MySQL* DBMS. Figure 44 illustrates the component diagram of SMPOpt.

![Component Diagram of SMPOpt](image)

**Figure 44 - Component Diagram of SMPOpt**

As Figure 44 illustrates five packages exist in SMPOpt; *UserInterface*, *UserInterface.Dialogs*, *Analysis*, *Data*, and *Data.PreProcessing*. This structure in SMPOpt follows three-tier architecture of *Interface*, *Analysis* and *Data*. Investors communicate with the system through the Interface layer. On the other side, imported data, stored in the local database, is accessed through the Data layer, and the Analysis layer in the middle contains all the core
functionalities in SMPOpt such as network construction, stock clustering, stock ranking, and portfolio construction.

As shown in Figure 44, some components such as Weka [72] and Jung [73] are integrated in SMPOpt. Both of these components are open source projects. Weka contains several data mining techniques such as various clustering algorithms which are used in the SMPOpt process, and Jung is used for graph visualization.

Figure 45 - Entity Relationship Diagram of SMPOpt

Figure 45 shows the Entity Relationship Diagram of SMPOpt. History entity is the core entity connected to three dimensions of Date, Stock and Feature Vector. Each instance in the history table represents a feature vector of a stock in a specific point of historical time. A feature Vector of stocks contains all the internal and external features imported from outsources of data. In addition to these features, there are some system features (for instance, target rank score in the
These features are stored in a separate table called System Feature. List of Markets including their country of origin and available stocks in them are stored as Market entity. Information about investors are captured in the Investor entity.

Different investors can login to SMPOpt, set their preferences and save their settings. User’s profile is stored in a text file with .pfl extension. The local path of this file is stored in the database, and next time when the user logs-in to the system, his/her stored profile is loaded automatically.

Figure 46 illustrates a sample user profile. In this example, selected stocks and features and the weight of each stock in the final constructed portfolio are stored. In addition, various analysis parameters for different steps, such as building the network and clustering, set by the user, are stored in this file.

A sample piece of code in SMPOpt, the CreateTrainingSet method which is to create the training set in the ranking process, is shown in Figure 47. This method retrieves the feature vectors and the target rank score (from the local database) of selected stocks in the specific period of time based on other parameters such as gap interval (available in the user profile). This
information is then used to build the training set in the acceptable format for the SVMRank algorithm.

```java
private boolean CreateTrainingSet()
{
    boolean written = false;
    try{
        Date endDate = new Date(userProfile.rankingSetting.currentDate.getTime() - (long)deltaT * DAY_IN_PS);
        ArrayList<String> requiredFeatures = new ArrayList<String>();
        for(int i=0; i<userProfile.selectedFeatures.size();i++){
            requiredFeatures.add(userProfile.selectedFeatures.get(i));
        }
        ArrayList<String> sysFeatures = new ArrayList<String>();
        userProfile.history.FillSelectedStocks_Features_SampledDates(stocksToRank, userProfile.rankingSetting.startDate, endDate, (String)null);
        int[] stockFlag = new int[stocksToRank.size()];
        while(userProfile.history.selected_stocks_features_SampledDates.next()){
            featureValues = new ArrayList<Double>();
            currentDate = userProfile.history.selected_stocks_features_SampledDates.get("date");
            currentScoreFeature = userProfile.history.selected_stocks_features_SampledDates.getDouble(rankScoreFeatureName);
            if(currentDate == pastDate)
            {
                instance = new TrainingSetInstance(qid++);
                for(int i=0;i<stocksToRank.size();i++)
                    stockFlag[i]=0;
                scores++;
                currentScoreFeature = instance.scoreValue.add(userProfile.history.selected_stocks_features_SampledDates.getDouble(i));
                if(pastScoreFeature != currentScoreFeature)
                    instance.AddStockScore(featureValues);
                completeTrainInstance = true;
                for(int i=0;i<stocksToRank.size();i++)
                {
                    if(stockFlag[i]!=0)
                        completeTrainInstance = false;
                    break;
                }
            }
            else
            {
                completeTrainInstance = true;
                instance.WriteInstanceIntoTrainingFile(currentDate);
                if(written) {
                    written = false;
                    userProfile.errorLogger.AddError("Error in creating training set in Ranking Process: ",e.getMessage());
                }
                return written;
            }
        }
    }catch(Exception e) {
        e.printStackTrace();
        userProfile.errorLogger.AddError("Error in creating training set in Ranking Process: ",e.getMessage());
    }
    return written;
}
```

**Figure 47 - Sample Code in SMPOpt**

### 7.2 Proposed Process in SMPOpt

Figure 48 illustrates the starting window, menus and panels available in SMPOpt. The financial data required for further analysis can be imported to the local database from various data sources such as *Yahoo! Finance* and *Bloomberg* (*File* menu). The available data in the local
The database is then shown to the user in checkbox lists of available stocks and available factors. For instance, two markets of Dow Jones and S&P400 containing their stocks and some factors such as OPER_ROE and PE_RATIO (mnemonics imported from Bloomberg) are shown in Figure 48. The user can select from these lists all items or a subset of items to be used by the system for further analysis. User preferences and settings constitute what is called the user profile which can be saved in a file and loaded later (File menu).

![Figure 48 - Main Screen of SMPOpt](image)

Various analysis functions are available in SMPOpt which are flexible to be executed independently or in sequence in order to perform the proposed optimization process. Options shown in the Analysis menu can be categorized into preprocessing and main steps of the optimization process. The three last options in this menu (Classification, Fill Missing Values, and Feature Construction) are the preprocessing steps. Their goal is to clean the available data and help the user to select appropriate features or create new features.
Functions like **Network Construction**, **Stock Clustering**, **Stock Ranking**, **Stock Selection**, and **Portfolio Construction** (**Analysis** menu) are the main steps of the proposed optimization process. The result of each analysis phase can be displayed to the user (**Visualization** menu) as shown in the **Analysis** panel or saved to the files (**File** menu). Each analysis step needs some
parameters to be set by using a user friendly graphical user interface available in SMPOpt. Figure 49 shows the whole process in SMPOpt as explained in Chapter 4. Figure 49(a), (b) and (c), respectively, show stock and feature selection, parameter setting, and process visualized output.
Chapter Eight: Summary, Conclusion and Future Directions

The last chapter of this thesis contains a brief summary of the whole chapters and the conclusion from them followed by future direction of this project.

8.1 Summary and Conclusion

Stock investment decision making, known as portfolio optimization, is a complicated process involving time-series multi-criteria analysis. Although there are several methodologies proposed by researchers after Modern Portfolio Theory in 1950s, still none of them are completely practical in real world. Initial financial theories were based on unrealistic assumptions leading to several modification proposed by other theorists in the following years. In addition to the financial theorists, computer scientists also have tried to propose heuristic solutions from the area of machine learning and data mining.

Since investment decision is a significant aspect of industry section, several improvement in this area occurred in private industry companies trying to get more benefit from their own investment decisions. Therefore, although there might be several achievements in this area in private industry section, they are not clearly worded to public. Thus, people need to pay the required cost for investment consulting services in case of being available to be purchased.

The objective of this thesis is to provide a heuristic solution for stock portfolio optimization problem based on a hybrid approach of machine learning and data mining techniques as an alternative solution. SMPOpt system, designed and implemented in this thesis, guides the investors in the process of stock investment decision through a semi-automated process based on their investment preferences.

In the proposed solution, the problem is broken down into separate phases each formulated in a way to be solved by a general purpose solution. In addition, based on temporal
characteristic of the financial datasets, these techniques are customized to fit the domain better.

An important contribution of this thesis is formulating the problem in a general way to be able to apply social network analysis, clustering and ranking algorithms. In the proposed structure, various algorithms with different parameters can be applied in each step. Furthermore, customizing some algorithms (such as SVMRank for ranking) to cover temporal behaviour in this domain is another significant contribution of this thesis.

SMPOpt, is designed in a way to accomplish a semi-automated process. Although it is possible to get benefit of a fully-automated decision process by a naive user, other types of users can communicate more with the system along the process to get more personalized result from the SMPOpt.

In this thesis the evaluation step has been done by setting up a series of experimental studies. Each of them focuses to evaluate the effectiveness of specific step(s). These experiments are on real stock market datasets in US country in the past recent decade(s). The treatment and outcome of each of these setups are different based on their objectives leading toward the general objective in this thesis which is using SMPOpt in stock investment decision making process help investors to come up with portfolios with high performance.

8.2 Future Directions

The result of the performed experiments confirm the above main objective. However, as discussed in Chapter 6, there are several threats to validity of these studies. These threats can be the main focus of future direction of this thesis. In other words, more evaluation is certainly required to empirically proof the strength of the SMPOpt. For instance, more markets from around the world, more internal and/or external features, and much more spread out period of
time can be used which have not been applied in the current thesis because of data access limitation.

The performance of portfolio constructed by SMPOpt in the experimental studies of this thesis is compared with the naive strategy of investing on the whole market. Although this strategy seems unpractical in reality, the reason of defining this benchmark is that as discussed before a literature review showed none of the proposed techniques have always overcome the performance of the whole market. Defining other benchmarks containing both financial and AI techniques is considered as future direction of this work. Furthermore, portfolio performance is measured in terms of Sharpe Ration which is a well-known metric for this purpose. However, there are other metrics for this purpose that need to be applied in future.

User study, containing various types of people in term of propensity for risk and financial expertise, is an important evaluation step for SMPOpt. Limitation in this research to find experts people made me to be contented by publicly available portfolios in StockPickr website to simulate expert people in one of the experimental setup and thus left this step as a future work.

Historical financial dataset required for analysis in SMPOpt, imported from outsources, might need more powerful pre-processing steps than the ones available in SMPOpt. For instance, there might be several missing values in the input dataset which cannot be handled by the simple technique explained in Chapter Four. In addition, efficacious financial indicators need to be discovered for further analysis in SMPOpt by a more powerful feature selection technique than the proposed one based on ranking accuracy.

The portfolio optimization problem defined in this thesis is simplified in a single buy decision in which investors are interested to find the optimum portfolio of stocks to buy, without considering their current portfolio. More complicated dynamic portfolio optimization need to be
defined to decide on both sell and buy perspectives with considering the important aspect of stock transaction cost for investors. Furthermore, stock weighting step is not completely focused in this work and thus simple existing metrics are used for this purpose. In future, more complicated weighting techniques need to be combined with the stock selection result of the proposed approach.
References


