Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 35 (2008) 19–29
Expert Systems with Applications
www.elsevier.com/locate/eswa

Mining stock category association and cluster on Taiwan stock market

Shu-Hsien Liao *, Hsu-hui Ho, Hui-wen Lin

Department of Management Sciences and Decision Making, Tamkang University, No. 151, Yingjuan Road, Danshuei Jen, Taipei 251, Taiwan, ROC

Abstract

One of the most important problems in modern finance is finding efficient ways to summarize and visualize the stock market data to give individuals or institutions useful information about the market behavior for investment decisions. The enormous amount of valuable data generated by the stock market has attracted researchers to explore this problem domain using different methodologies. This paper investigates stock market investment issues on Taiwan stock market using a two-stage data mining approach. The first stage Apriori algorithm is a methodology of association rules, which is implemented to mine knowledge and illustrate knowledge patterns and rules in order to propose stock category association and possible stock category investment collections. Then the *K*-means algorithm is a methodology of cluster analysis implemented to explore the stock cluster in order to mine stock category clusters for investment information. By doing so, this paper proposes several possible Taiwan stock market portfolio alternatives under different circumstances.

Keywords: Data mining; Association rule; Cluster analysis; Stock market analysis; Stock portfolio

1. Introduction

The stock market is one of the most popular forms of investment due to its high-expected profit. However, higher expected profit, also imply higher risk. Thus, numerous studies have proposed different analysis methods to assist investors in analysis and decision-making. On the other hand, many individual investors, stockbrokers, and financial analysts attempt to predict stock market price activities and their potential development. This mass behavior runs counter to the counsel of the many academic studies, which contend that the prediction of stock market development is ineffective. This point of view is codified as the generally called efficient markets hypothesis (Fama, 1991; Haugen, 1997).

There are three degrees of market efficiency. The first degree is the strong form of the efficient markets hypothesis, which states that all information that is knowable is immediately factored into the market's price for a security. If this is true, then all of those price predictors are definitely wasting their time, even if they have access to private information. The second degree is the semi-strong form of the efficient markets hypothesis, that all public information is considered to have been possessors of private information, which can use that information for profit. The third degree is the weak form, which holds only that any information gained from examining the security's past trading history is reflected in price. Indeed, the past trading history is public information implying that the weak form is a specialization of the semi-strong form, which itself is a specialization of the strong form of the efficient market hypothesis.

Due to the different degrees of market efficiency, academic researchers investigate the efficient market hypothesis by exploring the unknown and valuable knowledge from historical data, using techniques such as data mining. Enke and Thawornwong (2005) introduces an information gain technique used in machine learning for data mining to evaluate the predictive relationships of numerous financial and economic variables. Neural network models for level estimation and classification are then examined for their ability to provide an effective forecast of future values. Boginski, Butenko, and Pardalos (2006) propose a network representation of the stock market data referred to as the

* Corresponding author.
E-mail address: michael@mail.tku.edu.tw (S.-H. Liao).

0957-4174/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2007.06.001

market graph, which is constructed by calculating cross-correlations between pairs of stocks based on the opening price data over a certain period of time. Chun and Park (2005) proposes a learning technique, which extracts new case vectors using Dynamic Adaptive Ensemble CBR (DAE CBR). The main idea of DAE CBR originates from finding combinations of parameter and updating and applying an optimal CBR model to an application or domain area. These concepts are investigated against the backdrop of a practical application involving the prediction of a stock market index. In addition, Rapach and Wohar (2006) implement an analysis of in-sample and out-of-sample tests of stock return predictability in an effort to better understand the nature of the empirical evidence on return predictability. That study finds that certain financial variables display significant in-sample and out-of-sample predictive ability with respect to stock returns. Overall, most articles consider stock market analysis as a time series problem, and there have been few studies using stock market efficiency to explore the possible cause-and-effect relationships among different stock categories or the influence of outside factors.

This paper investigates stock market investment issues in the Taiwan stock market by implementing a two-stage data mining approach. First, the Apriori algorithm is a methodology of association rules that mines knowledge from historical data and this knowledge is illustrated as knowledge patterns and rules in order to propose stock category association and possible stock investment collections. Next, the *K*-means algorithm is a methodology of cluster analysis that explores the clustering of stock in order to mine this information for investment. Thus, using two different data mining approaches, this paper provides two aspects of data mining results in terms of presenting possible investment portfolio with stock market association and cluster knowledge. The rest of this paper is organized as follows. In Section 2, we describe the Taiwan stock market. Section 3 presents the research design. Section 4 introduces the proposed data mining system, which includes system framework, relational database design, and physical database design. Section 4 presents the data mining approach, including the Apriori and *K*-means algorithm. Section 5 describes the data mining results. Research findings and discussions are presented in Sections 6 and 7 presents a brief conclusion.

2. Taiwan stock market

TSEC, Taiwan Stock Exchange Corporation, maintains stock price indices to allow investors to grasp both overall market movement and different industrial sectors' performances conveniently. The indices may be grouped into market value indices and price average indices. The former are similar to the Standard and Poor's Index, weighted by the number of outstanding shares, and the latter are similar to the Dow Jones Industrial Average and the Nikkei Stock Average. The Taiwan Stock Exchange Capitalization Weighted Stock Index ("TAIEX") is the most widely quoted of all TSEC indices. The base year value as of 1966 was set at 100. TAIEX is adjusted in the event of new listings, de-listings and new share offerings to offset the influence on TAIEX owing to non-trading activities. TAIEX covers all of the listed stocks excluding preferred stocks, full-delivery stocks and newly listed stocks that have listed for less than one calendar month. The other market value indices are calculated and adjusted similarly to that of the TAIEX, but with different groupings of stocks included for calculation. Out of the TAIEX Component Stocks, the non-Finance Sub-Index, Non-Electronics Sub-Index, and Non-Finance Non-Electronics Sub-Index include stocks not in the financial sector, not in the electronics sector, and not in either sector. Similarly, the Industrial Sub-Indices are calculated for different industrial sectors. In 1986, eight Industrial Sub-Indices were introduced, i.e. Cement/Glass/Ceramics, Textiles, Foods, Plastics/Chemicals/Rubber, Electric Machinery/Electric Appliance/Cable/Electronics, Paper/Pulp, Construction, Finance. In 1995, the TSEC introduced additional 14 Industrial Sub-Indices, i.e. Cement, Plastics, Electric machinery, Electric appliance/cable, Automobile, Chemicals, Glass/ceramics, Iron/steel, Rubber, Electronics, Transportation, Tourism, Retail and others. This expansion was to give a broader perspective of industrial performance and a more comprehensive comparison with overall market trends. Total Return Indices add back cash dividends to the index calculations, and are published at the end of each trading day. This expansion can serve as a better indicator to measure the performance of funds.

The Industrial Price Average Index and the Composite Price Average Index contain 20 and 30 issues, respectively. The samples are chosen based on their representation in the market as a whole and are adjusted every year by taking considering the profitability, operational efficiency and trading liquidity of the shares, so that the indices can mirror the market trend. All of the TSEC indices (excluding Total Return Indices) are constantly computed and broadcast every minute during the trading hours through the TSEC MIS system and information vendors' networks. This information can be easily accessed on the systems of local and international information vendors, such as Reuters, Bridge, Quick, Bloomberg, Primark, etc. Monthly summaries of all the TSEC indices data are also available on TSEC website (http://www.tse.com.tw/en/).

3. Research design

3.1. Research procedure

In this article, we use the TAIEX indices, including 19 index categories and international stock indices such as the NIKKEI 225, KOSPI, Dow Jones, etc. and construct a database. This database provides the basis for data mining. Moreover, the database can be mined out of portfolio investment suggestions by using of association rule and