Expert Systems with Applications 37 (2010) 8793–8798

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Clustering Indian stock market data for portfolio management

S.R. Nanda, B. Mahanty, M.K. Tiwari *

Department of Industrial Engineering and Management, Indian Institute of Technology, Kharagpur, West Bengal, India

ARTICLE INFO

Keywords:
Portfolio management
Markowitz model
K-means clustering
Self organizing maps
Fuzzy *C*-means

ABSTRACT

In this paper a data mining approach for classification of stocks into clusters is presented. After classification, the stocks could be selected from these groups for building a portfolio. It meets the criterion of minimizing the risk by diversification of a portfolio. The clustering approach categorizes stocks on certain investment criteria. We have used stock returns at different times along with their valuation ratios from the stocks of Bombay Stock Exchange for the fiscal year 2007–2008. Results of our analysis show that *K*-means cluster analysis builds the most compact clusters as compared to SOM and Fuzzy *C*-means for stock classification data. We then select stocks from the clusters to build a portfolio, minimizing portfolio risk and compare the returns with that of the benchmark index, i.e. Sensex.

1. Introduction

One of the decision problems in the financial domain is portfolio management and asset selection. Under the extremely competitive business environment, in order to face the complex market competitions, financial institutions try their best to make an ultimate policy for portfolio selection to optimize the investor returns. A formal model for creating an efficient portfolio was developed by Markowitz (1952). In his model the return of an asset is its mean return and the risk of an asset is the standard deviation of the asset returns. Risk was quantified such that investors could analyze risk-return choices. Moreover, risk quantification enabled investors to measure risk reduction generated by diversification of investment. So diversification of investment is essential to create an efficient portfolio. The problem of selecting well diversified stocks can be tackled by clustering of stock data.

Clustering as defined by Mirkin (1996) is "a mathematical technique designed for revealing classification structures in the data collected in the real world phenomena". Clustering methods organize a data set into clusters such that data points belonging to one cluster are similar and data points belonging to different clusters are dissimilar. In this paper we demonstrate the implementation of stock data clustering using well known clustering techniques namely *K*-means, self organizing maps (SOM) and Fuzzy *C*-means. The stock market data is clustered by each of the above methods. The optimal number of clusters for the stock market data using each clustering technique is carried out. The stock data contains attributes as a series of its timely returns as well as the valuation ratios to present a clear position of their market value. These are the direct investment criteria that are being considered for stock selection. Thus the resulting clusters are a classification of high dimensional stock data into different groups in view of the difference between return series along with current market valuation of stocks. After clustering stock samples are selected from these clusters to create efficient portfolio. The process is simulated for certain iterations and average risk and return is found out. It is easy to get the portfolios with lowest risk for a given level of return, using certain optimization model which is demonstrated at the end of the paper.

In order to create efficient portfolios with Markowitz model, we use the clustering method to select stocks in the paper, called clustering-based selection in our paper. The remainder of the paper is organized as follows.

The remainder of this paper is organized as follows. Section 2 describes relevant literature review. Section 3 presents the clustering-based stock selection method. Section 4 shows problem description. Section 5 depicts experimental results. In Section 6, the conclusion is presented.

2. Literature review

In this section, portfolio management and clustering techniques are briefly reviewed.

2.1. Portfolio management

A model of creating efficient portfolio was developed by Markowitz (1952). In the Markowitz model, the return of a stock is the mean return and the risk of a stock is the standard deviation of the stock returns. The portfolio return is the weighted returns of stocks. The efficient frontier of portfolios is the set of portfolios that

* Corresponding author. Tel.: +91 3222 283746.
E-mail address: mkt09@hotmail.com (M.K. Tiwari).

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.06.026

offer the greatest return for each level risk (or equivalently, portfolios with the lowest risk for a given level of return). Investors measure risk reduction by diversification of investment. A lot of work has been done on portfolio management henceforth. Topaloglou, Vladimirou, and Zenios (2008) worked on a dynamic stochastic programming model for international portfolio management, a solution that determines capital allocations to international markets, the selection of assets within each market, and appropriate currency hedging levels. Genetic algorithms have been used for portfolio optimization for index fund management by Oh, Kim, and Min (2005). Fernandez (2005) states a stochastic control model that includes ecological and economic uncertainty for jointly managing both types of natural resources. Fuzzy models (Östermark, 1996) for dynamic portfolio management have also been implemented.

From the literatures reviewed we could see that there are very few studies on clustering stock data but there have been a lot of work for portfolio optimization. The initial cluster indexing of stock data can be helpful for optimization models thus improving their efficiency. Therefore, in this study we would like to focus on the stock data clustering and develop a model for diversified stock selection.

2.2. Clustering techniques

Various clustering techniques have been used in problems from various research areas of math, multimedia, biology, finance and other application domains. There are various studies within the literature that used different clustering methods for a given classification problem and compared their results. For instance Chiu, Chen, Kuo, and He (2009) applied *K*-means for intelligent market segmentation. Many variants of the normal *K*-means algorithm have also been used in various fields. Kim and Ahn (2008) used a GA version of *K*-means clustering in building a recommender system in an online shopping market. Kuo, Wang, Hu, and Chou (2005) developed a variant of *K*-means which modifies it as locating the objects in a cluster with a probability, which is updated by the pheromone, while the rule of updating pheromone is according to total within cluster variance (TWCV). Fuzzy *C*-means was a development by Bezdek (1981). The Fuzzy *C*-means clustering algorithm is a variation of the *K*-means clustering algorithm, in which a degree of membership of clusters is incorporated for each data point. The centroids of the clusters are computed based on the degree of memberships as well as data points. Over the time Fuzzy *C*-means has found an increasing use in data clustering. Ozkan, Türkşen, and Canpolat (2008) published a paper on analyzing currency crisis using Fuzzy *C*-means. A Fuzzy system modelling with Fuzzy *C*-means (FCM) clustering to develop perception based decision matrix is employed here. Tari, Baral, and Kim (2009) proposes a variant GO Fuzzy *C*-means which is a semi supervised clustering algorithm and it utilizes the Gene Ontology annotations as prior knowledge to guide the process of grouping functionally related genes.

Other popular clustering techniques use artificial neural networks for data clustering and one of the most popular is self organizing maps (SOM). Some of the recent works include use of self organizing maps in detection and recognition of road signs (Prieto & Allen, 2009), for clustering of text documents (Isa, Kallimani, & Lee, 2009), for classification of sediment quality (Alvarez-Guerra, González-Piñuela, Andrés, Galán, & Viguri, 2008) and many more.

There are papers showing comparison of different clustering methods (Budayan, Dikmen, & Birgonul, 2009; Delibasis, Mouravliansky, Matsopoulos, Nikita, & Marsh, 1999; Mingoti & Lima, 2006) and also adapting different clustering methods for a particular problem. In case based reasoning (CBR) (Chang & Lai, 2005; Jo & Han, 1996; Kim & Ahn, 2008) the problem of cluster indexing the case base to build a hybrid CBR has adapted many clustering methods.

In this paper we consider the *K*-means, Fuzzy *C*-means and self organizing maps for clustering stock data. We will use validity indexes in each case to find the optimal number of clusters.

3. Methodology

Through our literature survey we found that the problem of efficient frontier can be solved more efficiently by clustering the stocks and then choosing to enhance the criteria of diversification. We propose clustering of high dimensional stock data by the popular clustering methods *K*-means, SOM and Fuzzy *C*-means and then selecting stocks to build an efficient portfolio.

All the clustering methods are used to cluster financial stock data from Bombay Stock Exchange that consists of returns for variable period lengths along with the valuation ratios. Through the step of clustering, the aim of least diversity within a group and most difference among groups is be reached. The optimal number of clusters for each method is to be found out using certain internal validity indexes. The framework of our problem is shown in Fig. 1. A brief explanation of Markowitz model along with the clustering techniques is given below.

3.1. Markowitz model

As stated before Markowitz's has enabled investors to measure risk reduction generated by diversification of investment. We can say the return of a portfolio is the weighted return of the underlying stocks. If σ_p is the portfolio risk and n be the number of underlying stocks then

$$\sigma_p^2 = \sum_{i=1}^{n}\sum_{i=1}^{n} w_i w_j \sigma_{ij} \tag{1}$$

where σ_{ij} is the covariance between the stock price of i and j. w_i and w_j are the weights assigned to stock i and j. If return is fixed then the problem of minimization of risk can be stated as:

$$\min \quad \sigma_p^2 = w^T S w \tag{2}$$

$$\text{subject to} \quad w^T I = 1 \tag{3}$$

$$w^T R = R_E \tag{4}$$

where w is the weight vector which is a value between 0 and 1. S is the variance covariance matrix of the stocks and R_E is the expected return and R is the mean return of each stock defined as $R_t = \log\left(\frac{S_t}{S_{t-1}}\right)$. S_t is the price of the stock at time 't'.

3.2. K-means

K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. *K*-means starts with a single cluster with its center as the mean of the data. This cluster is split to two and the means of the new clusters are trained iteratively. These clusters again split and the process continues until the specified number of clusters is obtained. If the specified number of clusters is not a power of two, then the nearest power of two above the number specified is chosen. Then the least important clusters are removed and the remaining clusters are again iteratively trained to get the final clusters. This is a non hierarchical method.

3.3. Fuzzy C-means

In Fuzzy clustering methods data points can be assigned to more than one cluster with different degree of membership. In