Expert Systems with Applications 41 (2014) 1301–1314

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Stock market co-movement assessment using a three-phase clustering method

Saeed Aghabozorgi *, Ying Wah Teh

Department of Information System, Faculty of Computer Science & Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia

ARTICLE INFO

Keywords:
Stock market
Co-movement
Data mining
Clustering
Time-series

ABSTRACT

An automatic stock market categorization system would be invaluable to individual investors and financial experts, providing them with the opportunity to predict the stock price changes of a company with respect to other companies. In recent years, clustering all companies in the stock markets based on their similarities in the shape of the stock market has increasingly become a common scheme. However, existing approaches are impractical because the stock price data are high-dimensional data and the changes in the stock price usually occur with shift, which makes the categorization more complex. Moreover, no stock market categorization method that can cluster companies down to the sub-cluster level, which are very meaningful to end users, has been developed. Therefore, in this paper, a novel three-phase clustering model is proposed to categorize companies based on the similarity in the shape of their stock markets. First, low-resolution time series data are used to approximately categorize companies. Then, in the second phase, pre-clustered companies are split into some pure sub-clusters. Finally, sub-clusters are merged in the third phase. The accuracy of the proposed method is evaluated using various published data sets in different domains. We show that this approach has good performance in efficiency and effectiveness compared to existing conventional clustering algorithms.

1. Introduction

There are many works related to stock market analysis including clustering (Durante and Foscolo, 2013; Nanda et al., 2010) and prediction (Qiu et al., 2012; Svalina et al., 2013; Zarandia et al., 2009). Clustering is a data mining technique in which similar data are automatically placed into related groups without advanced knowledge of the group definitions. Clustering of companies in the stock market is very useful for managers, investors, and policy makers. It can be performed based on several factors, such as the size of the companies, their annual profit, and the industry category. For example, Nanda et al. (2010) used the returns of the stock for variable time intervals along with the validation ratios to cluster the companies listed in Bombay Stock Exchange (BSE). However, these features usually change over time; thus, they are improper for categorization purposes. Industry-based categorization is also not preferable due to evidence that financial analysts are biased by industry categorization (Krüger et al., 2012). Conversely, the closing prices of stocks related to each company are stored as time series data. Companies can be categorized by the clustering of their stock price time series. Clustering companies based on the time series of their stock price is particularly advantageous in co-movement assessment. Identifying homogeneous groups of stocks where the movement in one market affects stock prices in another market is called co-movement. The literature shows that the movement of a stock market in a country is affected by the movement of other stocks in that country or in other regions (Antoniou, 2003; Collins and Biekpe, 2003; Masih and Masih, 2001). As a result, numerous studies have been performed on the recognition of co-movements between different countries (Graham and Nikkinen, 2011; Norden and Weber, 2009; Rua and Nunes, 2009). For example, examining the co-movement of the stock markets of Taiwan and Hong Kong (Liao and Chou, 2013) or co-movement of Asia–Pacific with European and US stock market returns (Loh, 2013). However, most of these studies consider the co-movement of the stock market between different regions or countries but not among different industries or companies in a stock market, such as the Malaysian stock market. Assessment of the stock market co-movement between companies in a stock market can be very helpful for predicting the stock price, based on the similarity of a company to other companies in the same cluster. That is, given two stock market time series, *A* and *B*, we can show that whenever the price of *A* drops, the price of *B* will also drop, and vice versa. Therefore, clustering of time series related to the stock price of companies can answer the following questions:

* Corresponding author. Tel.: +60 (3) 79676431.
E-mail addresses: saeed@um.edu.my (S. Aghabozorgi), tehyw@um.edu.my (Y.W. Teh).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.08.028

1. Can a company category be determined given a historical record of its stock price?
2. How are the movements of a specific company in stock markets across the various companies?
3. How can the stock price be predicted based on its similarity to other co-movement companies?

Researchers have shown that clustering using the best-known conventional techniques, such as partitional and hierarchical algorithms, generate clusters with acceptable structural quality and consistency, and are partially efficient in terms of execution time and accuracy for static data (Jain et al., 1999). However, classic machine learning and data mining algorithms do not work well for time series due to their unique structure (Lin et al., 2004). In effect, the high dimensionality, very high feature correlation, and (typically) the large amount of noise that characterize time series data present difficult challenges for time series clustering (Keogh and Kasetty, 2003; Lin et al., 2004). Accordingly, massive research efforts have been made to present an efficient approach for time series clustering. However, focusing on the scalability of these methods to deal with time series data has come at the expense of losing the usability and effectiveness of clustering (Ratanamahatana, 2005). For example, Euclidean distance (ED) is adopted as a distance metric on most of the existing works due to its high efficiency, whereas it is not sufficiently accurate because it is only suitable for calculating the *similarity in time*(i.e., similarity in each time step) (Bagnall and Janacek, 2005; Ratanamahatana et al., 2005). Take the stock market data of several companies as an example. Fig. 1 shows the hierarchy clustering of several companies based on the similarity in time. Focusing on the inter-temporal co-movement of the daily closing prices in one month for two stocks, company B and company C may seem very different. In time series clustering analysis, these two stocks may be surmised to belong in two different clusters or that no linkage exists between them. However, a precise look at the stock price of these two stocks shows that they are co-moving but with some shifts or daily lead-lag relationships. Hence, the clustering results of these two stocks based on the similarity in time may fail to provide insightful information.

In this paper, a new **3**-**P**hase**Time series** **C**lustering model (3PTC) is proposed, which can construct the clusters based on similarity in shape. This model facilitates the accurate clustering of time series data sets and is designed specifically for very large time series data sets. In the first phase of the model, data are pre-processed, transformed into a low dimensional space, and grouped approximately. Then, the pre-clustered time series are refined in the second phase using an accurate clustering method, and are represented by some prototypes. Finally, in the third phase, the prototypes are merged to construct the ultimate clusters. To evaluate the accuracy of the proposed model, the 3PTC is tested extensively using published time series data sets from diverse domains.

This paper contributes to the existing literature by proposing a new model of time series clustering which (1) is more accurate than conventional approaches, (2) is scalable (on large datasets) due to the use of multi-resolution time series in different levels of clustering, and (3) can overcome the limitations of comparative clustering algorithms in finding the clusters of similar time series in shape. This salient feature is very advantageous for co-movement assessment in stock markets.

The rest of this paper is organized as follows. In Section 3, the related works are described. The proposed model is explained in Section 4. In Section 5, the algorithm is applied on diverse time series data sets, and the experimental results are reported. In Section 6, conclusions and future perspectives are drawn.

2. Literature review

A current research issue in the finance literature is co-movements of the world's national financial market indexes (Antonakakis et al., 2013; Chen and Wu, 2013; Chow et al., 2011; Graham et al., 2012; Liao et al., 2011; Madaleno and Pinho, 2012; Wahal and Yavuz, 2013). In recent years, many researchers have investigated the co-movement in stock markets by different data mining approaches. For instance, Liao and Chou (2013) employed association rules and clustering algorithm to investigate the co-movement in the Taiwan and China (Hong Kong) stock markets. They categorized the stock indexes into thirty clusters to perceive the behavior of stock index associations. In another study (Dutt and Mihov, 2013), the authors employed monthly stock indices to construct pairwise correlations of returns. They explained these correlations with risk-adjusted differences in industrial structure across 58 countries. The result of their study indicates that countries with similar industries exhibit higher stock market co-movements. Graham and Kiviaho (2013) investigated the short term and long term co-movement of MENA (Middle East and North Africa) region stock markets with the U.S. stock market and the regional co-movement among these markets. In this work, the wavelet squared coherency method was employed to examine the co-movement of stock markets. Likewise, using the wavelet coherence approach, Loh (2013) investigated the co-movement of 13 Asia–Pacific stock market returns with stock market of European and US stock market returns. Durante and Foscolo (2013) proposed an index to measure the contagion effects between a group of markets. This index was used by the authors to derive a

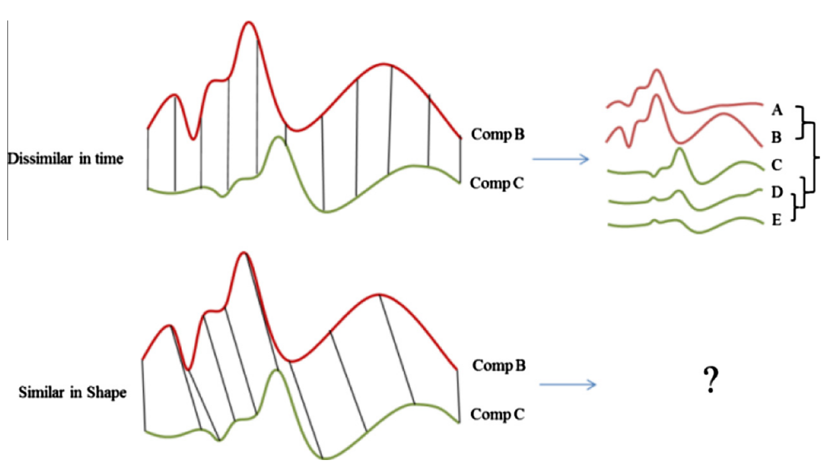

Fig. 1. separated clusters because of calculating the distance between time series based on similarity in time.