Expert Systems With Applications 86 (2017) 1–17

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Planning stock portfolios by means of weighted frequent itemsets

Elena Baralis, Luca Cagliero*, Paolo Garza

Dipartimento di Automatica e Informatica, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129, Torino, Italy

ARTICLE INFO

Article history:
Received 2 September 2016
Revised 19 May 2017
Accepted 20 May 2017
Available online 20 May 2017

Keywords:
Frequent itemset mining
Stock data analysis

ABSTRACT

Planning stock portfolios is a challenging task, because investors have to forecast stock market trends. To limit losses due to wrong forecasts a common strategy is diversification, which consists in buying stocks belonging to different sectors/markets to spread bets across different assets. Since the amount of stock market data is continuously growing, an appealing research strategy is to first apply data mining algorithms to discover significant patterns from potentially large stock datasets and then exploit them to support investor decision-making.

This article presents an itemset-based approach to supporting buy-and-hold investors in technical analyses by automatically identifying promising sets of high-yield yet diversified stocks to buy. Specifically, it investigates the use of itemsets to generate stock portfolios from historical stock data and recommend them for buy-and-hold investments. To achieve this goal, it analyzes stock market datasets, which contain for each stock the closing prices on different trading days. Datasets are enriched with (analyst-provided) taxonomies, which are used to classify stocks as the corresponding sectors. Unlike previous approaches, it generates a model composed of a subset of potentially interesting itemsets, which are then used to support investors in decision-making. The selected itemsets represent promptly usable stock portfolios satisfying expert's requirements on minimal average return and minimal level of diversification across sectors.

The experiments performed on real stock datasets acquired under different market conditions demonstrate the effectiveness of the proposed approach compared to real stock funds.

1. Introduction

Financial data mining entails the application of data mining techniques to analyze very large financial datasets (Han & Kamber, 2006). It focuses on (i) understanding and managing financial risks (Lin, Lee, Kao, & Chen, 2008; Strong, Steiger, & Wilson, 2009), (ii) supporting intra-day trading (Schreck, Tekušová, Kohlhammer, & Fellner, 2007; Tsai & Quan, 2014), and (iii) performing credit rating, loan management, bank customer profiling, and money laundering (Lu, Tsai, Chen, Hung, & Li, 2012; Stubblebine, Syverson, & Goldschlag, 1999). A challenging financial task is stock portfolio planning and management based on buy-and-hold strategies. Buy-and-hold is an investment strategy that focuses on buying market stocks at a given price and holding them for a relatively long time, because their prices are likely to increase (Chen, Kao, Lyuu, & Wong, 1999). Investors' decisions are commonly driven by either fundamental or technical analyses (Williams & Turton, 2014). Fundamental analyses entail the study of the overall state of a company or a business (e.g., production, earnings, employment, housing, manufacturing, management), while technical analyses rely on the study of stock prices, which are assumed to reflect all external influences, with the help of ad hoc statistics-based indicators (e.g. moving averages). When markets are open stock prices continuously vary over time. Hence, to maximize returns investors need automated tools to support the analysis of large market stock datasets. Therefore, an appealing research strategy is to first apply data mining algorithms to discover significant patterns from potentially large stock datasets and then exploit them to support investor decision-making.

Diversification is one of the most established methodologies for limiting losses in case forecasts turn out to be wrong (Markowitz, 1991). It consists in buying anti-correlated or independent stocks to spread bets across a wide range of assets. For example, a common strategy to mitigate losses due to underperforming market sectors is to buy stocks belonging to various sectors (Meric, Ratner, & Meric, 2008; Ratner & Leal, 2005). Some attempts to adopt data mining techniques to support investor decision-making have been made. For example, supervised approaches relying on classi-

* Corresponding author.
E-mail addresses: elena.baralis@polito.it (E. Baralis), luca.cagliero@polito.it, luca.cagliero84@gmail.com (L. Cagliero), paolo.garza@polito.it (P. Garza).

http://dx.doi.org/10.1016/j.eswa.2017.05.051
0957-4174/

fication or regression analyses (e.g., Leung, Daouk, & Chen, 2000; Tilakaratne, Mammadov, & Morris, 2007; Zhang, Jiang, & Li, 2004) have already been proposed to forecast financial stock market signals (e.g., buy, hold, sell). In parallel, unsupervised approaches based on clustering or time series analyses (e.g., Aguiar & Sales, 2010; Chung, Fu, Luk, & Ng, 2002; Guo, Jia, & Zhang, 2008; Xiu, Hong, & Zhen, 2009) addressed stock data characterization and trend detection. Some research efforts (e.g., Chen, Mabu, Hirasawa, & Hu, 2007; Kantardzic, Sadeghian, & Shen, 2004; Parque, Mabu, & Hirasawa, 2009) have also been devoted to identifying the subset of stocks that maximizes returns while minimizing risks due to under-diversification. The position of this work with respect to the state-of-the-art is thoroughly discussed in Section 2.

This article addresses the problem of supporting the planning and management of diversified stock portfolios in a buy-and-hold scenario. Specifically, it presents an itemset-based approach to supporting investors in planning diversified stock portfolios. We investigate the use of frequent itemsets to automatically generate stock portfolios from historical data and then recommend them for buy-and-hold investments. Frequent itemset mining is an exploratory unsupervised data mining technique that has largely been used to discover significant correlations between items occurring in large datasets (Agrawal, Imieliński, & Swami, 1993). Itemsets are extracted from weighted stock datasets, which contain the closing prices of the stocks on different days. In our context, itemsets are sets of stocks whose corresponding daily prices are strongly correlated with each other.[1] Even though the proposed approach is general and can be applied to any price (e.g. opening, closing, daily maximum) or technical indicator, hereafter we will target our analyses on the stock closing price, which is commonly used as reference market factor (Zhang, Jiang, & Li, 2004). More specifically, a weighted stock dataset consists of a (potentially large) set of rows, each one corresponding to a different timestamp. Each row contains a set of pairs (*stock, return*), called items, where *stock* is a stock identifier, while *return* is the relative return of the corresponding stock at the given timestamp, which is computed as the percentage difference between the closing price of the stock at the current timestamp and a reference price. To differentiate investments across different sectors/markets stock datasets are enriched with analyst-provided taxonomies (i.e., is-a hierarchies), which classify stocks as the corresponding categories. To generate taxonomies, stocks can be semi-automatically clustered into well-known categories (e.g., the market sectors) or, alternatively, they can be grouped by exploiting more advanced clustering strategies targeted to portfolio diversification (e.g., Aguiar & Sales, 2010; Gandhmal, Parihar, & Argiddi, 2011; Rostoker, Wagner, & Hoos, 2007; Xiu, Hong, & Zhen, 2009). Sector-based categorization has already been exploited to diversify stock investments (e.g., Meric, Ratner, & Meric, 2008; Ratner & Leal, 2005).

Example. Let us consider an investor who would like to plan stock investments in 2014 by analyzing the daily closing prices of the U.S. NASDAQ index stocks in 2013. The analyzed dataset contains as many rows as the number of trading days in 2013. To compute for each stock the daily relative returns in 2013, let us consider as reference price the corresponding closing price on the last trading day of 2012. Table 1 reports the daily relative returns achieved by five representative market stocks (*a, b, c, d, e*) over a time period of six days (identified by the corresponding timestamps $t_1 - t_6$). Stocks in Table 1 are clustered into sectors according to the taxonomy reported in Fig. 1. For example, stocks *a* and *b* are classified as *Healthcare*.

[1] We study the correlation between the closing prices of multiple stocks in a given time period rather than analyzing the temporal sequences of stock prices.

Table 1
Example of weighted stock dataset.

Time stamp	Weighted stock transaction
t_1	⟨*a*, 5%⟩ ⟨*b*, 5%⟩ ⟨*c*, −1%⟩ ⟨*d*, 7%⟩ ⟨*e*, 5%⟩
t_2	⟨*a*, 2%⟩ ⟨*b*, 6%⟩ ⟨*c*, 0%⟩ ⟨*d*, 2%⟩ ⟨*e*, 2%⟩
t_3	⟨*a*, 4%⟩ ⟨*b*, 5%⟩ ⟨*c*, −2%⟩ ⟨*d*, 4%⟩ ⟨*e*, 5%⟩
t_4	⟨*a*, 4%⟩ ⟨*b*, 2.5%⟩ ⟨*c*, −4%⟩ ⟨*d*, 10%⟩ ⟨*e*, 4%⟩
t_5	⟨*a*, 1%⟩ ⟨*b*, 4%⟩ ⟨*c*, −2%⟩ ⟨*d*, 7%⟩ ⟨*e*, 1%⟩
t_6	⟨*a*, −1%⟩ ⟨*b*, 6%⟩ ⟨*c*, 0⟩ ⟨*d*, 1%⟩ ⟨*e*, −1%⟩

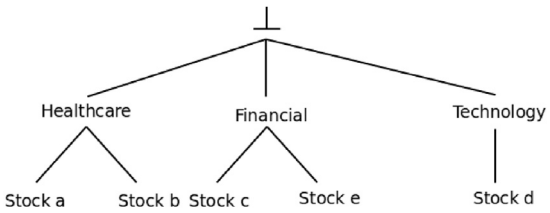

Fig. 1. Example of taxonomy built over the weighted stock dataset.

A model composed of a subset of potentially interesting itemsets is generated and exploited for supporting investor decision-making. Itemsets are extracted by using a variant of a frequent weighted itemset mining algorithm (Cagliero & Garza, 2014), which has been modified to generate only the subset of itemsets of interest. More specifically, in our context itemsets represent combinations of correlated stocks of arbitrary size. Thus each itemset represents a candidate stock portfolio. As discussed in Section 2, to the best of our knowledge this work is the first attempt to use frequent itemsets to generate stock portfolios. To generate profitable yet diversified portfolios experts are asked to set (i) a minimum average return *minret* gained by all the stocks in the portfolio (e.g. *minret*=2% means that the average, computed on historical data, of the least per-day relative returns of the corresponding stocks is above 2%). (ii) a minimum level of diversification *mindiv*, which is expressed by the percentage of stocks belonging to different sectors (e.g., *mindiv*=80% means that at least 8 out of 10 stocks in the portfolio belong to different sectors). Constraint (i) ensures that the average daily profit, on the considered time frame, of the least performing stock in the portfolio is above a given threshold. Hence, investors can pay off the investment with an acceptable return if they need promptly available funds in any time. Constraint (ii) ensures that portfolio stocks are spread across a large enough number of market sectors thus guaranteeing diversification over different assets. Itemsets satisfying both the aforesaid constraints represent valuable hints for buy-and-hold investors because they protect investors against negative outcomes of single sectors while guaranteeing a minimal per-stock profit according to historical data.

To support expert decisions the mined itemsets are ranked in order of decreasing length and profit. Itemsets with maximal length are placed first because their corresponding portfolios maximize investment diversification over stocks, i.e., they spread bets on a larger number of assets. Note that the best portfolio size (i.e., the number of recommended stocks to buy) is automatically defined by the system according to the itemset-based model. On equal terms, itemsets with maximal average profit are preferred because their corresponding stocks yield maximal gain on historical data. The top ranked itemset is deemed as the most interesting hint for buy-and-hold (long-term) investors.

Table 1 reports the itemsets mined by enforcing a minimum relative return threshold equal to 3% and minimum diversification level equal to 60%, ranked by decreasing length and average profit. Based on itemset raking a portfolio consisting of stocks *a, d*, and *e* is recommended to domain experts because it is the largest combination of stocks that satisfies both minimum quality constraints