Expert Systems with Applications 42 (2015) 1325–1339

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Developing an approach to evaluate stocks by forecasting effective features with data mining methods

Sasan Barak [a],*, Mohammad Modarres [b]

[a] Young Researchers and Elite Club, Ardabil branch, Islamic Azad University, Ardabil, Iran
[b] Department of Industrial Engineering, Sharif University of Technology, Tehran, Iran

ARTICLE INFO

Article history:
Available online 23 September 2014

Keywords:
Stock market
Data mining
Classification algorithm
Feature selection
Function-based clustering method

ABSTRACT

In this research, a novel approach is developed to predict stocks return and risks. In this three stage method, through a comprehensive investigation all possible features which can be effective on stocks risk and return are identified. Then, in the next stage risk and return are predicted by applying data mining techniques for the given features. Finally, we develop a hybrid algorithm, on the basis of filter and function-based clustering; the important features in risk and return prediction are selected then risk and return re-predicted. The results show that the proposed hybrid model is a proper tool for effective feature selection and these features are good indicators for the prediction of risk and return. To illustrate the approach as well as to train data and test, we apply it to Tehran Stock Exchange (TSE) data from 2002 to 2011.

1. Introduction

Of the most important concerns of market practitioners is future information of the companies which offer stocks. A reliable prediction of the company's financial status provides a situation for the investor to more confident investments and gaining more profits (Huang, 2012). One can refer to different studies about share gaining and return prediction, for example, time series stock price prediction model (Araújo & Ferreira, 2013), buy–hold–sell prediction model (Wu, Yu, & Chang, 2014; Zhang, Hu, Xie, Zhang, et al., 2014), Index prediction model with Anfis (Svalina, Galzina, Lujić, & Šimunović, 2013) or MARS and SVR (Kao, Chiu, Lu, & Chang, 2013), profit gaining (Ng, Liang, Li, Yeung, & Chan, 2014). However, unlike the return, risk has been rarely considered for prediction, while customers usually balance their return for a proper level of risk, then clearly both risk and return are important factors in financial decision making (Barak, Abessi, & Modarres, 2013; Tsai, Lin, Yen, & Chen, 2011). Without risk evaluation the portfolio efficient frontier does not make sense. Thus, this paper implements the forecasting of both risk and return of stocks which has tremendous effect on price setting. Also, up-down prediction of stock movement such as (Patel, Shah, Thakkar, & Kotecha, 2014; Yu, Chen, & Zhang, 2014; Zhang, Hu, Xie, Wang, et al., 2014) cannot result in precision view of stock future and investors gaining. While classifying the amount of risk and return to different categories like our method gives more specific and clear knowledge.

Therefore, in this study, the simultaneous prediction of risk and return classes with different classification algorithms is investigated.

To predict risk and return variables accurately, the effective factors need to be identified. In fact, one of the key issues of stock prediction design lies on how to select representative features for prediction (Zhang, Hu, Xie, Wang, et al., 2014).

Most studies in this area focus on technical features, financial ratios or macroeconomic indicators. For example, Tsai and Hsiao (2010) studied 8 financial ratios and 16 macroeconomic indicators as the main features to predict stock return by back propagation in Taiwan stock market. Cheng, Chen, and Lin (2010) conducted a comprehensive study on macroeconomic and technical features and studied 8 financial ratios and 10 macroeconomic indicators to investigate their effect on return variation in Taiwan stock market. By applying probabilistic back propagation algorithm, rough set and C4.5 Tree, they achieved 76% accuracy. de Oliveira, Nobre, and Zárate (2013) use 15 technical indicators and 11 fundamental indexes to prediction of stocks movement in Petrobras with artificial neural networks and obtain 87.50% for direct prediction. Tsai et al. (2011) considered 19 financial ratios and 11 macroeconomic indicators in Taiwan stock market by combining logistic regression algorithm, MLP back propagation and CART Tree to investigate their effect urn (negative or positive) on the stock

* Corresponding author at: No. 15, Shahriar 2 Alley, Danesh Street, Ardebil, Iran. Tel.: +98 9356546404; fax: +98 4517723386.
E-mail address: Sasan.barak@gmail.com (S. Barak).

http://dx.doi.org/10.1016/j.eswa.2014.09.026
0957-4174/© 2014 Elsevier Ltd. All rights reserved.

return and achieved 66.67% accuracy based on bagging and voting algorithms. In majority of studies, as mentioned, the focus is mostly on financial ratios, macroeconomic indicators, and technical indicators based on experts' ideas to predict returns. However, this paper presents a systematic and efficient methodology for comprehensive searching the potential representative features on stock market in 3 categories of financial ratio, profit and loss reports, and stock pricing models and not arbitrarily choosing likely effective features.

Furthermore, many studies have claimed and verified that feature selection (FS) is the key process in stock prediction modeling (Tsai & Hsiao, 2010). Zhang, Hu, Xie, Wang, et al. (2014) use a causal feature selection (CFS) algorithm to find effective features in Shanghai stock exchanges. The idea in their model is about causalities based feature selection algorithm. They assert that CFS represents direct influences between various stock features, while correlation based algorithms cannot distinguish direct influences from indirect ones. Wu et al. (2014) use textual and technical features to improve prediction accuracy of stock market. They use SVR algorithm and trend segmentation method to forecast trends and generate trading signals, respectively. Their feature selection algorithm is stepwise regression analysis. Although there are a variety of studies in the area of feature selection, almost all of them use a single feature selection model.

In this research, a novel hybrid feature selection algorithm on the basis of filter and function-based clustering method is applied to select the important features. What makes our proposed approach different from the previous ones is that we consider the combination of 9 different feature selection algorithms with function-based clustering algorithm. Hybrid model of our paper enjoys the power and advantage of correlation based algorithms like Chi-square, One-R in addition to the power of classified errors based, interval based, and information based algorithms like SVM, Relief-f, and Gini index/gain ration algorithms respectively. The effectiveness of our model is illustrated with the prediction of both risk and return of stocks and then analyzing the results with and without implementing of our hybrid feature selection algorithms.

To sum up, in the first stage of paper, a complete list of likely effective features on the stocks risks and returns are identified. After developing an appropriate database in the second stage, different classification algorithms are used to predict the risk and return. We also scrutinize on the effect of their results to our data base based on feature-oriented view point. Finally, in the third stage, a novel hybrid feature selection algorithm on the basis of filter and function-based clustering method is applied to select the important features which affect the prediction of risk and return.

The contribution of the paper is summarized as follow:

- A comprehensive and systematic study to identify the likely effective features in risk and return prediction.
- Stock risks as well as return prediction with different classification methods.
- Designing a hybrid feature selection algorithm on the basis of filter and function-based clustering.
- Finally, each algorithm with a feature-oriented view point is analyzed. The results indicate the factors which cause strength and weakness of that algorithm. As a result the nature of each feature is provided according to the amount of interference variable in their prediction.

The rest of the article is organized as follows. In Section 2, the proposed model is presented which has three stages. In Section 3, to illustrate the approach, we implement it with some real data from Tehran Stock Exchange (TSE). The results are analyzed in which the predictions with and without considering important effective features are also compared. Then in Section 4, a discussion on real return and risk prediction with important features has been represented. Finally, some conclusion and future research directions are provided in Section 5.

2. Proposed model

Our proposed algorithm which consists of three stages is shown in Fig. 1. In the first stage a database is developed and data is preprocessed. Non-systematic risk as well as real return is predicted with classification algorithms in the next stage. A hybrid feature selection algorithm is also presented in the third stage and risk and return are re-predicted based on selected features.

2.1. First stage: developing financial database

This stage we utilize the concepts and techniques of input features, response variables, and preprocessing models.

2.1.1. Input features

First we analyze and gather important features from the company's financial ratios and the profit and loss reports, as well as stock pricing models (Table 1).

- Financial ratio: to have a complete list of effective features we gather 4 general groups of financial ratio as a part of input variables of companies' database. The importance of these features is discussed in many studies (see (Bauer, Guenster, & Otten, 2004; Bernstein & Wild, 1999; Carnes & College, 2006; Huang, 2012; Omran & Ragab, 2004; Sadka & Sadka, 2009; Soliman, 2008)), also see financial ratio's part of Table 1.
- Stock pricing models: we review different stock pricing models (capital asset pricing model (CAPM), Gordon, Walter, Campbell–Shiller, and Fama–French) and obtain other important factors which effective on the risk and return prediction of stocks, see Table 2 (Kaplan & Ruback, 1995; Brealey, Myers, & Allen, 2007; Fama & French, 1993; Fama & French, 2012; Gordon, 1982; Hjalmarsson, 2010; Lee, Tzeng, Guan, Chien, & Huang, 2009; Lewellen, 2004; Mukherji, Dhatt, & Kim, 1997).
- Company's profit and loss reports: by using the profit and loss reports of companies, the other added factors are extracted. In Table 1, all input variables of financial model are provided.

2.1.2. Response variables

The most important response variables in our model are real return and non-systematic risk, as follows:

$$R = \sqrt[n]{\left(1+\frac{r_1}{100}\right)\left(1+\frac{r_2}{100}\right)\dots\left(1+\frac{r_n}{100}\right)} \quad (5)$$

where $r_1,\dots,r_n$ = real return of $1,\dots,n$th periods.

Non-systematic risk is defined as the standard deviation of the stock return, as follows.

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=0}^{n}\left(r_i - E(r)^2\right)} \quad (6)$$

2.1.3. Data pre-processing

Data preparing stage is an important part of the approach. Furthermore, it is time consuming in data mining process, described as follows.

- Removing high correlation features: features with higher than a predefined correlations percent on the basis of Pearson test are removed.