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:

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

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.

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

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.

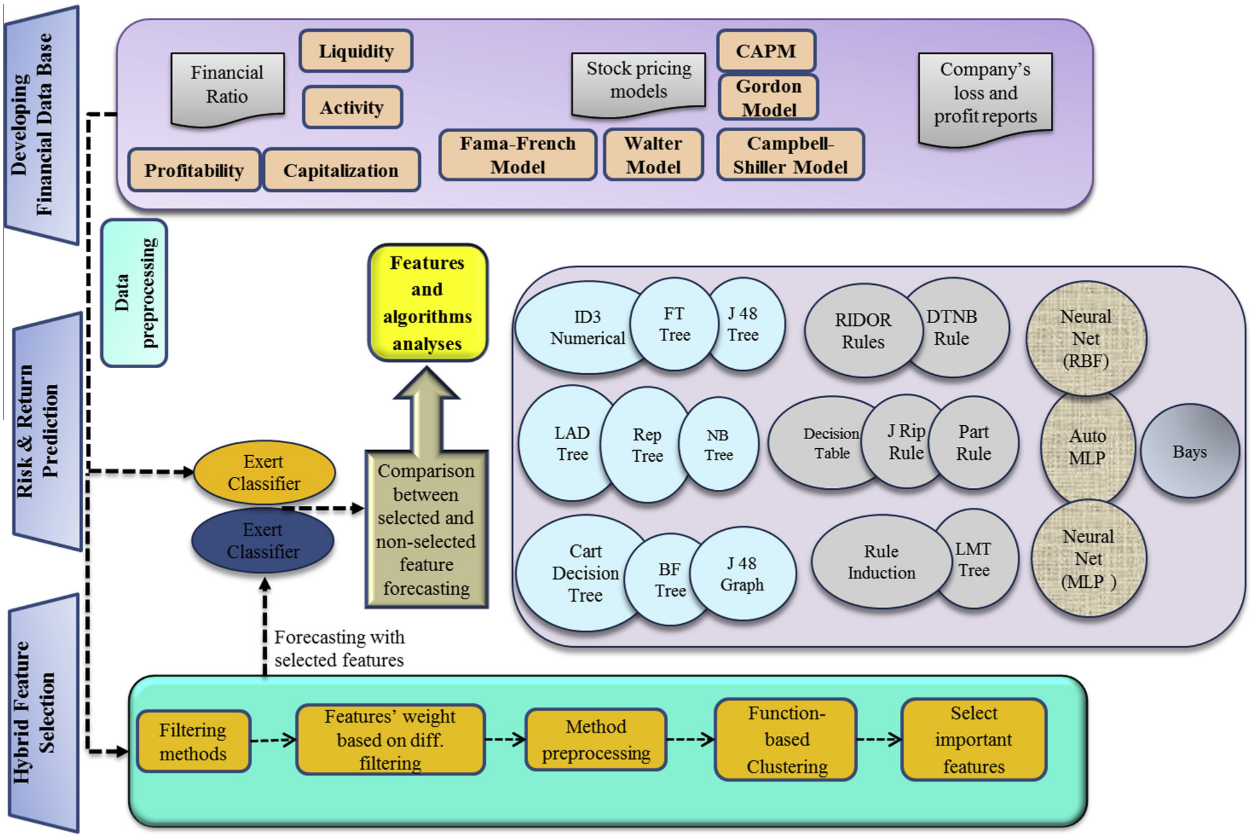

Fig. 1. Conceptual design of the proposed model.

Table 1
Financial input features.

Category		Features
Financial ratio	Liquidity ratios	Current ratio, quick ratio, current assets ratio, net working capital, liquidity ratios
	Activity ratio	Average payment period, current assets turn over, fixed asset turnover, total asset turnover
	Capitalization ratio	Equity ratio, debt coverage ratio, debt to total assets ratio, debt to equity ratio, long-term debt to equity ratio, current debt to equity ratio
	Profitability ratio	Percentage of net profit to sale, percentage of operating profit to sale, percentage of gross profit to sale, percentage of net profit to gross profit, return on asset (after tax) ROA, return on equity (after tax) ROE, working capital return percentage, fixed assets return percentage, assess the loan usefulness
Stock pricing models	Capital asset pricing model	r = return ration without risk β = stock beta coefficient (systematic risk) r_m = expected return from market
	Gordon model	EPS, DPS, EPS prediction, EPS cover, prediction difference percentage of EPS with the real amount, EPS growth ration in compare to the previous fiscal year
	Campbell–Shiller Model	P/E, P/S
	Walter model	Stock cumulative profit
	Fama–French Model	Company's capital (investment), stock book value, stock market value
Company's loss and profit reports	Total predicted income (last income prediction in the current fiscal year), total income growth % (total real income/(total real income – total predicted income)), predicted profit margins (last profit ratio/company's income in the current fiscal year), profit margin growth rate (real profit margin/(real profit margin – predicted profit margin)) and Efficiency (percent of daily trading volume/company's daily value in the before period)	

- Missing data: defected records caused by incomplete information of company or the company's negligence in reporting are also deleted from the database. Some Decision Tree algorithms and K Nearest neighbor techniques do not need to replace the missing data.
- Finding outlier data: to find outlier data in database, we use the distance-based approach which is based on data intervals (Knorr & Ng, 1999), density approach (Breunig et al., 2000) in which a parameter named Local Outlier Factor (LOF) is specialized to each sample based on K-Nearest neighbor density.

Table 2
Stock pricing models.

CAPM	$R = r_f + \beta(r_m - r_f)$	(1)	This model explains the connection between expected return and risk and it is used for bonds pricing with risk (Kaplan & Ruback, 1995) R: expected return, r_f: rate of return without risk, β: systematic risk, r_m: market expected return. A brief description about this formula can be found in Appendix A. On the basis of this model, r_m, r_f, β are added to database
Gordon model	$P = \frac{\text{DPS}}{k-g}$ DPS = EPS × DPR	(2)	Gordon has suggested this model using the investment of retained earnings to stock pricing (Gordon, 1982)
			g: stock profit increase, K: shareholder's expected return ratio. From this model two important factors EPS and DPS are achieved. ROE feature have been mentioned also in financial ration before. In addition to this, four other features that were obtained from EPS have been regarded and inserted into the model as follows: EPS prediction of companies in fiscal year, EPS coverage, prediction difference percentage of EPS with the real amount, and EPS growth ration in compare to the previous fiscal year (Hjalmarsson, 2010). Gordon model is used in different capital market discussions like (Lee et al., 2009)
Walter model	$P = \frac{\text{DPS}+(\text{EPS}-\text{DPS})r/k}{k}$	(3)	P: stock market price of each stock, r: internal rate of return, EPS–DPS: cumulative profit per share, K: capital rate cost (Brealey et al., 2007). According to this model, cumulative profit per share is known as a criterion in the database
Campbell–Shiller model	P/E, P/S ratio		This model calculates the stock P/E average by using Market data (Brealey et al., 2007). From the literature, it was clarified that P/E parameter is very important for analyzing and predicting the stock price, and it is inserted to database (Hjalmarsson, 2010; Lewellen, 2004). In addition to this, P/S ratio that is result of stock price divided to each stock sale is also inserted to database (Mukherji et al., 1997)
Fama–French model	$r_i = \alpha + \beta(r_m) + \beta_{size}(Size) + \beta_{\frac{\beta}{M}}\left(\frac{\beta}{M}\right)$	(4)	Fama–French offer β, size and book value to market value's model with the help of CAPM model as a multivariate regression to study the factors affecting portfolio returns (Fama & French, 1993, 2012) The first part of the model is similar to sharp model. The second part shows the company size which is a factor showing the company's capital and third part indicates the book value to market value. By using this model, company's capital, stock Book value and stock Market value are inserted to the database as 3 important factors. The other models like Glassman–Host and kernel are derived from the introduced methods and they do not help this research in finding new features

Samples with high LOF are known as outlier points, clustering approach (Hong & Wu, 2011) within the use of k means clustering algorithm, and deviation method (Hong & Wu, 2011).

2.2. Second stage: risk and return prediction with classification methods

Generally, researchers and scholars are seeking to achieve a more scientific model, ranging from Portfolio Theory by Markowitz in 1952 and Sharp assets pricing models in 1964, to Fama–French in 1992. However, they cannot solely evaluate price, risk, and return well. Bartholdy and Peare (2005) compared CAPM and Fama–French model while it appears that the latter can better explain the return deviation and can give better evidences. But regarding the real data, none of them can explain return well. Cao, Leggio, and Schniederjans (2005) concluded that the neural network is much more powerful than Fama–French model in stock return prediction. Dastgir and Afshari (2004) compared Walter, Gordon and current value of future cash flow stock pricing models in Tehran Stock Exchange and observed that real prices and prices obtained by models were not equal. As these studies show the traditional methods cannot necessarily estimate properly. Thus, it is necessary to apply some methods to be able to determine the complexity of the data. Some researchers have used different methods like neural networks and statistical methods. Among these results, the conclusion gained by machine learning algorithm and data mining are prominent (Patel et al., 2014).

Ou and Wang (2009) and Lai, Fan, Huang, and Chang (2009) concluded that Decision Tree methods have outstanding performance in stock return prediction. In addition, what is important is the rules obtained from the rule based algorithms and trees, since these rules conduct investors to buy and select the portfolio.

On the other hand, the output of the methods that are applied in this area (like SVM and NN) which do not use rules for prediction is not appropriate for practitioners. Decision Tree structure is more comprehensive, transparent and rational. On the basis of what was discussed, our study focuses on tree and rule based algorithms in order to be more appropriate for investors and analysts. Levin and Zahavi (2001) concluded that data correlation problem in tree algorithms is more transparent than statistical algorithms, and it can be solved by Pruning algorithms. Chang (2011) compared CART, back propagation, and CART–back propagation hybrid method from the point of view of stock price prediction based on fundamental data and concluded that back propagation and Decision Tree accuracy perform better than the hybrid methods.

In this study, by using different classification methods, risk and return are predicted on the basis of the given features and database. A comparison between different methods is performed. Actually, this section is done for two times. In the first time the prediction is done with all features but in the second time, the best selected features from hybrid feature selection algorithm are predicted. A comprehensive comparison between these two predictions is also done. In other words, in this paper we compare the accuracy of risk and return forecasts with and without feature selection, based on different classification methods and explain the effect of feature selection on classification methods. The classification algorithms are shown in Fig. 1.

2.2.1. Testing strategy

In order to get robustness prediction, we perform 10-fold cross-validation model on the predictors (duda, Hart, & Strok, 2001). This method has been proved to be statistically good enough in evaluating the performance of the predictive model (Mitchell, 1997). In 10-fold cross-validation, the training set is equally divided into 10 different subsets. Nine out of 10 of the subsets are used to train the classifier and the tenth subset is used as the test set. The

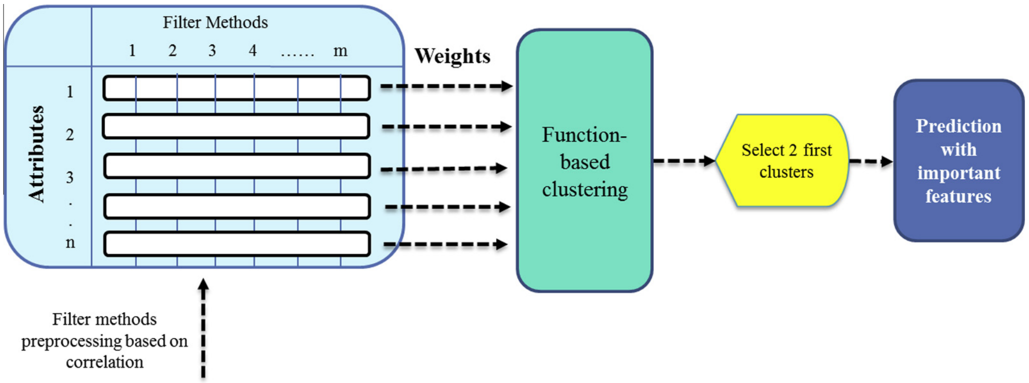
Filter Methods
1 2 3 4 …… m
Attributes
1
2
3
.
.
n
Weights
Function-based clustering
Select 2 first clusters
Prediction with important features
Filter methods preprocessing based on correlation

Fig. 2. Hybrid feature selection.

Table 3
Confusion matrix for five classes.

		Predicted class				
		Very low	Low	Normal	High	Very high
Actual class	Very low	a_1	b_1	c_1	d_1	e_1
	Low	a_2	b_2	c_2	d_2	e_2
	Normal	a_3	b_3	c_3	d_3	e_3
	High	a_4	b_4	c_4	d_4	e_4
	Very high	a_5	b_5	c_5	d_5	e_5

procedure is repeated 10 times, with a different subset being used as the test set and the best result has been chosen.

In order to reliably evaluate the predictors, we consider not only prediction accuracy but sensitivity and specificity. The accuracy of a predictor on a given test set is the percentage of test set tuples that are correctly predicted by the predictor. Prediction accuracy for five classes can be measured by a confusion matrix shown in Table 3 with formula (1).

$$Accuracy = \frac{a_1 + b_2 + c_3 + d_4 + e_5}{\sum_{i=1}^{5} a_i + b_i + c_i + d_i + e_i} \tag{7}$$

Sensitivity is also referred to the proportion of positive tuples that are correctly identified while specificity is the proportion of negative tuples that are correctly identified (Han & Kamber, 2006).

2.3. Third stage: hybrid feature selection

Based on special conditions in stock exchange, occasionally we encounter with many attributes whereas some of them no longer have useful information and just complicate the condition. For this reason feature selection is one of the very crucial aspect that has a highly regarded recommendation (Huang, 2012; Huang, Yang, & Chuang, 2008; Tsai & Hsiao, 2010).

In this section to investigate the features which have greater effect on risk and return and better analysis of algorithms results a novel feature selecting method in 2 levels is established. It should be noted that feature selecting in capital markets issues has double importance. The reason is that we encounter with so many features that are either useless or have low information value. Thus, dealing with these features is time wasting without any gain. Feature selection methods are generally divided into three categories: (1) filter methods, (2) wrapper methods, and (3) hybrid methods (Chen & Cheng, 2012). In this approach we use a hybrid model based on combination of filter and function-based clustering method to extract a set of efficient features as a follow (see Fig. 2).

2.3.1. Filter methods

According to Witten and Frank (2011) 7 algorithms were defined as Filter method: Chi square (Kononenko, 1994), Info Gain (Dumais, Platt, Heckerman, & Sahami, 1998), Gain Ratio (duda et al., 2001), Relief-f (Kononenko, 1994), consistency (Liu & Setiono, 1996), One R (Holte, 1993) and CFS (Hall, 1998). In addition, Symmetrical Uncertainty and SVM algorithm are also used for weighting the features (Chen & Cheng, 2012). In this section, to compare the importance of each feature using mentioned methods, a comprehensive analysis was conducted on the features and eventually the weightings of features are presented.

2.3.2. Function-based clustering method

After attaining the features weights by different filter based algorithms we have n attributes with m attributes' weight and then we need a model to determine the important features' clustering between these weighted attributes. In this section we develop Li (2006a) function-based clustering method. This model is based on hierarchical divisive clustering method which begins with one cluster including all objects, ($\mathbf{X}_{n\times m}$).

For the object $\mathbf{x}_1,\ldots,\mathbf{x}_n$, we denote the vector of group membership of objects as $\mathbf{z} = (z_1,\ldots,z_n)^T$, where $\mathbf{z} \in \mathbf{Z}$, and $\mathbf{Z}$ is the space of sign vectors defined to be

$$\mathbf{Z} = \{\mathbf{z} = (z_1,\ldots,z_n)^T | z_i = \pm 1\} \tag{8}$$

All objects that are associated with an entry of 1 in $\mathbf{z}$ are classified into one group, whereas the others with an entry of -1 are classified into the other group.

Then by using the model of multivariate analysis of variance defined to be

$$\mathbf{x}_i = \boldsymbol{\mu} + z_i\boldsymbol{\gamma} + \boldsymbol{\varepsilon}_i, \quad i = 1,2,\ldots,n \tag{9}$$

where the error vectors ε_i are assumed to be normally distributed with a zero mean and a common covariance matrix V, i.e. $N(0,V)$. In addition ε_i and ε_j $(i \neq j)$ are assumed to be independent. Then by maximum likelihood, the clustering problem is formulated as a least squares optimization problem.

$$\min_{\alpha,\boldsymbol{\beta},\mathbf{z}\in\mathbf{Z}} \left\{(\mathbf{z} - \alpha\mathbf{1} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{z} - \alpha\mathbf{1} - \mathbf{X}\boldsymbol{\beta})\right\} \tag{10}$$

Simultaneously the unknown vector of cluster membership and the coefficients of the linear clustering function are estimated. The computation of the clustering-function-based method will be converted to that of sign analysis (Li, 2006b), and by problem solving two clusters is achieved.

Next, one of these groups based on higher within-group dispersion matrix is further divided into two dissimilar subgroups. The process continues until some stopping criterion has been satisfied.

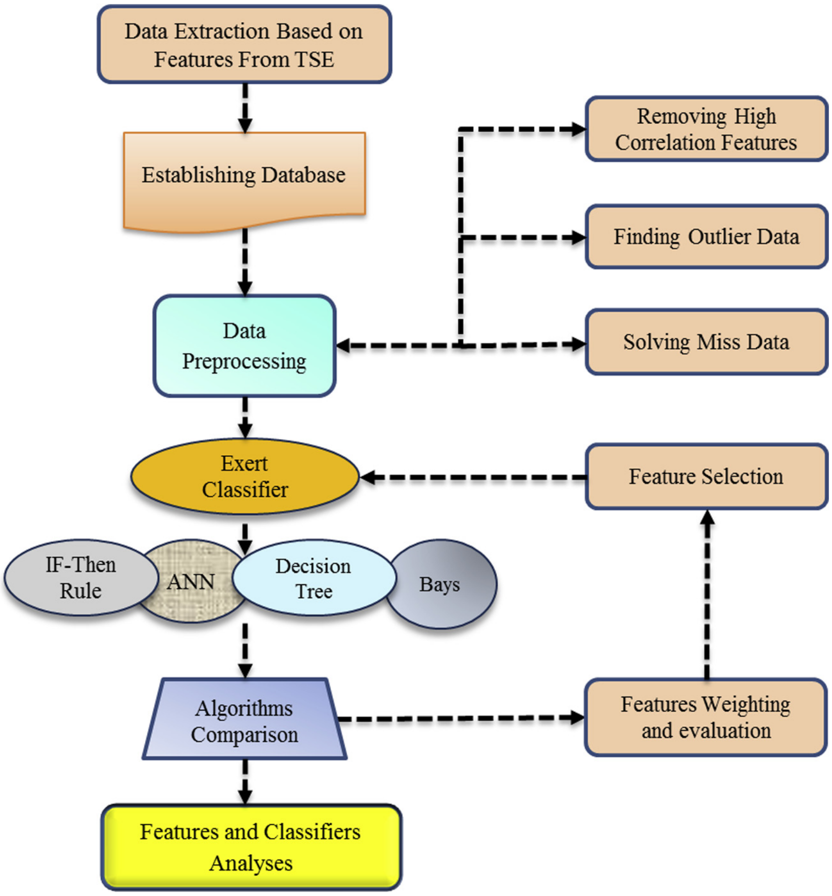

Fig. 3. Experimental results process.

Most of the stopping criteria are based on within-group dispersion and/or between-group dispersion matrices.

By this approach we use the advantages of different filter methods and use these weighting attributes by function-based clustering method to make more accurate decision of effective feature.

3. Experimental results and analysis

In this study a database including 44 input features and 2 goal features are gathered from TSE data from 2003 to 2012. The resulting database has 1963 records for 400 companies.

According to a group of experts, 5 intervals were introduced for the real return: very high with a range higher than 9.3, high with the range of 4–9.3, average with a range of 1.14–4, low with the range of −1.3 to 1.14 and very low that lower than −1.3. Risk is also classified in 3 intervals: high in range of higher than 15.5, average in range of 6.3–15.5 and low in rage of lower than 6.3. According to literature, it is found out that to predict return, negative and positive return(Tsai et al., 2011; Wang & Chan, 2006) and negative and positive return trend (Enke & Thawornwong, 2005) are used (see also (Patel et al., 2014; Yu et al., 2014; Zhang, Hu, Xie, Wang, et al., 2014)). For more accuracy, we increased the prediction intervals. These intervals give more information to investors and they can develop a balance between the share price and the future gained return. In fact, the information which is limited to the company's profitability or losses does not help them very much. Beside this, risk has been rarely mentioned in prediction field of stock exchange. We can conclude whether the proposed return range is optimal or not just by knowing the risk amount. The previous studies of the field have just focused on return prediction while these 2 features together show the portfolio efficient frontier and investors can use it to select the best optimal portfolio. The process is illustrated in Fig. 3.

3.1. Data pre-processing

- *Removing high correlation features*: features with higher than 0.95% correlations on the basis of Pearson test were removed. Therefore, features including: gross profit to sale percentage, assess the loan usefulness, stock cumulative profit, fixed assets return percentage, debt to equity ratio, and current debt to equity ratio are removed. Their correlation with return and risk variables is higher than 95% in comparison with the other features. The correlation between real return and risk with other features has been illustrated in Tables 4 and 5 respectively.
- *Finding outlier data and miss data*: to find outlier data in database, at first we used the distance-based approach and by analyzing remote records we concluded that some are very large governmental companies that are not applicable in our study and in fact they are not outlier data. Other outliers were also deleted. By using the density approach 12 records were known as outlier points, in which 7 of them were large companies. Thus they remained in the database. However, others were omitted, mostly because they did not provide accurate information. With clustering approach we determined some outlier data that were in none of the clusters. As a result 6 samples were identified as outlier data. By analyzing input feature of the company, no suspected case was found and no company was omitted. Finally, by using the techniques based on the deviation, 5 records were known as outlier points. Analyzing records

Table 4
Correlation between real return and other features.

Debt to total assets ratio	Net profit to sale	Operating profit to sale	Net profit to Gross profit	ROA	ROE	Current assets turn over	Fixed asset turnover	Current ratio	Quick ratio
0.0023	0.017	0.0488	−0.0131	−0.1494	−0.0113	0.0005	−0.0019	−0.0304	0.0103
Return from market	Net working capital	Average payment period	Current assets ratio	Working capital return	Total asset turnover	Equity ratio	Predicted profit margins	Long-term debt to equity ratio	Debt coverage ratio
0.0218	0.0251	−0.0113	0.0128	−0.0021	−0.0937	0.0977	0.0066	0.0011	0.0246
Efficiency	DPS	EPS	Capital	EPS prediction	EPS difference with real EPS	Return ratio without risk	EPS growth	Total predicted income	
0.0135	−0.1168	−0.1402	0.0059	−0.1211	0.0037	−0.0384	0.0194	−0.0547	
Total income growth	Profit margin growth rate	EPS cover	Liquidity ratios	*P*/*E*	Stock book value	*P*/*S*	Stock market value	Beta coefficient	
−0.0231	0.0135	0.0171	−0.0378	0.0058	−0.0377	0.0109	0.0651	0.0153	

Table 5
Correlation between risk and other features.

Debt to total assets ratio	Net profit to sale	Operating profit to sale	Net profit to Gross profit	ROA	ROE	Current assets turn over	Fixed asset turnover	Current ratio	Quick ratio
0.0203	−0.0301	0.0444	0.0148	−0.0307	−0.0176	−0.0055	−0.0062	0.05	−0.0015
Return from market	Net working capital	Average payment period	Current assets ratio	Working capital return	Total asset turnover	Equity ratio	Predicted profit margins	Long-term debt to equity ratio	Debt coverage ratio
0.0414	0.0056	0.0099	−0.0057	−0.0021	−0.0195	−0.0112	0.0296	0.0212	−0.0126
Efficiency	DPS	EPS	Capital	EPS prediction	EPS difference with real EPS	Return ratio without risk	EPS growth	Total predicted income	
0.0135	0.0099	0.0222	−0.0317	0.02251	−0.0156	−0.0935	0.0159	−0.0247	
Total income growth	Profit margin growth rate	EPS cover	Liquidity ratios	*P*/*E*	Stock book value	*P*/*S*	Stock market value	Beta coefficient	
−0.0369	−0.0031	−0.0313	−0.0238	−0.0114	−0.0089	−0.0149	0.0251	−0.0283	

Table 6
Algorithms Comparison for real return variable.

Algorithm	Accuracy	Sensitivity	Specificity	Number of rules	Tree size	Number of leaves
LAD Tree[a] (Hall et al., 2009)	78.00	77.15	75.29	–	31	15
Cart Decision Tree[b]	76.50	74.27	74.3	–	13	7
DTNB rule[c] (Hall & Frank, 2008)	76.00	75.08	73.55	998	–	–
Decision table[d]	75.50	75.14	72.44	56	–	–
Rep Tree[e]	75.00	74.09	71.64	–	33	17
RIDOR rules (Witten & Frank, 2011)	75.00	75.3	73.07	208	–	–
J Rip Rule (Witten & Frank, 2011)	74.90	73.18	74.02	9	–	–
BF Tree[f] (Shi, 2007)	74.50	78.49	74.28	–	9	5
Part Rule[g] (Frank & Witten, 1998)	72.60	67.84	69.15	104	–	–
J 48 Graph	71.50	69.68	66.38	–	1619	810
NB Tree (Witten & Frank, 2011)	71.00	70.2	69.5	–	97	49
LMT Tree (Witten & Frank, 2011)	70.50			–	30	20
Neural Net (RBF)	70.00	67.3	66.4	–	–	–
Neural Net (MLP[h])	69.00	66.3	67.1	–	–	–
Auto MLP[i]	70.00	67.01	66.02	–	–	–
Rule Induction[j]	68.50	69.27	66.76	56	–	–
FT Tree	68.50	66.3	64.8	–	45	28
J 48 Tree	67.39	66.2	65.6	–	303	152
ID3 Numerical	61.50	58.42	57.47	–	1905	979
Bays	60.00	58	62.28	–	–	–

[a] It uses AD Tree with boosting for prediction and cross validation to select training data and class label decisions are done on the basis of this algorithm most votes.

[b] The used tree's split has been done by Gini index algorithm and the Pruning is done based on cost – complexity after constructing the tree.

[c] At first, these models determined the important variables by using Naive Bays algorithm (18 attribute achieve) and then offer classification prediction rules are provided by Decision Tree.

[d] Algorithm first uses the Forward election algorithm to determine the input variables. After 375 implementing the algorithm, ROE, Net working capital, EPS prediction and return are known as effective features and then based on best first (BF) algorithm the model is constructed.

[e] This algorithm is based on information gain and its Pruning is based on prediction error minimization.

[f] BF Tree uses binary pruning to construct a tree based on selecting the first important feature as nodes point. Based on this, the model has found the feature that best predicts the output variable among other input variables. Then a binary tree based on this variable is constructed. Best–First (BF) Decision Tree just use return variable to prediction real return. (Just this input is considered in this method's tree).

[g] This method has used C4.5 algorithm in every implantation and has used the best of them as a new rule in the model rules.

[h] Feed forward multi layout perceptron based on Levenberg–Marquardt algorithm with 12 neuron in hidden layer.

[i] Number of neurons and hidden layers are optimized.

[j] This algorithm gains the rules based on the information gain and first gained rule is: if return without risk $\leqslant 11.967$ and return without risk > -12.164 and EPS > 51.447 and EPS coverage percent > 164.500 then low. This algorithm, gains the rule based on the information gain and based on decrease amount in model accuracy while constructing the rules we prune them. And will construct the model till there is no other variable to be added to the model or the error amount is more than 0.5.

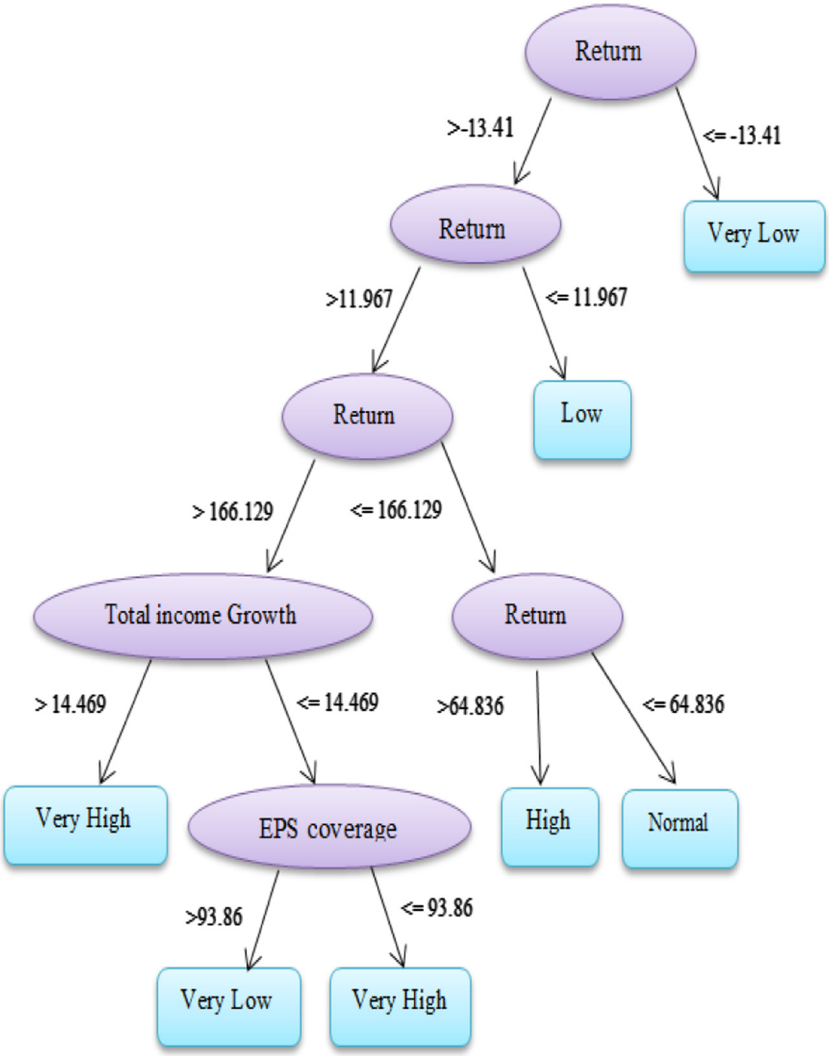

Fig. 4. Real return prediction – Cart Decision Tree.

clarified that input feature of these records were pertained to the previous fiscal year and then were removed. Also, among 1963 records, 12 records because of miss value derived from lack of information are deleted.

3.2. Comparison of algorithms

Table 6 shows the results of Decision Trees, rule base algorithms, and neural networks accuracy for real return prediction. As it is clear (from Table 6), LAD Tree algorithm has achieved a higher accuracy in the prediction of return. The other algorithms, like SVM and K-Nearest neighbor, had accuracy of close to 60%. Thus, due to their low accuracy, we did not apply them for analysis. Low accuracy of SVM algorithm can be because of its high sensitivity to the missed data. In fact, it is generally an algorithm to predict 2 class outputs, while we are dealing with multi-class data and many missed data.

From investigating the results it can be stated that generally denser trees have shown better accuracy than big ones. This is clear in the cases of ID3 Numerical, J48 Graph, and J48 Tree. Due to the importance of pruning after tree construction, since these models have no pruning stage or their pruning algorithm is not efficient, their accuracy is not acceptable.

Trees with the size of less than 33 for real returns have an accuracy of higher than 70%. Despite the medium accuracy of these trees compared to larger trees, they have higher accuracy in test data.

DTNB Rule algorithm has the highest accuracy in comparison with the other “If-Then Rules” algorithms on test data. On average, the accuracy of “If-Then Rules” algorithms is better than trees, however the best prediction is obtained by LAD Tree. Some algorithms of tree types are shown in Figs. 4–7 (returns in all figures’ nodes are return ration without considering risk).

Similarly, risk is predicted as shown in Table 7.

It can be stated that generally larger trees (for example higher than 300) and smaller trees have no prominent results in comparison to medium sized trees. In comparison to “If-Then Rules” algorithms, the highest prediction accuracy for test data is gained from DTNB, similar to real return prediction. For risk prediction also “If-Then Rules” algorithms accuracy is also better than those gained by tree, but the best prediction is gain by LAD Tree. To predict risk, neural network results have lower accuracy in comparison with the prediction of return. In Fig. 8 LAD Tree for risk prediction is depicted.

3.3. Prediction after hybrid feature selection

In this section, we develop a hybrid feature selection to evaluate each feature. In the first stage by using the above mentioned filter methods, a comprehensive analysis of the features is conducted to

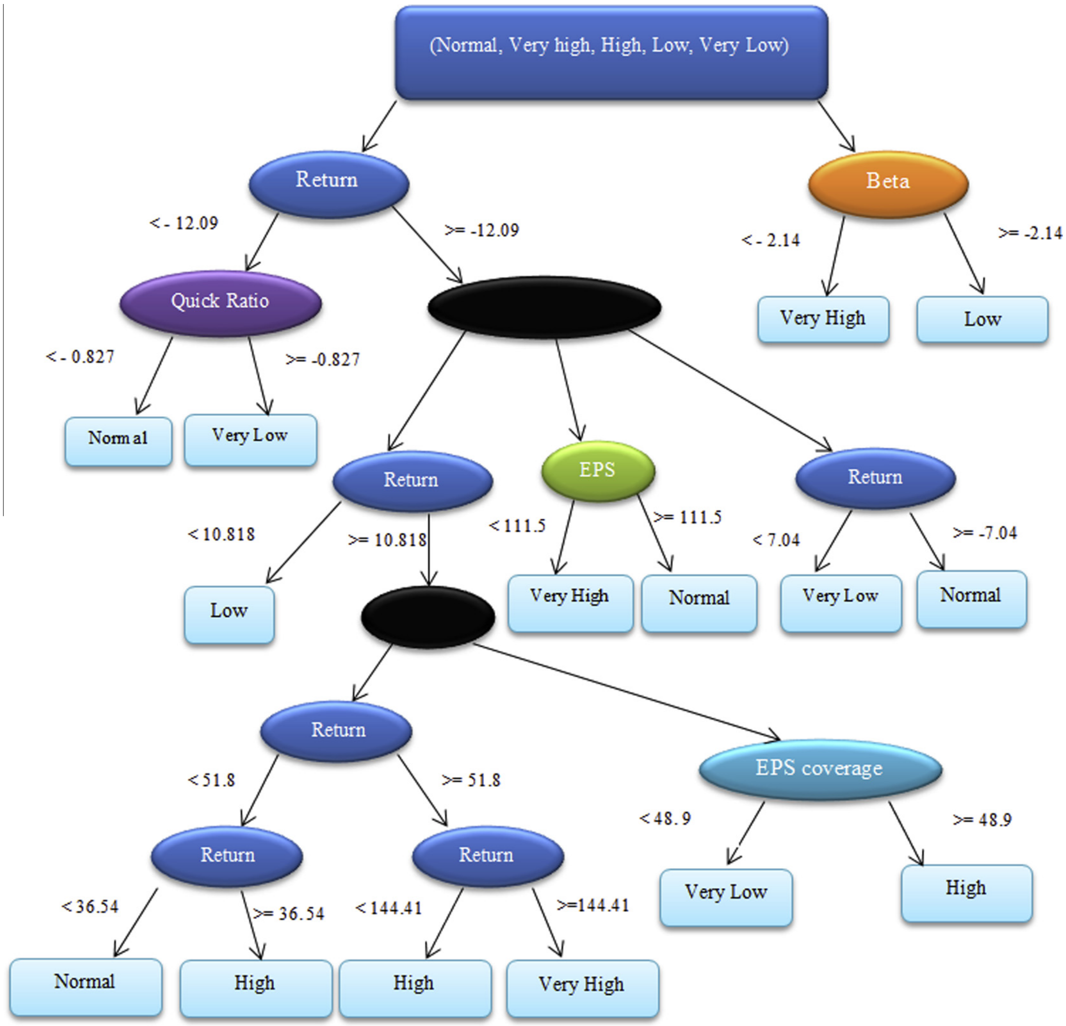

Fig. 5. Real return prediction – LAD Tree.

```
0 Return< -13.10281: very low (365.0/62.0)
0 Return >= -13.10281
1 Return < 10.81819: Low (392.0/96.0)
1 Return >= 10.81819
2 Return < 53.97851: normal (326.0/126.0)
2 Return >= 53.97851
3 Return < 136.66997: high (184.0/63.0)
3 Return >= 136.66997: Very high (72.0/27.0)
```

Fig. 6. Best-first Decision Tree.

assign suitable weights for features. (All features even high correlation features are considered.)

Depending on the evaluation function, we applied the following filter methods:

- Based on interval method: Relief-f method.
- Based on information: Gini index, Information Gain Ratio, Information gain, Symmetrical Uncertainty methods.
- Based on correlation: Chi square, One R methods.
- Based on consistency: consistency method.
- Based on classified errors: SVM method.

The CFS method did not have any acceptable result because of dependence and was eliminated in method pre-processing step.

We have applied Rapid Miner and Weka software to implement the algorithms (Hofmann & Klinkenberg, 2013). Furthermore, the algorithm which is used by Weka is mentioned, like Weka IG, Weka Chi-2....

The resulting weights for features of risk parameter and real return parameter are shown in Tables 8 and 9, respectively.

After attaining the features weights by different filter based algorithms we have 13 columns (m) with 44 attributes' weight (n) and then because of obtain accurate clusters, we use preprocessing on this data set. Therefore, the seventh column (Weka IG) of the Tables 4 and 5 is put aside from the analysis, because of its correlation with other ones. This way we do the grouping for 12 columns and 44 features.

Then, we use the clustering-function-based method for clustering attributes to predict the most important features. Usually the first and second clusters are the effective ones and we choose them (Li, 2006b).

For real return, in the first step of clustering, return and market return is separated from the other attributes. In other words, they are more important with higher weights compared with other attributes. In the second step, 6 features out of 43 remaining ones are separated. Finally, eight features in the first and second clusters are considered as important features. Similarly, for risk parameter in the first step, three features, return, beta coefficient and efficiency of 44 features can be separated. In

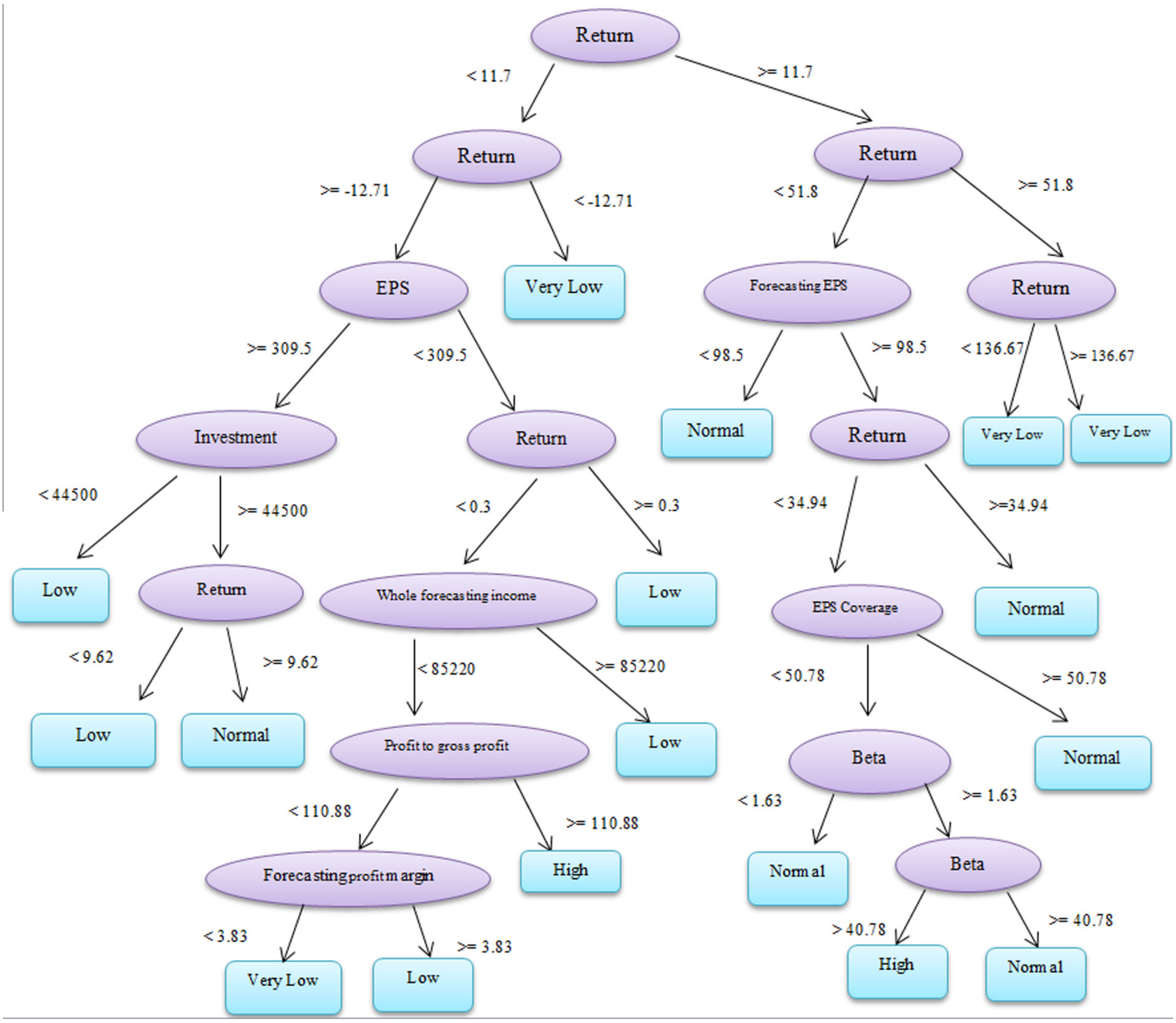

Fig. 7. Real return prediction – Rep Tree.

Table 7
Algorithm comparison for risk variable.

Algorithm[a]	Accuracy	Sensitivity	Specificity	Number of rules	Tree size	Number of leaves
LAD Tree	78.24	69.62	80.24	–	31	20
DTNB rule	77.41	69.44	78.51	426	–	–
Decision table	76.57	65.38	81.32	297	–	–
BF Tree[b]	76.15	67.41	74.71	–	109	55
J Rip Rule	74.90	73.7	74.3	9	–	–
J 48 Graph	73.64	64.68	71.66	–	721	361
Part Rule	73.64	65.36	71.64	55	–	–
Rep Tree	72.80	67.13	69.39	–	77	39
Rule Induction	71.55	64.45	70.25	59	–	–
J 48 Tree	71.55	70.9	72.5	–	313	157
FT Tree	67.78	65.6	64.8	–	63	32
NB Tree	66.95	66.3	64.7	–	7	4
Neural Net (MLP)	59.00	61.2	59.22	–	–	–
ID3 Numerical	57.00	55.1	54.2	–	553	403
Bays	55.65	57.3	50.2	–	–	–

[a] Some algorithms that used in real return prediction have low prediction accuracy in risk prediction and we do not report their results.
[b] The beta coefficient is known as the first leaf.

the second step, 12 features out of the 42 remaining ones can be separated. Finally 15 features from 44 features are selected as important ones. The results for risk and real return parameters are presented in the Table 10.

As can be seen, more features were selected for the risk variable than with real return. The classification results with selected features show in parenthesis at Table 11, in which "Deviation = Accuracy base on selected feature – Accuracy base

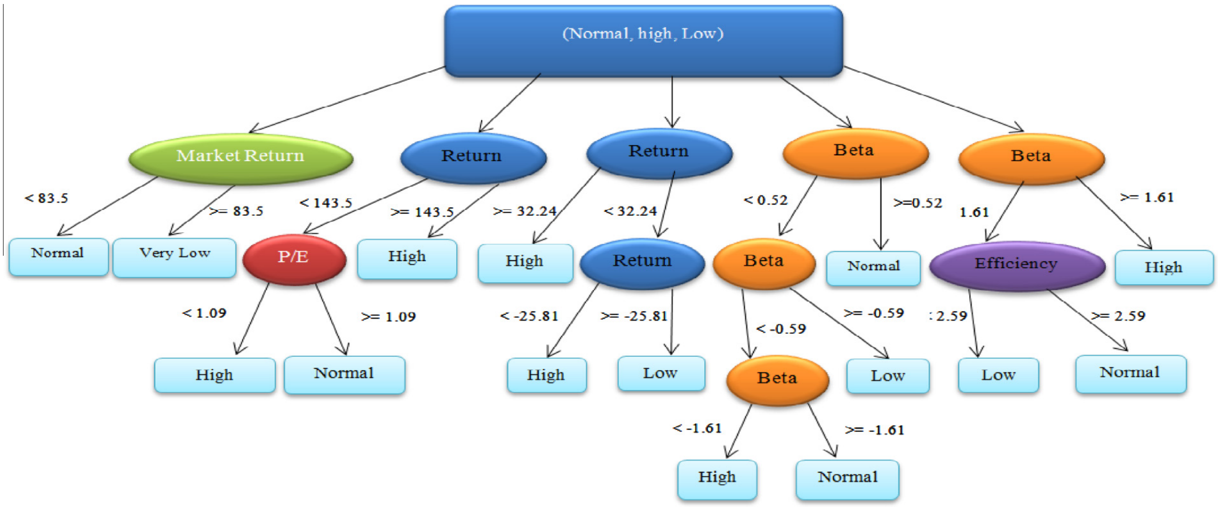

Fig. 8. Risk prediction – LAD Tree.

Table 8
Weighting for risk parameter.

Attributes	Chi-2	IG Ratio	Info Gain	R-f	SVM	Consistency	Weka IG	Weka chi-2	Weka con	Weka IGR	Weka R-f	Gini index	Weka One R
Return	1	0.69	0.61	1	0.024	0.87	0.513	0.437	0.513	0.63	0.9	0.725	0.759
Beta	0.188	0	0.062	0.397	0.079	0.163	0.097	0.073	0.097	0.132	0.519	0.061	0.124
Efficiency	0.091	0.707	0.076	0.071	0.083	0.099	0.056	0.044	0.056	0.112	0.093	0.093	0.127
Market return	0.058	0.67	0.072	0.019	0.023	0.058	0.106	0.092	0.106	0.157	0.021	0.064	0.117
EPS prediction %	0.008	0.615	0.099	0	0.076	0.005	0.05	0.039	0.05	0.17	0.002	0.112	0.111
Long-term debt to equity	0.032	0.487	0.076	0.011	0.029	0.039	0.085	0.078	0.085	0.136	0.015	0.077	0.118
Total income growth %	0.021	0.707	0.055	0.005	0.045	0.016	0.03	0.025	0.03	0.16	0.002	0.075	0.095
EPS growth %	0.025	0.707	0.046	0.005	0.061	0.016	0.026	0.021	0.026	0.136	0.002	0.061	0.095
ROE	0.058	0.328	0.074	0.054	0.07	0.073	0.039	0.031	0.039	0.135	0.046	0.092	0.089
DPS	0.046	0.328	0.078	0.047	0.025	0.063	0.041	0.033	0.041	0.15	0.045	0.089	0.123
Debt to total assets ratio	0.07	0.338	0.057	0.035	0.065	0.086	0.03	0.024	0.03	0.141	0.034	0.057	0.1
Profit margin growth rate	0.01	0.615	0.036	0.005	0.084	0.008	0.021	0.017	0.021	0.136	0.009	0.045	0.055
EPS	0.046	0.421	0.024	0.063	0.088	0.045	0.016	0.013	0.016	0.056	0.131	0.033	0.093
P/*E*	0.027	0.328	0.064	0.013	0.108	0.041	0.032	0.025	0.032	0.102	0.002	0.075	0.127
Predicted profit margin	0.053	0.319	0.067	0.071	0.023	0.07	0.036	0.028	0.036	0.11	0.037	0.063	0.047
Stock market value	0.029	0.615	0.023	0.013	0.043	0.029	0.016	0.012	0.016	0.065	0.011	0.027	0.049
ROA	0.033	0.422	0.038	0.01	0.038	0.046	0.023	0.019	0.023	0.152	0	0.056	0.049
Current debt to equity ratio	0.039	0.381	0.046	0.003	0.043	0.049	0.026	0.018	0.026	0.121	0.002	0.039	0.092
Debt to equity	0.036	0.419	0.047	0	0.012	0.05	0.041	0.039	0.041	0.114	0.005	0.026	0.043
Book value	0.015	0.615	0.013	0.004	0.131	0.016	0	0	0	0	0.001	0.021	0.048
Assess the loan usefulness	0.039	0.421	0.026	0.01	0.031	0.052	0.018	0.016	0.018	0.122	0.006	0.018	0.075
Equity ratio	0.018	0.615	0.003	0.001	0.099	0.019	0	0	0	0	0	0.011	0.04
Gross profit to sale	0.021	0.107	0.055	0.005	0.034	0.016	0.03	0.025	0.03	0.134	0.002	0.074	0.059
Current ratio	0.016	0.615	0.008	0.002	0.073	0.018	0	0	0	0	0	0.013	0.032
Operating profit to sale	0.017	0.615	0	0.001	0.09	0.02	0	0	0	0	0.004	0.002	0.023
P/*S*	0.014	0.615	0.01	0.002	0.06	0.015	0	0	0	0	0	0.018	0.037
Fixed asset turnover	0.032	0.421	0.006	0.027	0.059	0.024	0	0	0	0	0.128	0.011	0.061
Fixed assets return	0.016	0.615	0.008	0.002	0.058	0.018	0	0	0	0	0	0.013	0.032
Debt coverage ratio	0.005	0.366	0.041	0.002	0.045	0.008	0.024	0.019	0.024	0.073	0	0.052	0.088
Liquidity ratio	0.015	0.421	0.028	0.011	0.018	0.022	0.033	0.027	0.033	0.083	0.006	0.025	0
Net profit to sale	0.023	0.377	0.021	0.004	0	0.04	0.015	0.012	0.015	0.046	0.001	0.03	0.068
Working capital return percentage	0.002	0.319	0.027	0.012	0.094	0.003	0.017	0.015	0.017	0.064	0.001	0.02	0.049
EPS deviation	0	0.338	0.026	0.006	0.033	0	0.017	0.014	0.017	0.06	0	0.038	0.067
Capital	0.034	0.371	0.016	0.01	0.04	0.036	0	0	0	0	0.025	0.022	0.052
Current assets ratio	0.034	0.371	0.016	0.009	0.024	0.037	0	0	0	0	0.025	0.022	0.041
Total predicted income	0.009	0.347	0.024	0.003	0.026	0.015	0.016	0.013	0.016	0.051	0	0.034	0.019
Net profit to gross profit	0.016	0.338	0.009	0.03	0.068	0.019	0	0	0	0	0.027	0.015	0.013
EPS coverage percent	0.008	0.328	0.016	0.014	0.061	0.016	0	0	0	0	0.003	0.024	0.063
Net working capital	0.015	0.358	0.009	0.019	0.011	0.025	0	0	0	0	0.001	0.013	0.064
Average payment	0.018	0.319	0.011	0.034	0.024	0.034	0	0	0	0	0.013	0.007	0.053
Total asset turnover	0.015	0.358	0.008	0.019	0.011	0.025	0	0	0	0	0.001	0.013	0.054
Quick ratio	0.027	0.328	0.009	0.001	0.043	0.038	0	0	0	0	0.01	0.015	0.027
Stock cumulative profit	0.012	0.338	0.004	0.005	0.016	0.018	0	0	0	0	0.001	0.008	0.033
Current assets turnover	0.012	0	0	0.01	0.073	0.019	0	0	0	0	0.007	0	0.051

Table 9
Weighting for real return parameter.

Attributes	Chi-2	IGR	IG	R-f	SVM	Consistency	Weka IG	Weka chi-2	Weka con	Weka IGR	Weka R-f	Gini index	Weka oneR
Return	1	0.69	0.61	1	0.024	0.87	0.513	0.437	0.513	0.63	0.9	0.725	0.759
Market return	0.188	0	0.062	0.397	0.079	0.163	0.097	0.073	0.097	0.132	0.519	0.061	0.124
ROA	0.091	0.707	0.076	0.071	0.083	0.099	0.056	0.044	0.056	0.112	0.093	0.093	0.127
Beta	0.058	0.67	0.072	0.019	0.023	0.058	0.106	0.092	0.106	0.157	0.021	0.064	0.117
ROE	0.008	0.615	0.099	0	0.076	0.005	0.05	0.039	0.05	0.17	0.002	0.112	0.111
EPS growth %	0.032	0.487	0.076	0.011	0.029	0.039	0.085	0.078	0.085	0.136	0.015	0.077	0.118
Profit margin growth rate	0.021	0.707	0.055	0.005	0.045	0.016	0.03	0.025	0.03	0.16	0.002	0.075	0.095
Operating profit to sale	0.025	0.707	0.046	0.005	0.061	0.016	0.026	0.021	0.026	0.136	0.002	0.061	0.095
EPS	0.058	0.328	0.074	0.054	0.07	0.073	0.039	0.031	0.039	0.135	0.046	0.092	0.089
DPS	0.046	0.328	0.078	0.047	0.025	0.063	0.041	0.033	0.041	0.15	0.045	0.089	0.123
EPS prediction %	0.07	0.338	0.057	0.035	0.065	0.086	0.03	0.024	0.03	0.141	0.034	0.057	0.1
Predicted profit margin	0.01	0.615	0.036	0.005	0.084	0.008	0.021	0.017	0.021	0.136	0.009	0.045	0.055
Net profit to sale	0.046	0.421	0.024	0.063	0.088	0.045	0.016	0.013	0.016	0.056	0.131	0.033	0.093
EPS deviation	0.027	0.328	0.064	0.013	0.108	0.041	0.032	0.025	0.032	0.102	0.002	0.075	0.127
Efficiency	0.053	0.319	0.067	0.071	0.023	0.07	0.036	0.028	0.036	0.11	0.037	0.063	0.047
P/*S*	0.029	0.615	0.023	0.013	0.043	0.029	0.016	0.012	0.016	0.065	0.011	0.027	0.049
Net profit to gross profit	0.033	0.422	0.038	0.01	0.038	0.046	0.023	0.019	0.023	0.152	0	0.056	0.049
Quick ratio	0.039	0.381	0.046	0.003	0.043	0.049	0.026	0.018	0.026	0.121	0.002	0.039	0.092
Equity ratio	0.036	0.419	0.047	0	0.012	0.05	0.041	0.039	0.041	0.114	0.005	0.026	0.043
Stock market value	0.015	0.615	0.013	0.004	0.131	0.016	0	0	0	0	0.001	0.021	0.048
Book value	0.039	0.421	0.026	0.01	0.031	0.052	0.018	0.016	0.018	0.122	0.006	0.018	0.075
Long-term debt to equity	0.018	0.615	0.003	0.001	0.099	0.019	0	0	0	0	0	0.011	0.04
Gross profit to sale	0.021	0.107	0.055	0.005	0.034	0.016	0.03	0.025	0.03	0.134	0.002	0.074	0.059
Debt to equity ratio	0.016	0.615	0.008	0.002	0.073	0.018	0	0	0	0	0	0.013	0.032
Debt coverage ratio	0.017	0.615	0	0.001	0.09	0.02	0	0	0	0	0.004	0.002	0.023
Current debt to equity	0.014	0.615	0.01	0.002	0.06	0.015	0	0	0	0	0	0.018	0.037
Net working capital	0.032	0.421	0.006	0.027	0.059	0.024	0	0	0	0	0.128	0.011	0.061
Assess the loan usefulness	0.016	0.615	0.008	0.002	0.058	0.018	0	0	0	0	0	0.013	0.032
Stock cumulative profit	0.005	0.366	0.041	0.002	0.045	0.008	0.024	0.019	0.024	0.073	0	0.052	0.088
P/*E*	0.015	0.421	0.028	0.011	0.018	0.022	0.033	0.027	0.033	0.083	0.006	0.025	0
Liquidity ratio	0	0.615	0.007	0.007	0.025	0	0	0	0	0	0.001	0.012	0.02
Working capital return percentage	0.023	0.377	0.021	0.004	0	0.04	0.015	0.012	0.015	0.046	0.001	0.03	0.068
Total income growth %	0.002	0.319	0.027	0.012	0.094	0.003	0.017	0.015	0.017	0.064	0.001	0.02	0.049
Current ratio	0	0.338	0.026	0.006	0.033	0	0.017	0.014	0.017	0.06	0	0.038	0.067
Current assets ratio	0.034	0.371	0.016	0.01	0.04	0.036	0	0	0	0	0.025	0.022	0.052
Debt to total assets ratio	0.034	0.371	0.016	0.009	0.024	0.037	0	0	0	0	0.025	0.022	0.041
Current assets turn over	0.009	0.347	0.024	0.003	0.026	0.015	0.016	0.013	0.016	0.051	0	0.034	0.019
Total asset turn over	0.016	0.338	0.009	0.03	0.068	0.019	0	0	0	0	0.027	0.015	0.013
EPS coverage percent	0.008	0.328	0.016	0.014	0.061	0.016	0	0	0	0	0.003	0.024	0.063
Fixed assets turn over	0.015	0.358	0.009	0.019	0.011	0.025	0	0	0	0	0.001	0.013	0.064
Total predicted income	0.018	0.319	0.011	0.034	0.024	0.034	0	0	0	0	0.013	0.007	0.053
Fixed assets return	0.015	0.358	0.008	0.019	0.011	0.025	0	0	0	0	0.001	0.013	0.054
Average payment	0.012	0.338	0.004	0.005	0.016	0.018	0	0	0	0	0.001	0.008	0.033
Capital	0.012	0	0	0.01	0.073	0.019	0	0	0	0	0.007	0	0.051

Table 10
Selected features for risk and real return parameters.

Selected features of the first and second cluster, based on function clustering method for risk parameter	Return, beta coefficient, efficiency, market return, EPS prediction, percent of growth EPS, DPS, *P*/*E*, EPS, equity ratio, stock book value, debt to total assets ratio, predicted profit margin, *P*/*S*, total incomes growth
Selected features of the first and second cluster, based on function clustering method for real return parameter	Return, market return, beta coefficient, return on asset (ROA), percent of growth EPS, EPS, predicted profit margin, EPS coverage percent

on all feature". If deviation is positive, it means that, use of important feature will improve the prediction results and vice versa.

As results show from Table 11, by this hybrid method we can get better prediction in some methods with fewer numbers of features.

4. Discussion

4.1. The real return results in prediction with selected features

If for denser structure trees all effective features in first prediction are selected by the proposed hybrid model, results in better accuracy, such as "BF Tree", "LAD Tree", and "FT Tree". Otherwise, it is possible that accuracy drops, like "CART and Rep" TREEs. The selected features have different effect on the accuracy of forecasting. Some trees with large structure, such as J48 Graph and J48 Tree are get lower accuracy, while some get a higher accuracy such as ID3 Numerical. Higher accuracy of all algorithms is due to the fact that the hybrid feature selection model, as a pruning algorithm, is used to reduce over training error. Bays algorithms for both real return and risk obtained weak prediction. Thus, DTNB and also NB Tree output for both real return and risk achieved lower accuracy. The results show that the rule base algorithms with average number of rules, such as Part Rule, decision table and Rule Induction, obtain better results. On the other hand, the accuracy of the algorithms with fewer rules like J Rip Rule has descended. The accuracy of the neural network for each output has increased, because of not getting stuck in local optimum points.

Table 11
Algorithms deviation.

Algorithm	Risk accuracy deviation	Return accuracy deviation
LAD Tree	%2 (80.24%)	%1 (79%)
Cart Decision Tree	%1.5 (66.5%)	%−4.5 (72%)
DTNB rule	%−0.9 (76.51%)	%−1 (75%)
Decision table	%−1.2 (75.37%)	%0.07 (76.20%)
BF Tree	%−2 (74.15%)	1.5 (76%)
J Rip Rule	%0 (74.90%)	%−1.2 (73.7%)
J 48 Graph	%−1.83 (71.81%)	%−2 (69.50%)
Part Rule	%1.91 (75.55%)	%2 (74.6%)
Rep Tree	%0.77 (73.52%)	%−2 (73%)
Rule Induction	%0.95 (72.50%)	%1.5 (70%)
J 48 Tree	%−1.5 (70. 05%)	%−0.50 (66.89%)
FT Tree	%2 (69.18%)	%2 (70.5%)
NB Tree	%−4.20 (62.75%)	%−1 (70%)
Neural Net (MLP)	%2.00 (61.00%)	%2.5 (71.5%)
ID3 Numerical	%2.5 (59.5%)	%2 (63.5%)
Bays	%−1.40 (54.15%)	%−2.00 (58.00%)

4.2. The risk results in prediction with selected features

Due to the large number of features extracted from the hybrid feature selection algorithm for risk, the moderate size tree, such as Rep Tree, FT Tree, and LAD Tree have better accuracy than before. However, BF Tree accuracy has been decreased because of removing 2 effective attribute.

With this analysis it is also clear that the algorithms, such as LAD Tree, which use the features beta coefficient, market return, *P*/*E*, and the efficiencies, obtained the highest accuracy which has improved up to 80.24%. Large trees such as J 48 Tree and J 48 Graph get lower prediction accuracy but ID3 Numerical results are improved. As said before, the prediction result of bays based algorithms like DTNB and NB Tree have been decreased but the large drop in NB Tree prediction is because of its dense structure.

For other rule base algorithms the prediction result have been improved or remained stable, except decision table algorithm. This is derived from the average number of rules that are covered by the selected features.

Moreover, by using weight of features obtained from Chi-2 or IG Ratio or Info Gain algorithms (without using the hybrid model), the return and market returns features to predict real return get the highest weight (90% of cumulated weight). Maybe, the high percentages predicted by BF Tree and LAD Tree algorithms are due to these two features. Also to risk parameter, return, beta coefficient, and efficiency features get the highest weight (90% of cumulated weight). Thus the high accuracy of LAD Tree, FT Tree, Rep Tree, and Rule Induction algorithms could be due to these three features. Because these algorithms emphasize these high weight features more than others.

A comparison between our method and similar researches is illustrated in Table 12. Six hybrid methods which have a brilliant accuracy in return forecasting in different country stock exchange compared based on input data, base classifier, feature selection, hybrid prediction model and the best accuracy as follow:

We also exerted data dimension reduction methods including: Principle Component Analyses (PCA), Independent Component Analysis (ICA), Factor Analysis (FA), Discrete Wavelet Transform (DWT), and Discrete Fourier Transform (DFT) methods on data set. Our results on this methods show that despite of long runtime the accuracy of prediction algorithms highly decreased. As an instance, after the reduction of dimensionality from 44 to 11 with PCA algorithm, the LAD Tree prediction results get 52.7% which is a very low accuracy.

Moreover, the data reduction process time in this data is very high and as an instance, based on Rapid Miner Software it takes 11 h and 32 min in DWT algorithms. Although by using MATLAB algorithm, the execution time is less than before, but the accuracy of the results will not differ much. Among these 5 algorithms, ICA results despite of long execution time (approximately 31 h with Rapid Miner Software) obtain better prediction accuracy and the accuracy predicted based on LAD Tree is 71%.

5. Conclusions

In this study, an approach for simultaneous prediction of risk and real return were developed by applying data mining technique

Table 12
Comparison results with other studies.

Author /year	Stock exchange	Input data	Base classifier	Feature selection	Hybrid model	The best accuracy %
Tsai et al. (2011)	Electronic Industry in Taiwan	19 financial ratios and 11 macroeconomic indicators	MLP–Cart–logistic regression	–	Bagging–Voting	66.67
Huang (2012)	30 special companies in Taiwan	14 financial ratios	SVR–GA	–	–	85–76.71
Cheng et al. (2010)	Taiwan	10 technical indexes and 8 macroeconomic indicators	PNN–C4.5–rough Set	–	Hybrid	76
Huang et al. (2008)	South-Korea and Taiwan	23 technical indexes	SVM–K–NN–Cart–logistic regression–back propagation	Wrapper	Voting	76.06
Tsai and Hsiao (2010)	Taiwan	8 fundamental index and 11 macroeconomic indicators	–	GA–PCA–Cart	Back propagation	80.28 79
Tsai, Lu, and Yen (2012)	Taiwan	61 intangible assets value variable	MLP	PCA–stepwise regression–Decision Trees–association rules–GA	MLP	75
Zhang et al. (2014)	Shanghai stock exchanges	50 financial and fundamental feature	NB–SVM–J48–LR–NN	Casual feature selection	–	55
Recent work	Return forecasting in TSE-Iran	44 financial ratios and fundamental index	Cart, Rep Tree, LAD Tree,...	Function based Clustering	Hybrid	80.24
Recent work	Risk forecasting in TSE-Iran	44 financial ratios and fundamental index	DTNB, BF Tree, LAD Tree,...	Function based clustering	Hybrid	79.01

as well as fundamental data set. To do this, first through a comprehensive study, the features which can be potentially effective on risk and return were investigated. Then, after developing an appropriate database the preprocessing of database step was taken. To predict the real return and risk, 20 and 15 different prediction algorithms were applied respectively. Then, the strength and weakness of each one was investigated by analyzing the size and leaves of tree algorithms or/and "If-Then Rules" gains of rule based algorithms. In the next step, by using hybrid feature selection algorithm on the basis of 9 different filter algorithms and function-based clustering method, important features were selected and re-prediction with selected features was performed. The results show that for real return parameter, the number of effective features are usually less than the number of effective features on risk parameter. With the help of these features, the results in most algorithms were improved. In this way, this hybrid feature selection method is capable of identify effective features. The high accuracy of prediction results indicates that the extracted features explain the behavior of market very well and can be considered as a suitable database for the future research. Our findings can enable the investors to analyze the market and gain high accurate results with fewer features, and not getting confused in the market by many features which are not necessarily effective. This study is differed from the previous ones by considering the combination of 9 different feature selection algorithms with function-based clustering algorithm. This hybrid model can enjoy the advantages of all feature selection algorithms and make a robust and accurate decision. 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. While almost none of the relevant studies in this field pay attentions to prediction of risk feature. Furthermore, we design 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 arbitrary choosing likely effective features.

Finally, investigating each algorithm with a feature-oriented view point indicates the factors which cause strength and weakness of that algorithm. Therefore, by searching about property of data base, we can choose a proper algorithm without implementation of all methods. This idea can be further extended not only in quantitative investment, but also in other field of studies where expert systems and machine learning techniques are used.

The limitation of this method is that collecting all data and information may be difficult for some real cases.

Future research directions of paper include but are not limited to

1. Combining prediction methods in the framework of fusion models or optimize the classification algorithms by applying some metaheuristics algorithms to improve the prediction results.
2. Predicting the other important variable (in addition to risk and return) such as liquidity (Barak et al., 2013).
3. Using technical features and textual information, in addition to fundamentals features, in order to have a more comprehensive features and to be able to predict short term situation of stocks.
4. Customizing the proposed approach for the prediction of risk and return in a particular industry or investigating the accuracy of the procedure by data from other popular stock markets, such as US stock market which may result in new dimensions in this procedure.
5. Applying different clustering models to our feature selection data set and compare results by considering new feature selection methods, such as CFS (Zhang, Hu, Xie, Wang, et al., 2014), density based clustering (Shamshirband et al., 2014)or entropy-based clustering for feature selection (Lin, 2013).

Appendix A. CAPM model

- β coefficient is the amount of changes in the stock return to market and accounted as follow.

$$\beta = (Co\,v\,(\text{Market return}^{*}\text{stock return}))/Var\,(\text{Market return}). \tag{A$_1$}$$

- Market expected return, shows the amount of market return in a definite time which is gained with this formula:

$$r_m = \frac{P_t - P_{t-1}}{P_{t-1}} \tag{A$_2$}$$

In this p_t = the market indicator at the end of period (for example 2013/12/28) and P_{t-1} = the market indicator at the beginning of period (for example 2013/1/1).

- Return without risk also can be done through this formula.

$$r_{ft} = \frac{(p_t - p_{t-1}) + D_t}{p_{t-1}} \times 100 \tag{A$_3$}$$

In this p_t is end of period stock price, P_{t-1} = beginning of period stock price and D_t = benefits of stock ownership which has belonged to shareholder in period t. if we have capital increase in period of investment from savings or receivables and cash income then the formula will change as a follow:

$$r_{ft} = \frac{D_t + p_t(1+\alpha+\beta) - (p_{t-1} + c\alpha)}{p_{t-1} + c\alpha} \times 100 \tag{A$_4$}$$

In this α = the percent of capital increase of the receivables and cash income, β = the percent of capital increase of savings, c = nominal amount paid by investor to increase the capital of the receivables and cash income. We use this formula for calculate the r_m, r_f and β.

References

Araújo, R. d. A., & Ferreira, T. A. (2013). A morphological-rank-linear evolutionary method for stock market prediction. *Information Sciences, 237*, 3–17.

Barak, S., Abessi, M., & Modarres, M. (2013). Fuzzy turnover rate chance constraints portfolio model. *European Journal of Operational Research, 228*, 141–147.

Bartholdy, J., & Peare, P. (2005). Estimation of expected return: CAPM vs. Fama and French. *International Review of Financial Analysis, 14*, 407–427.

Bauer, R., Guenster, N., & Otten, R. (2004). Empirical evidence on corporate governance in Europe: The effect on stock returns, firm value and performance. *Journal of Asset Management, 5*, 91–104.

Bernstein, L., & Wild, J. (1999). *Analysis of financial statements* (5th ed.). McGraw-Hill.

Brealey, R. A., Myers, S. C., & Allen, F. (2007). *Principles of corporate finance* (9th ed.). McGraw-Hill.

Breunig, M. M., Kriegel, H-. P., Ng, R. T., & Sande, J. (2000). LOF: Identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on management of data* (pp. 93–104). Publishing, Dallas.

Cao, Q., Leggio, K. B., & Schniederjans, M. J. (2005). A comparison between Fama and French's model and artificial neural networks in predicting the Chinese stock market. *Computers & Operations Research, 32*, 2499–2512.

Carnes, T. A., & College, B. (2006). Unexpected changes in quarterly financial-statement line items and their relationship to stock prices. *Academy of Accounting and Financial Studies Journal, 10*, 99–116.

Chang, T.-S. (2011). A comparative study of artificial neural networks, and decision trees for digital game content stocks price prediction. *Expert Systems with Applications, 38*, 14846–14851.

Chen, Y.-S., & Cheng, C.-H. (2012). A soft-computing based rough sets classifier for classifying IPO returns in the financial markets. *Applied Soft Computing, 12*, 462–475.

Cheng, J.-H., Chen, H.-P., & Lin, Y.-M. (2010). A hybrid forecast marketing timing model based on probabilistic neural network, rough set and C4.5. *Expert Systems with Applications, 37*, 1814–1820.

Dastgir, M., & Afshari, M. H. (2004). Evaluating stock pricing models on Tehran stock exchange assets. *Accounting Studies, 3*, 60–94.

de Oliveira, F. A., Nobre, C. N., & Zárate, L. E. (2013). Applying artificial neural networks to prediction of stock price and improvement of the directional prediction index – case study of PETR4, Petrobras, Brazil. *Expert Systems with Applications, 40*, 7596–7606.

duda, R. o., Hart, P. E., & Strok, D. G. (2001). *Pattern classification*. Wiley.

Dumais, S., Platt, J., Heckerman, D., & Sahami, M. (1998). Inductive learning algorithms and representations for text categorization. In *Proceedings of the international conference on information knowledge management* (pp. 148–155).

Enke, D., & Thawornwong, S. (2005). The use of data mining and neural networks for forecasting stock market returns. *Expert Systems with Applications, 29*, 927–940.

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics, 33*, 3–56.

Fama, E. F., & French, K. R. (2012). Size, value, and momentum in international stock returns. *Journal of Financial Economics, 105*, 457–472.

Frank, E., & Witten, I. H. (1998). Generating accurate rule sets without global optimization. In *Fifteenth international conference on machine learning* (pp. 144–151).

Gordon, M. J. (1982). *The investment, financing, and valuation of the corporation* (vol. 52). Greenwood Press Reprint.

Hall, M. (1998). Correlation-based feature subset selection for machine learning. University of Waikato.

Hall, M., & Frank, E. (2008). Combining Naive Bayes and decision tables. In *Proceedings of the 21st Florida artificial intelligence society conference (FLAIRS)*. Florida.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter, 11*, 10–18.

Han, J., & Kamber, M. (2006). *Data mining: Concepts and techniques*. The Morgan Kaufmann.

Hjalmarsson, E. (2010). Predicting global stock returns. *Journal of Financial and Quantitative Analysis, 45*, 49–80.

Hofmann, M., & Klinkenberg, R. (2013). *RapidMiner: Data mining use cases and business analytics applications*. Chapman and Hall/CRC.

Holte, R. C. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning, 11*, 63–90.

Hong, T.-P., & Wu, C.-W. (2011). Mining rules from an incomplete dataset with a high missing rate. *Expert Systems with Applications, 38*, 3931–3936.

Huang, C.-F. (2012a). A hybrid stock selection model using genetic algorithms and support vector regression. *Applied Soft Computing, 12*, 807–818.

Huang, C. F. (2012b). A hybrid stock selection model using genetic algorithms and support vector regression. *Applied Soft Computing, 12*, 807–818.

Huang, C.-J., Yang, D.-X., & Chuang, Y.-T. (2008). Application of wrapper approach and composite classifier to the stock trend prediction. *Expert Systems with Applications, 34*, 2870–2878.

Kao, L.-J., Chiu, C.-C., Lu, C.-J., & Chang, C.-H. (2013). A hybrid approach by integrating wavelet-based feature extraction with MARS and SVR for stock index forecasting. *Decision Support Systems, 54*, 1228–1244.

Kaplan, S. N., & Ruback, R. S. (1995). The valuation of cash flow forecasts: An empirical analysis. *The Journal of Finance, 50*, 1059–1093.

Knorr, E. M., & Ng, R. T. (1999). Finding intentional knowledge of distance-based outliers. In *Proceedings of the 25th international conference on very large data bases* (pp. 211–222). Edinburgh, Scotland.

Kononenko, I. (1994). Estimating attributes: analysis and extensions of relief. In *Proceedings of the seventh European conference on machine learning* (pp. 171–182).

Lai, R. K., Fan, C.-Y., Huang, W.-H., & Chang, P.-C. (2009). Evolving and clustering fuzzy decision tree for financial time series data forecasting. *Expert Systems with Applications, 36*, 3761–3773.

Lee, W.-S., Tzeng, G.-H., Guan, J.-L., Chien, K.-T., & Huang, J.-M. (2009). Combined MCDM techniques for exploring stock selection based on Gordon model. *Expert Systems with Applications, 36*, 6421–6430.

Levin, N., & Zahavi, J. (2001). Predictive modeling using segmentation. *Journal of Interactive Marketing, 15*, 2–22.

Lewellen, J. (2004). Predicting returns with financial ratios. *Journal of Financial Economics, 74*, 209–235.

Li, B. (2006a). A new approach to cluster analysis: The clustering-function-based method. *Journal of the Royal Statistical Society, Series B (Statistical Methodology), 68*, 457–476.

Li, B. (2006b). Sign eigenanalysis and its applications to optimization problems and robust statistics. *Computational Statistics & Data Analysis, 50*, 154–162.

Lin, H.-Y. (2013). Feature selection based on cluster and variability analyses for ordinal multi-class classification problems. *Knowledge-Based Systems, 37*, 94–104.

Liu, H., & Setiono, R. (1996). A probabilistic approach to feature selection – a filter solution. In *Proceedings of the 13th international conference on machine learning* (pp. 319–327).

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.

Mukherji, S., Dhatt, M. S., & Kim, Y. H. (1997). A fundamental analysis of Korean stock returns. *Financial Analysts Journal, 53*, 75–80.

Ng, W. W., Liang, X.-L., Li, J., Yeung, D. S., & Chan, P. P. (2014). LG-trader: Stock trading decision support based on feature selection by weighted localized generalization error model. *Neurocomputing*.

Omran, M., & Ragab, A. (2004). Linear versus non-linear relationships between financial ratios and stock returns: Empirical evidence from Egyptian firms. *Review of Accounting and Finance, 3*, 84–102.

Ou, P., & Wang, H. (2009). Prediction of stock market index movement by ten data mining techniques. *Modern Applied Science, 3*, 28–42.

Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2014). Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert Systems with Applications*.

Sadka, G., & Sadka, R. (2009). Predictability and the earnings–returns relation. *Journal of Financial Economics, 94*, 87–106.

Soliman, M.-T. (2008). The use of DuPont analysis by market participants. *The Accounting Review, 83*, 823–853.

Shamshirband, S., Amini, A., Anuar, N. B., Kiah, M. L. M., Wah, T. Y., & Furnell, S. (2014). D-FICCA: A density-based fuzzy imperialist competitive clustering algorithm for intrusion detection in wireless sensor networks. *Measurement*.

Shi, H. (2007). Best-first decision tree learning. *Annals of statistics, 2*, 337–407.

Svalina, I., Galzina, V., Lujić, R., & Šimunović, G. (2013). An adaptive network-based fuzzy inference system (ANFIS) for the forecasting: The case of close price indices. *Expert Systems with Applications, 40*, 6055–6063.

Tsai, C.-F., & Hsiao, Y.-C. (2010). Combining multiple feature selection methods for stock prediction: Union, intersection, and multi-intersection approaches. *Decision Support Systems, 50*, 258–269.

Tsai, C.-F., Lin, Y.-C., Yen, D. C., & Chen, Y.-M. (2011). Predicting stock returns by classifier ensembles. *Applied Soft Computing, 11*, 2452–2459.

Tsai, C.-F., Lu, Y.-H., & Yen, D. C. (2012). Determinants of intangible assets value: The data mining approach. *Knowledge-Based Systems, 31*, 67–77.

Wang, J.-L., & Chan, S.-H. (2006). Stock market trading rule discovery using two-layer bias decision tree. *Expert Systems with Applications, 30*, 605–611.

Witten, I. H., & Frank, E. (2011). *Data mining: Practical machine learning tools and techniques* (3rd ed.). San Francisco: Morgan Kaufmann.

Wu, J.-L., Yu, L.-C., & Chang, P.-C. (2014). An intelligent stock trading system using comprehensive features. *Applied Soft Computing, 23*, 39–50.

Yu, H., Chen, R., & Zhang, G. (2014). A SVM stock selection model within PCA. *Procedia Computer Science, 31*, 406–412.

Zhang, X., Hu, Y., Xie, K., Wang, S., Ngai, E., & Liu, M. (2014). A causal feature selection algorithm for stock prediction modeling. *Neurocomputing*.

Zhang, X., Hu, Y., Xie, K., Zhang, W., Su, L., & Liu, M. (2014). An evolutionary trend reversion model for stock trading rule discovery. *Knowledge-Based Systems*.