Applied Soft Computing 22 (2014) 205–212

Contents lists available at ScienceDirect

Applied Soft Computing

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

Short term stock selection with case-based reasoning technique

Huseyin Ince*

Faculty of Business Administration, Gebze Institute of Technology, Cayırova Fab. Yolu No:101 P.K:141, 41400 Gebze, Kocaeli, Turkey

ARTICLE INFO

Article history:
Received 21 June 2011
Received in revised form
27 September 2013
Accepted 17 May 2014
Available online 27 May 2014

Keywords:
Stock selection
Case-based reasoning
Intelligent system
Computational intelligence
Genetic algorithms
Earning analysis

ABSTRACT

Stock selection is an important decision making problem. Trading strategies and rules based on fundamental and technical analysis can be used for decision making process. In this paper, we propose an intelligent stock selection method, which is called case-based reasoning (CBR). This technique uses the fundamental and technical indicators to identify the winning stocks around the earning announcements. CBR method is compared with other artificial intelligence techniques such as multi layer perceptron (MLP), decision trees (QUEST, Classification and Regression Trees, C5), generalized rule induction (GRI) and logistic regression. We show that the performance of CBR is better than the performance of other techniques in terms of classification accuracy, average return, Sharpe ratio and ideal profit.

1. Introduction

Stock selection is regarded as a challenging task for portfolio optimization. With the growing importance in the role of equities, the selection of attractive stocks for the short and long term investment has been the most important decision. Therefore, a reliable tool in the selection process can be of great assistance to investors. An effective and efficient system gives investors the competitive edge over others as they can identify the performing stocks with minimum effort. Trading strategies and rules based on fundamental and technical analysis have been used by both academics and practitioners for decision making process. Trading strategies can be transformed to computer language to exploit the logical processing power of the computer. This greatly reduces the time and effort to find attractive stocks.

Many papers have focused on fundamental indicators to understand how they affect future earnings and stock prices [8,64,23,18,60]. Developing investment strategy based on fundamental indicators result in significant abnormal returns. In addition to this, analysts' recommendations, stock market rumors and earning surprises can lead to abnormal returns. Many studies find that stock prices respond positively to the announcements of increase in earnings and negatively to the announcements of decrease in earnings for the U.S. firms [59,29,10,16]. Some researchers prove evidence of the informational content of earning announcements in a number of non U.S. markets [5,15]. Based on these findings, we can develop a model for selecting winning stocks around the earning announcements.

There have been many studies using machine learning (ML) techniques in stock selection and stock price prediction. Most of these studies have focused on stock market index and individual stock prediction [7,29,28,6,17,13,14]. Recent studies have presented encouraging results on stock selection using data mining techniques such as rule induction, neural network, and combination of classifiers [20,24,25,58,4,31,27,14,34].

CBR technique is one of the popular methodologies in knowledge-based systems. It is a novel paradigm that solves a new problem by recalling and reusing specific knowledge from past experience [1]. Concurrently, it is already an established and powerful methodology for intelligent problem solving and has been used for developing a variety of applications. Due to its strengths, researchers have successfully applied CBR to many areas: supply chain management and scheduling [44,50,45], bond rating [54,33], business failure prediction [43], business control system development [11], bankruptcy prediction and credit analysis [3,49,32], and stock market prediction [20,21,36].

This study proposes a short term stock selection model based on CBR. A combination of technical and fundamental indicators is used as an input the CBR model for effective stock selection around the earning announcement. Portfolio managers focus on long-term portfolio management. Therefore, they try to choose fundamentally strong stocks with low volatility. On the other hand, some investors,

* Tel.: +90 5333011279.
E-mail addresses: ince@yahoo.com, h.ince@gyte.edu.tr

http://dx.doi.org/10.1016/j.asoc.2014.05.017
1568-4946/

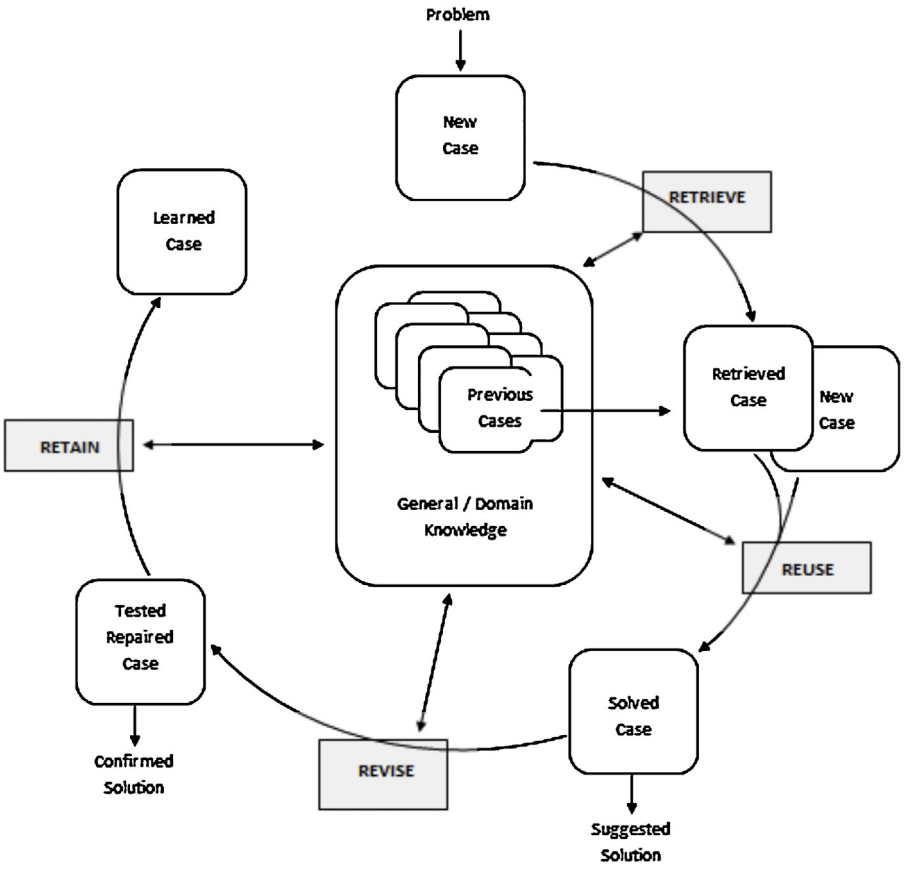

Fig. 1. A CBR cycle.

Source: Adopted from Aamodt and Plaza [1].

especially small investors, focus on short term stock selection. They try to develop investment strategy based on fundamental and technical analyses. We propose an intelligent stock selection model around the earning announcement periods that uses combination of fundamental and technical indicators. There are some researches that use the CBR approach to forecast the stock indices and direction of movement [37,35]. As far as we know, there are no researches applying the CBR technique to short term stock selection or stock classification problem. The primary contribution of this study is to show that CBR can be used for short term stock selection. Another contribution is that we combine the fundamental and technical indicators.

The rest of this paper is organized as follows: The next section briefly describes the CBR approach. Section 3 presents the proposed stock selection model. Section 4 explains the experimental design and the results of the evaluation experiment. The final section presents the conclusion.

2. Literature review and proposed model

2.1. Case-based reasoning approach

Case-based reasoning (CBR) technique is one of the popular methodologies in knowledge-based systems. While other artificial intelligence techniques depend on generalized relationships between problem descriptors and conclusions, CBR utilizes specific knowledge of previously experienced problem situations. It solves a problem by retrieving, reusing, revising and retaining past cases based on their degree of match and usefulness to the current situation. This is done by partial matching of the past cases with the current case, and by ranking across case dimensions until a smaller set of matching and useful cases is retrieved [3,1]. The usefulness of past cases for the current situation, on the other hand, may be assessed by assigning weights. A high degree of similarity or usefulness presents a good reason for adaptation. CBR methodology has been used in a broad range of domains to capture and organize past experience and to learn how to solve new situations from previous solutions.

In general, a CBR system can be viewed as a composition of two modules, i.e., a case library and a problem solver [3]. The case library, which contains historical problems and their corresponding solutions, acts as a source of knowledge. Given a new problem, the problem solver performs two actions, i.e., (i) retrieves similar cases from the case library based on some similarity measure; and (ii) adapts the retrieved cases so that a solution to the new problem can be proposed.

CBR system is composed of four sequential steps which are called into action each time that a new problem is to be solved [3,1,38]. Conceptually CBR is commonly described by the CBR cycle shown in Fig. 1. It involves four major steps which are recalled every time that a problem needs to be solved [3].

i. Retrieve the most relevant case(s).
ii. Reuse the case(s) to attempt to solve the problem.
iii. Revise the proposed solution if necessary.
iv. Retain the new solution as a part of a new case.

The purpose of the retrieval step is to search the case base to select existing cases sharing significant features with the new case. The key issues in this step are computing case similarity to match

the best case, and adapting a similar solution to fit the new problem. Thus, the success of a CBR system largely depends on an effective retrieval of useful prior cases for the problem. The nearest neighbor method has been widely used for case retrieval. The method involves the assessment of similarities between stored cases and the new input case, based on matching a weighted sum of features. Once one or more cases are identified in the case base as being very similar to the new problem, they are selected for the solution of this particular problem. The CBR system tries to reuse the information and knowledge of the previously retrieved cases for solving the new problem. Once matching cases are retrieved from the case base, they should be adapted to the requirements of the current case. This process is called the revision process for CBR. This solution is revised (if possible) and finally the new case is stored. Cases can also be deleted if they prove to be inaccurate; they can be merged together to create more generalized ones and they can be modified. In the final step, the new solution is retained as part of a new case likely to be useful for future problem solving [61].

Efficiency and accuracy of case retrieval highly depend on the determination of weight for each feature. In many cases, subjective weighting values are given by the user, and thus the retrieved solutions cannot always be guaranteed. Therefore, several case indexing methods have been proposed for effective case retrieval [3]. These are nearest neighbor, induction, fuzzy logic, rough set theory, kernel methods and database technology. Nearest neighbor is the most commonly used case indexing method. It is a direct method that uses a numerical function to compute the degree of similarity. In this study, we use a numeric evaluation function which measures the distance taking into account the importance of features to compute the degree of match in retrieval. A typical numerical function is shown in the following formula [42]:

$$DIS_{ab} = \sqrt{\sum_{i=1}^{n} w_i \cdot (f_{ai} - f_{bi})^2} \tag{1}$$

where DIS is the matching function using Euclidian distance between cases, w_i the weight of the feature i, and n is the number of features.

In nearest-neighbor, every feature in the input case is matched to its corresponding feature in the stored case, and the degree of match of each pair is computed using the matching function given in Eq. (1). Then, based on the importance assigned to each feature, an aggregate match score is computed. Cases are ordered according to their scores [56].

The importance of each field shows us how much attention to pay to the respective match. Although, researchers suggest several ways of assigning the importance values such as knowledge of human expert, statistical evaluation, machine learning techniques and fuzzy logics [56,60,49], it is difficult to tell a priori regarding which set of weights would be the most effective to solve a specific problem. According to Shin and Han [54], one way is to have a human expert assign the importance values as the case library. The expert is expected to have the knowledge and experience required to decide which dimensions make good predictors. Another way to assign importance values is to do a statistical evaluation of known cases to determine which dimensions predict the solutions best. The correlation coefficient between each input and the output in the reference set can be used to weigh each input when computing the distance measure for a new example. Machine learning can be used as an alternative approach to learn the optimal weights from historical cases using evolutionary search technique. By evaluating the fitness of different weight vectors, good solutions can be found for CBR system.

2.2. Genetic algorithms for case based reasoning

Several researchers have suggested using genetic algorithms (GAs) to determine the most appropriate feature weights of CBR approach ([3,47,54]). According to Moon and Sohn [47], the GA is able to improve the search results by constantly examining various possible solutions with the reproduction, crossover and mutation operations.

GAs can be used as an alternative approach to compute optimal weights from old cases. Good solutions can be found by evaluating the fitness of different weight vectors. GAs explore a complex space in an adaptive way, guided by biological evolution mechanisms of reproduction, crossover and mutation to generate a new population of problem solutions and select the best solution for the problem. More information about GAs can be found in ([3,47,19]). We need to specify the parameters and their adjustable ranges, potential constraints and objective or fitness function to evaluate the performance. The parameters that are coded binary represent the weight vectors for nearest neighbor matching. We do not use any constraints for this problem. Defining the fitness function is the most critical step. The objective of GAs is to determine the set of weighting values that can best formalize the match between the input case and the previously stored cases. In other words, our objective is to retrieve more relevant cases that can lead to the correct solution. This can be achieved by increasing the classification accuracy. In this study, the fitness function is defined as the classification accuracy of the training set. The fitness function is expressed as:

$$\begin{aligned} &\max \quad CR = \frac{1}{n}\sum_{i=1}^{n} CA_i \\ &s.t. \quad CA_i = 1 \quad \text{if} \quad O(T_i) = O(S_{j*(i)}), \\ &CA_i = 0, \quad \text{otherwise}, \\ &S_{j*(i)} = \min_{j \in R} \left(\sqrt{\sum_{k=1}^{l} w_k (T_{ik} - R_{jk})^2} \right), \end{aligned} \tag{2}$$

for a given i (i = 1,2,. . .n). CR is the classification accuracy rate of the test set; CA_i the classification of the ith case of the test set denoted by 1 and 0; $O(T_i)$ the target output of ith case of the training set; $O(S_{j^*(i)})$ the output of jth case of reference set that has the minimum distance with the ith case of the training set; $S_{j(i)}$ the distance between ith case of the training set and jth case of the reference set; T_{ik} the kth feature of the ith case of the training set(T); R_{jk} kth feature of the jth case of the reference set(R); w_k the importance weight of the kth feature of case; l the number of features and n the number of the training cases [53].

The key parameters consisting of population size, crossover rate, mutation rate and the stopping criteria have to be defined first when developing the algorithm. The population size is determined according to size of the problem. The common view is that a larger population takes longer to settle on a solution, but is more likely to find a global optimum because of its more diverse gene pool. The crossover and mutation rates prevent the output from falling into the local optima. The crossover rate ranges between 0.5 and 0.8 and the mutation rate ranges between 0.06 and 0.1 for this experiment. As a stopping condition, we use 5000 trials.

The structure of the algorithm is shown in Fig. 2. The hybrid GA-CBR algorithm is given as follows:

Step 1: We search an optimal weight vector with precedent cases for which the classification output is determined.
Step 2: The weight vector obtained in step 1 is applied to the case indexing scheme for the retrieval process. We also evaluate

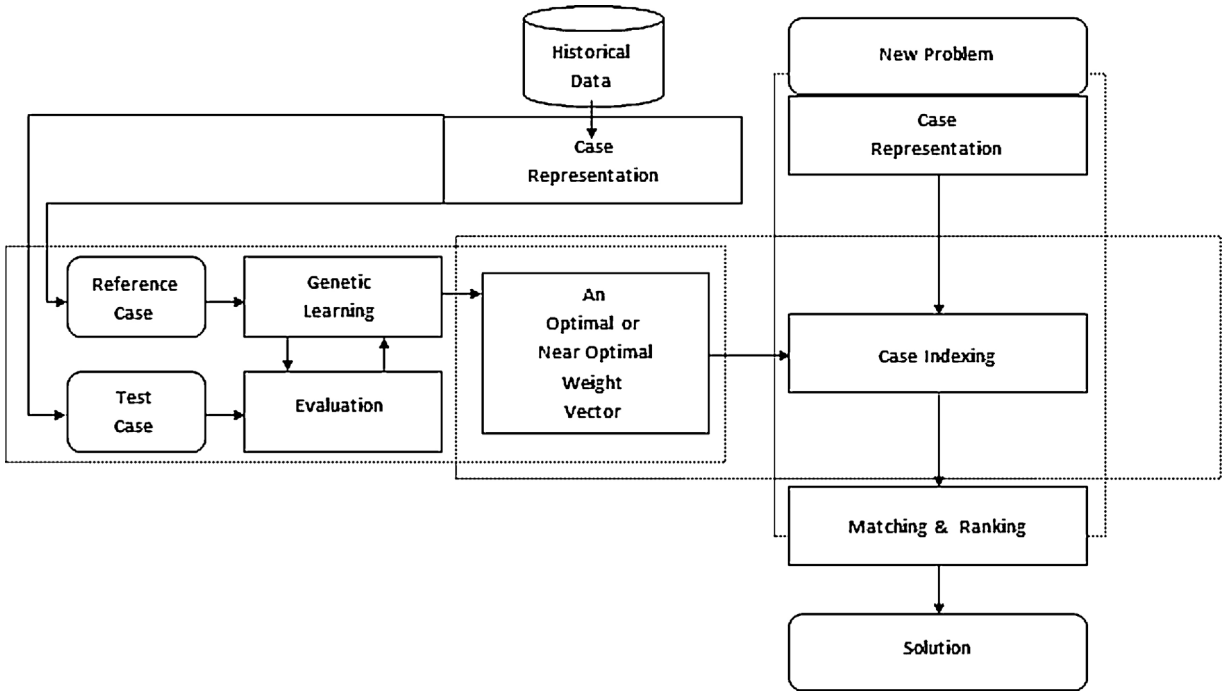

Fig. 2. Hybrid structure of a GA–CBR system.

Source: Adopted from [53].

the resulting model with additional validation cases for which the outcome is known. Useful cases are ranked and retrieved by using a weight vector in the nearest neighbor matching function. As the validation cases are not used for parameter optimization, the prediction performance tested by these cases would be the closest to the current or future cases. If the project is successful, this leads to production.

Step 3: In step 3, new data are presented for the model to solve the problem.

The weakness of the CBR systems is that there is some human interaction. This is a current weakness of CBR systems and one of their major challenges. Corchado and Aiken [22] have proposed a method of automating the process of case adaptation for the solution of problems in which the cases are characterized predominantly by numerical information. In this study, Corchado and Aiken's [22] approach is utilized.

3. Stock selection model based on fundamental and technical indicators

Investors are usually faced with an enormous amount of stocks in the market. A crucial part of their decision process is the selection of stocks to invest in. The identification of winning stocks remains to be one of the major problems for investors in making effective decisions. There exists an immense body of work on the mathematical analysis regarding the behavior of stock prices, stock markets and successful strategies for stock selection. In recent times, a variety of different approaches have been tried for identifying stocks with higher returns. High return (loss) occurs with new information such as earning announcements, company news, and dividend announcements. Market reaction to earning surprises has been investigated by several researchers (see [39,52] for a recent review). They show that there is a positive association between earning surprises and abnormal returns around the earning announcements. Stock return volatilities generally occur around earning announcements [59].

Two approaches, technical and fundamental analysis, are commonly used by academicians and market professionals for stock selection and predicting stock prices [8,9,59,26,57]. Fundamental analysis involves using the financial and related data (fundamentals) of a firm to determine stock value and forecast future stock price movements. Fundamental analysts believe that an investment instrument has its intrinsic value that can be derived from the behavior and performance of its company. The fundamental approach utilizes quantitative tools, mainly the financial ratios compiled from financial statements as well as qualitative indicators, such as management policy, marketing strategy, and product innovation, to determine the value of an investment instrument [40]. Technical analysis uses knowledge from the past behavior of a stock price series to predict the future. The technical approach tries to identify turning points, momentum, levels, and directions of an investment instrument, using tools such as charting, relative strength index, moving averages, on balance volume, momentum and rate of change, breadth advance decline indicator, directional movement indicator, and de-trended price oscillator [46,41].

A combination of technical and fundamental indicators can be used for effective stock selection. Portfolio managers use different techniques to identify which stocks to add to their portfolio. Generally, they focus on long-term portfolio management. Therefore, they try to choose fundamentally strong stocks with low volatility. On the other hand, some investors, especially small investors, focus on short term stock selection. They try to develop investment strategy based on fundamental and technical analyses. We propose an intelligent stock selection model around the earning announcement periods that uses combination of fundamental and technical indicators.

Based on the previous studies [8,30,40,42,57], we select 12 fundamental and technical indicators as predictor variables. These are the last four quarters' earnings per share (eps), insider buying and selling, institutional buying and selling, 10-day moving average (MA), 50-day moving average, 10-day moving average over 50-day moving average and current quarter eps estimate. Technical indicators are defined as categorical variables. For example, the 10-day

MA is defined as "Bull (1)" if the closing price is higher than 10-day MA value; otherwise this variable is defined as "Bear(−1)". Other indicators are also categorized in the same manner.

For each quarter we assign a class to every stock to indicate its performance. Since different companies have different report cycles, the stock returns are calculated individually using the price data for the five days following the date of the quarterly earning announcement. All price data is obtained from the Yahoo finance and the adjusted stock price is used. If the performance of a stock is greater or equal to 5% then, it is labeled as exceptional high return stock (Class +**1**) and the others are labeled as unexceptional return (Class −**1**).

This stock selection problem is then formulated as a two-class pattern recognition task. We represent the fundamental and technical indicators for the *i*th firm as a vector of predictor variables $\mathbf{x_i} = (x_1, x_2, \ldots, x_n)$ (for our case $n = 12$). The expected future return of the stock will be a binary dependent variable $y_i = \pm 1$, where +1 represents exceptional high return stocks, and −1 as normal stocks. Therefore a training set ($\mathbf{x}$, y) of l firms will be the following pairs:

$$\{(x_1, y_1), (x_2, y_2), \ldots (x_l, y_l)\} \subset \Re^n x \pm 1 \quad (3)$$

In this study, we present stock selection as a two-class classification problem. This problem can be solved by using case based reasoning approach.

4. Research data and experiments

First, we explain multilayer perceptron (MLP), decision trees (classification and regression trees (CART), QUEST, C5), logistic regression and generalized rule induction (GRI).

4.1. Multi-layer perceptron

Multilayer perceptron (MLP) is the commonly used technique for classification and prediction. The MLP also known as backpropagation neural network (BPN) is feed forward neural network trained with the standard back propagation algorithm. Backpropagation algorithm uses the gradient steepest descent method to minimize the total square error of the output computed by the net. A multilayer perceptron is made up of several layers of neurons. Each layer is fully connected to the next one. The MLP in this paper was a three-layer, single-output network with twelve input nodes, 10 hidden-layer nodes and one output node. Varying the number of hidden nodes from 4 to 15, it was determined that using ten hidden nodes in the BPN gives the best performance using the criteria given by Tay and Cao [57]. The selection of the learning rate as 0.01 and the momentum term as 0.9 is because a BPN with these settings as the learning parameters has the best prediction performance with the least number of epochs [57]. The sigmoid function was used at each of the hidden layer nodes and the output node. The number of training iterations was set to 5000.

4.2. Decision tree algorithms

Decision trees form an integral part of 'machine learning' an important sub-discipline of artificial intelligence. Almost all the decision tree algorithms are used for solving classification problems. Decision trees are becoming increasingly more popular for data mining because they are easy to understand and interpret, require little data preparation, handle numerical and categorical data, and they perform very well with a large data set in a short time. Decision tree algorithms induce a binary tree on a given training data, resulting in a set of 'if–then' rules. These rules can be used to solve the classification or regression problem. Decision trees produce excellent visualizations of results and their relationships. There are many algorithms of decision trees. Their main difference is the way to decide the sequence of attributes that should be used for each branch node. ID3 [51] is a famous decision tree algorithm, which uses the information gain for the choice of the sequence of attributes. However, a bias may develop when each attribute or variable has different magnitudes. C4.5 and C5 [63] is an extended version of ID3, which uses the gain ratio instead of the information gain to avoid this problem. A decision tree that is able to deal with continuous data is a regression tree such as the classification and regression tree (CART). CART [12] uses the Gini index, which is the sum of the squares of the proportions of the categories, for the choice of the sequence of attributes. The Quick, Unbiased, Efficient Statistical Tree (QUEST) algorithm is a binary-split decision tree algorithm for classification. It is similar to the CART algorithm. However, there are some minor differences. For instance, QUEST employs an unbiased variable selection method, uses imputation for dealing with missing values instead of surrogate splits, and handles categorical variables with many categories. In this study, the C5, CART, and QUEST algorithms are the most commonly used ones.

4.3. Logistics regression

Logistic regression is a widely used statistical modeling technique in which the probability of a dichotomous outcome is related to a set of potential predictor variables in the form:

$$\log\left[\frac{1}{1-p}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i \quad (4)$$

where p is the probability of the outcome of interest, β_0 is the intercept term, and β_i ($i = 1,2,\ldots,n$) represents the β coefficient associated with the corresponding explanatory variable x_i ($i = 1,2,\ldots,n$) [48]. The dependent variable is the logarithm of the odds, $\{\log[p/(1-p)]\}$; which is the logarithm of the ratio of two probabilities of the outcome of interest. These variables are usually selected for inclusion by using some form of backward or forward stepwise regression technique [48]. The maximization of the likelihood function is usually applied as the convergent criterion to estimate the coefficients of corresponding parameters when the logistic regression models are utilized. In this study, the stepwise logistic regression procedure is used in building the model.

4.4. Generalized rule induction (GRI)

An association rule algorithm is capable of producing rules that describe associations (affinities) between attributes of a symbolic target. The association rule shows relationships among items in a transaction of a database. These patterns or rules have been used for various purposes. The GRI induction process can be divided into two steps. First the database is scanned to extract all the itemsets that satisfy a user-specified minimum support criterion. Then each association rule that describes an association between items in a transaction that occur frequently is extracted based on a user-specified minimum confidence criterion [2]. The process usually results in a large number of ARs. In order to reduce the number of rules generated, the minimum confidence threshold value can be increased. However, setting the minimum confidence level too high may prevent the identification of some important ARs. The Generalized Rule Induction (GRI) is one of most commonly used methods for AR induction. The GRI generates rules to summarize patterns in the data using a quantitative measure for the interestingness of rules. This measure provides a method for ranking competing rules and allows the system to constrain the search space for useful rules, as well as identifying the best or most interesting rules describing a database. GRI is based on the ITRule algorithm [55] and extends that algorithm with added functionality.

Table 1
Case attributes of stock selection and descriptive statistics.

Variable name	Range	Mean	Std. deviation
Quarterly EPS change (%)	804.44	17.19	101.94
EPS current quarter	39.75	0.22	1.18
EPS 1 quarter ago	15.43	0.2	0.62
EPS 2 quarter ago	7.56	0.14	0.45
EPS 3 quarter ago	8.66	0.15	0.57
EPS 4 quarter ago	13.72	0.13	0.52
Current quarter EPS estimate	4.62	0.2	0.35
Net insider activities (difference between buying and selling)	84141	−436.13	3721.79
Net Institution activities (difference between buying and selling)	246268	1157.39	9058.18
10-Day moving average[a]	–	–	–
50-Day moving average[a]	–	–	–
10 over 50 day moving average[a]	–	–	–

[a] Categorical variable (1: Bullish, −1: Bearish)

4.5. Experimental results

We examine the stocks that are traded in NASDAQ. The data set contains 1200 observations. As we mentioned before, 12 fundamental and technical indicators are used as predicted variables. Technical indicators, which are 10-day moving average (MA10) and 50-day moving average (MA50) are computed one day before earning announcements and categorized as "Bull" and "Bears" as we explained in section 3. The 10 over 50-day moving average (MA1050) is categorized as "Yes (1)" and "No (−1)". These indicators are used widely in practical. Quarterly fundamental indicators, which are, quarterly eps change, current quarter eps and its estimate, and last four quarter eps are used as inputs (see Table 1). In addition to these variables, net insider and institution activities (difference between buying and selling) are used to reflect the opinion of the insiders and institutions. Stock prices are used to calculate the stock returns around the quarterly earning announcements. Given the stock prices and indicators, it is a prediction problem that involves discovering useful patterns in the dataset and applying that information to classify the stocks.

To evaluate the effectiveness of the CBR approach, we compare the return generated by the selected stock from CBR with other techniques such as multilayer perceptron (MLP), decision trees (classification and regression trees (CART), QUEST, C5), logistic regression and generalized rule induction (GRI). Next, we explain these techniques briefly. Table 2 shows classification rates for all methods for the training and testing data set.

Table 2 provides classification rates for all methods, for the training and testing set. The CBR approach produces better results for training and testing set, with around 70% of stocks correctly classified and about 57% of "High Return" stocks correctly classified. Given that the top 25% of stocks are assigned to the "High Return" class, this is well in excess of what might be expected by chance. Neural Networks (NN), logistic regression and decision trees (C5, QUEST, CART methods) produce comparable results. The performance of general rule induction (GRI) technique is highly lower than the others. It classifies about 49% of the stocks correctly.

For small investors, it is more feasible to buy shares than to sell them short. Therefore, our goal is to identify high-performing (Class +1) stocks reliably. The last column of Table 2 shows the percentage of shares predicted to be high performing (Class +1) and those that are actually high performing (Class +1). If the classifiers had no skill and assigned stocks to the Buy group at random, then this would be 25%. The figures achieved by all classifiers are in the range 33–59%,

Table 2
Actual versus predicted classification rates for training and testing dataset.

Method	Predicted	Actual training			Actual testing		
		High return	Normal return	Correct (%)	High return	Normal return	Correct (%)
CBR	High return	211	78	**73.01**	78	54	**59.09**
	Normal return	124	390	75.88	57	155	73.11
	Correct (%)	62.99	83.33	74.84	57.78	74.16	67.73
MLP	High return	134	117	**53.39**	46	65	**41.44**
	Normal return	145	407	73.73	69	164	70.39
	Correct (%)	48.03	77.67	67.37	40	71.62	61.05
Logistic regression	High return	101	65	**60.84**	34	43	**44.16**
	Normal return	178	459	72.06	81	186	69.66
	Correct (%)	36.2	87.6	69.74	29.57	81.22	63.95
C5	High return	121	69	**63.68**	43	42	**50.59**
	Normal return	69	455	86.83	72	187	72.2
	Correct (%)	63.68	86.83	71.73	37.39	81.66	66.86
QUEST	High return	214	217	**49.65**	80	105	**43.24**
	Normal return	65	307	82.53	35	124	77.99
	Correct (%)	76.7	58.59	64.88	69.57	54.15	59.3
CART	High return	114	60	**65.52**	33	43	**43.42**
	Normal return	165	464	73.77	82	186	69.4
	Correct (%)	40.86	88.55	71.98	28.7	81.22	63.66
GRI	High return	239	368	**39.37**	88	178	**33.08**
	Normal return	40	156	79.59	27	51	65.38
	Correct (%)	85.66	29.77	49.19	76.52	22.27	40.41

Notes: Bolds are the percentage of shares predicted to be High return class and those that are actually high return class.

Table 3
Performance evaluation of the classification techniques.

Tests	Methods						
	CBR	MLP	Logistic regression	C5	QUEST	CART	GRI
Average return	0.134	0.103	0.103	0.096	0.101	0.100	0.095
Sharpe ratio	0.295	0.188	0.121	0.136	0.100	0.111	0.048
Ideal profit	0.421	0.234	0.141	0.169	0.126	0.142	0.063

Table 4
McNemar values for the pair-wise comparison of performance between models.

	GRI	CART	C5	Logistic reg.	MLP	CBR
QUEST	0.8571[a]	6.1489**	1.0638	48.0769**	6.2051**	10.2558**
GRI	–	4.3128*	4.2056*	13.9655**	5.7339*	22.4536**
CART	–	–	1.4938	19.6600**	11.5783**	18.3333**
C5	–	–	–	0.2941	3.5555*	16.9000**
Logistic reg.	–	–	–	–	5.5000*	16.3203**
MLP	–	–	–	–	–	8.1000**

* Significant at 5%.
** Significant at 1%.
[a] Chi-square value.

with the CBR performing best and the GRI classifier worst. Given our large sample size, all these ratios are again very significantly higher than the benchmark figure of 25%.

We have used the Sharpe ratio, average return and ideal profit ratio for out-of-sample (test sample) comparison. The Sharpe ratio can be defined as the mean return of the trading strategy by its standard deviation. In other words, The Sharpe ratio measures return to the risk taken and higher positive values are preferred. Higher value of the Sharpe ratio indicates higher return and lower volatility. The ideal profit ratio measures the return of the corresponding method against a perfect predictor. The range of the ideal ratio is [−1, 1], yet a positive value is desirable [62].

Table 3 presents the Sharpe ratio, ideal profit ratio and average return values for each techniques. All measures are positive. The average return values vary between 0.09 and 0.134. CBR approach has the highest average return while GRI method has the lowest average return. Ideal profit ratio indicates that CBR approach is better than the other techniques. The Sharpe ratio values for each technique are positive and the highest value is given for CBR approach. Based on these three tests, we can conclude that CBR approach outperforms the other techniques for short term stock selection.

In addition to profit comparison, we conducted McNemar test to examine whether or not the classification performance of the CBR approach is significantly higher than that of other techniques. The McNemar test is a nonparametric test for two related samples and assesses the significance of the difference between two dependent samples when the value of interested variable is a dichotomy. Since we are interested in the correct classification of cases, the measure for testing is classification accuracy (the number of correct classifications to the total number of holdout samples). Table 4 shows the results of McNemar testing in comparing with the classification ability between GA based CBR approach and other techniques for holdout samples.

The CBR method performs significantly better than the other techniques at the 1% level. In addition to this, MLP technique performs significantly better than QUEST, CART at the 1% level, and GRI, C5 logistic regression at the 5% level. Logistic regression is significantly better than QUEST, GRI, CART and C5 at the 1% level. There are no differences in terms of performances among the other techniques. From the empirical results, we conclude that CBR approach is an effective approach that can be used for short term stock selection problem.

5. Conclusion

Stock selection problem is an important and widely studied topic due to its significant impact on profitability of portfolio. This paper presents a new case-based reasoning technique, namely CBR, which is an efficient and robust artificial intelligent (AI) paradigm for problem solving. AI techniques have received significant attention in recent years to solve the financial problems. GA is used to assign relative importance of the feature weights for case indexing and retrieving.

The literature contains numerous varieties of learning techniques such as neural networks, induction and decision trees. Experimental results showed that CBR was significantly better than other models in terms of classification accuracy. We have shown that this approach increases overall classification accuracy rate significantly. This can be explained by the following results. Since GA defines a domain specific fitness function, the knowledge acquired by problem domains supports the retrieval of similar cases in solving the problem. Furthermore, GA is an effective knowledge extraction method. Therefore, we can obtain near optimal weights for each feature.

A promising direction for the future is to integrate the CBR with other learning techniques for financial market. A hybrid approach can improve the performance of the CBR. Feature selection approaches can be used to select the important features.

References

[1] A. Aamodt, E. Plaza, Artificial intelligence communications, AI Commun. 7 (1) (1994) 39–59.
[2] C.C. Aggarwal, P.S. Yu, Online generation of association rules, in: Proceedings of the 14th International Conference on Data Engineering, IEEE Computer Society Press, Los Alamitos, CA, 1998, pp. 402–411.
[3] H. Ahn, K.J. Kim, Bankruptcy prediction modeling with hybrid case-based reasoning and genetic algorithms approach, Appl. Soft Comput. 9 (2) (2009) 599–607.
[4] G. Albanis, R. Batchelor, Combining heterogeneous classifiers for stock selection, Intell. Syst. Account. Financ. Manage. 15 (1–2) (2007) 1–21.
[5] A. Alford, J. Jones, R. Leftwich, M. Zmijewski, The relative informativeness of accounting disclosures in different countries, J. Account. Res. 31 (1993) 183–223.
[6] G. Armano, M. Marchesi, A. Murru, A hybrid genetic-neural architecture for stock indexes forecasting, Inf. Sci. 170 (1.) (2004) 3–33.
[7] G.S. Atsalakis, E.M. Dimitrakakis, C.D. Zopounidis, Elliott wave theory and neuro-fuzzy systems, in stock market prediction: the WASP system, Expert Syst. Appl. 38 (8) (2011) 9196–9206.

[8] G.S. Atsalakis, K.P. Valavanis, Forecasting stock market short-term trends using a neuro-fuzzy based methodology, Expert Syst. Appl. 36 (7) (2009) 10696–10707.
[9] G.S. Atsalakis, K.P. Valavanis, Surveying stock market forecasting techniques – Part II: Soft computing methods, Expert Syst. Appl. 36 (3 (Part 2)) (2009) 5932–5941.
[10] R. Ball, S. Kothari, Security returns around earnings announcements, Account. Rev. 66 (1991) 718–738.
[11] M.L. Borrajo, J.M. Corchado, E.S. Corchado, M.A. Pellicer, J. Bajo, Multi-agent neural business control system, Inf. Sci. 180 (2010) 911–927.
[12] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Wadsworth and Brooks/Cole, Monterey, 1984.
[13] A.C. Briza, P.C. Naval, Stock trading system based on the multi-objective particle swarm optimization of technical indicators on end-of-day market data, Appl. Soft Comput. 11 (1) (2011) 1191–1201.
[14] S. Chakravarty, P.K. Dash, A PSO based integrated functional link net and interval type-2 fuzzy logic system for predicting stock market indices, Appl. Soft Comput. 12 (2) (2012) 931–941.
[15] K. Chan, W. Fong, B. Kho, R. Stulz, Information, trading and stock returns: lessons from dually-listed securities, J. Bank. Financ. 20 (1996) 1161–1187.
[16] V. Chari, R. Jagannathan, A. Ofer, Seasonalities in security returns: the case of earnings announcements, J. Financ. Econ. 21 (1988) 101–121.
[17] Y. Chen, X. Dong, Y. Zhao, Stock index modelling using EDA based local linear wavelet neural network, in: Proceedings of International Conference on Neural Networks and Brain, 2005, pp. 1646–1650.
[18] M.C. Chiang, I.C. Tsai, C.F. Lee, Fundamental indicators, bubbles in stock returns and investor sentiment, Q. Rev. Econ. Financ. 51 (1) (2011) 82–87.
[19] C. Chiu, A case-based customer classification approach for direct marketing, Expert Syst. Appl. 22 (2) (2002) 163–168.
[20] S.H. Chun, Y.J. Park, Dynamic adaptive ensemble case-based reasoning: application to stock market prediction, Expert Syst. Appl. 28 (3) (2005) 435–443.
[21] S.H. Chun, Y.J. Park, A new hybrid data mining technique using a regression case based reasoning: application to financial forecasting, Expert Syst. Appl. 31 (2) (2006) 329–336.
[22] J.M. Corchado, J. Aiken, Hybrid artificial intelligence methods in oceanographic forecast models, IEEE Trans. Syst. Man Cybern. C 32 (4) (2002) 307–313.
[23] Z. Da, E. Schaumburg, Relative valuation and analyst target price forecasts, J. Financ. Mark. 14 (1) (2011) 161–192.
[24] D.H. Dorr, A.M. Denton, Establishing relationships among patterns in stock market data, Data Knowl. Eng. 68 (3) (2009) 318–337.
[25] L. Dymova, P. Sevastianov, P. Bartosiewicz, A new approach to the rule-base evidential reasoning: stock trading expert system application, Expert Syst. Appl. 37 (8) (2010) 5564–5576.
[26] R. Gençay, Optimization of technical trading strategies and the profitability in security markets, Econ. Lett. 59 (1998) 249–254.
[27] T.J. Hsieh, H.F. Hsiao, W.C. Yeh, Forecasting stock markets using wavelet transforms and recurrent neural networks: an integrated system based on artificial bee colony algorithm, Appl. Soft Comput. 11 (2) (2011) 2510–2525.
[28] W. Huang, Y. Nakamori, S.Y. Wang, Forecasting stock market movement direction with support vector machine, Comput. Oper. Res. 32 (10) (2005) 2513–2522.
[29] H. Ince, T.B. Trafalis, A hybrid model for exchange rate prediction, Decis. Support Syst. 42 (2) (2006) 1054–1062.
[30] H. Ince, T.B. Trafalis, Kernel principal component analysis and SVMs for stock price prediction, IIE Trans. 39 (6) (2007) 629–637.
[31] O. Jangmin, J. Lee, J.W. Lee, B.T. Zhang, Adaptive stock trading with dynamic asset allocation using reinforcement learning, Inf. Sci. 176 (2006) 2121–2147.
[32] H. Jo, I. Han, H. Lee, Bankruptcy prediction using case-based reasoning, neural networks and discriminant analysis, Expert Syst. Appl. 13 (2) (1997) 97–108.
[33] K.J. Kim, I. Han, Maintaining case-based reasoning systems using a genetic algorithms approach, Expert Syst. Appl. 21 (2001) 139–145.
[34] H. Kim, K. Shin, A hybrid approach based on neural networks and genetic algorithms for detecting temporal patterns in stock markets, Appl. Soft Comput. 7 (2) (2007) 569–576.
[35] K.J. Kim, Financial time series forecasting using support vector machines, Neurocomputing 55 (2003) 307–319.
[36] K. Kim, Toward global optimization of case-based reasoning systems for financial forecasting, Appl. Intell. 21 (3) (2004) 239–249.
[37] K.S. Kim, I. Han, The cluster-indexing method for case-based reasoning using self-organizing maps and learning vector quantization for bond rating cases, Expert Syst. Appl. 21 (3) (2001) 147–156.
[38] J. Kolodner, Case-Based Reasoning, Morgan Kaufmann, San Mateo, CA, 1993.
[39] S.P. Kothari, Capital markets research in accounting, J. Account. Econ. 31 (2001) 105–231.
[40] M. Lam, Neural network techniques for financial performance prediction: integrating fundamental and technical analysis, Decis. Support Syst. 37 (2004) 567–581.
[41] W. Leigh, R. Hightower, N. Modani, Forecasting the New York stock exchange composite index with past price and interest rate on condition of volume spike, Expert Syst. Appl. 28 (1) (2005) 1–8.
[42] W. Leigh, W. Purvis, J.M. Ragusa, Forecasting the NYSE composite index with technical analysis, pattern recognizer, neural network, and genetic algorithm: a case study in romantic decision support, Decis. Support Syst. 32 (2002) 361–377.
[43] H. Li, J. Sun, Gaussian case-based reasoning for business failure prediction with empirical data in China, Inf. Sci. 179 (2009) 89–108.
[44] H. Li, J. Sun, J. Wu, X.J. Wu, Supply chain trust diagnosis (SCTD) using inductive case-based reasoning ensemble (ICBRE): the case of general competence trust diagnosis, Appl. Soft Comput. 12 (8) (2012) 2312–2321.
[45] J. Lima, R. Francisco, L. Osiro, L.C.R. Carpinetti, A fuzzy inference and categorization approach for supplier selection using compensatory and non-compensatory decision rules, Appl. Soft Comput. (2013), http://dx.doi.org/10.1016/j.asoc.2013.06.020.
[46] C.J. Lu, T.S. Lee, C.C. Chiu, Financial time series forecasting using independent component analysis and support vector regression, Decis. Support Syst. 47 (2) (2009) 115–125.
[47] T.H. Moon, S.Y. Sohn, Case-based reasoning for predicting multiperiod financial performances of technology based SMEs, Appl. Artif. Intell. 22 (6) (2008) 602–615.
[48] F.C. Pampel, Logistic regression: a primer, in: Sage University Papers Series on Quantitative Applications in the Social Sciences, Series No. 07-132, Sage, Thousands Oaks, CA, 2000.
[49] Y.J. Park, E. Choi, S.H. Park, Two-step filtering datamining method integrating case-based reasoning and rule induction, Expert Syst. Appl. 36 (1) (2009) 861–871.
[50] I. Pereira, A. Madureira, Self-optimization module for scheduling using case-based reasoning, Appl. Soft Comput. 13 (3) (2013) 1419–1432.
[51] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1) (1986) 81–106.
[52] S. Ramnath, S. Rock, P. Shane, The financial analyst forecasting literature: a taxonomy with suggestions for further research, Int. J. Forecast. 24 (1) (2008) 34–75.
[53] K.S. Shin, I. Han, Case-based reasoning supported by genetic algorithms for corporate bond rating, Expert Syst. Appl. 16 (2) (1999) 85–95.
[54] K.S. Shin, I. Han, A case-based approach using inductive indexing for corporate bond rating, Decis. Support Syst. 32 (1) (2001) 41–52.
[55] P. Smyth, R.M. Goodman, An information theoretic approach to rule induction from databases, IEEE Trans. Knowl. Data Eng. 4 (4) (1992) 310–316.
[56] N. Stéphane, R. Hector, L.L.J. Marc, Effective retrieval and new indexing method for case based reasoning: application in chemical process design, Eng. Appl. Artif. Intell. 23 (6) (2010) 880–894.
[57] F.E.H. Tay, L.J. Cao, Modified support vector machines in financial time series forecasting, Neurocomputing 48 (2002) 847–861.
[58] H.J. Teoh, et al., Fuzzy time series model based on probabilistic approach and rough set rule induction for empirical research in stock markets, Data Knowl. Eng. 67 (1) (2008) 103–117.
[59] B. Trueman, M.H.F. Wong, X.J. Zhang, Anomalous stock returns around Internet firms' earnings announcements, J. Account. Econ. 34 (1–3) (2003) 249–271.
[60] B.J. Vanstone, G.R. Finnie, C.N.W. Tan, Evaluating the application of neural networks and fundamental analysis in the Australian stock market, in: Proceedings of Computational Intelligence, 2005, p. 487.
[61] B.S. Yang, T. Han, Y.S. Kim, Integration of ART-Kohonen neural network and case-based reasoning for intelligent fault diagnosis, Expert Syst. Appl. 26 (3) (2004) 387–395.
[62] V. Zakamouline, S. Koekebakker, Portfolio performance evaluation with generalized Sharpe ratios: beyond the mean and variance, J. Bank. Financ. 33 (7) (2009) 1242–1254.
[63] H. Zhao, A multi-objective genetic programming approach to developing Pareto optimal decision trees, Decis. Support Syst. 43 (2007) 809–826.
[64] R.T. Zhou, R.N. Lai, Herding and information based trading, J. Empir. Financ. 16 (3) (2009) 388–393.