Proceedings of KFIS Autumn Conference 2007 Vol. 17, No. 2.

자료편집기법과 사례기반추론을 이용한 재무예측시스템

Financial Forecasting System using Data Editing Technique and Case-based Reasoning

김경재

서울시 중구 필동 3-26 동국대학교 경영정보학과
E-mail: kjkim@dongguk.edu

Abstract

This paper proposes a genetic algorithm (GA) approach to instance selection in case-based reasoning (CBR) for the prediction of Korea Stock Price Index (KOSPI). CBR has been widely used in various areas because of its convenience and strength in complex problem solving. Nonetheless, compared to other machine learning techniques, CBR has been criticized because of its low prediction accuracy. Generally, in order to obtain successful results from CBR, effective retrieval of useful prior cases for the given problem is essential. However, designing a good matching and retrieval mechanism for CBR systems is still a controversial research issue. In this paper, the GA optimizes simultaneously feature weights and a selection task for relevant instances for achieving good matching and retrieval in a CBR system. This study applies the proposed model to stock market analysis. Experimental results show that the GA approach is a promising method for instance selection in CBR.

Key Words : Instance selection; Genetic algorithms; Case-based reasoning, Stock market prediction

1. Introduction

Case-based reasoning(CBR) is a popular inference technique and has been applied to many business problems. The basic idea of CBR is to find a solution to new problems by adopting solutions that have been used in the past. Although most artificial intelligence techniques pursue generalized relationships between problem descriptors and conclusions, it just refers to specific knowledge of previously experienced, concrete problem situations, so it is effective for complex and unstructured problems and easy to update (Shin & Han, 1999). The general process of CBR is shown graphically in Figure 1 (adopted from Turban & Aronson, 2001). Among the steps of the flowchart, the second process, case retrieval, is the most important step because the performance of CBR systems usually depends on it (Kolodner, 1993).

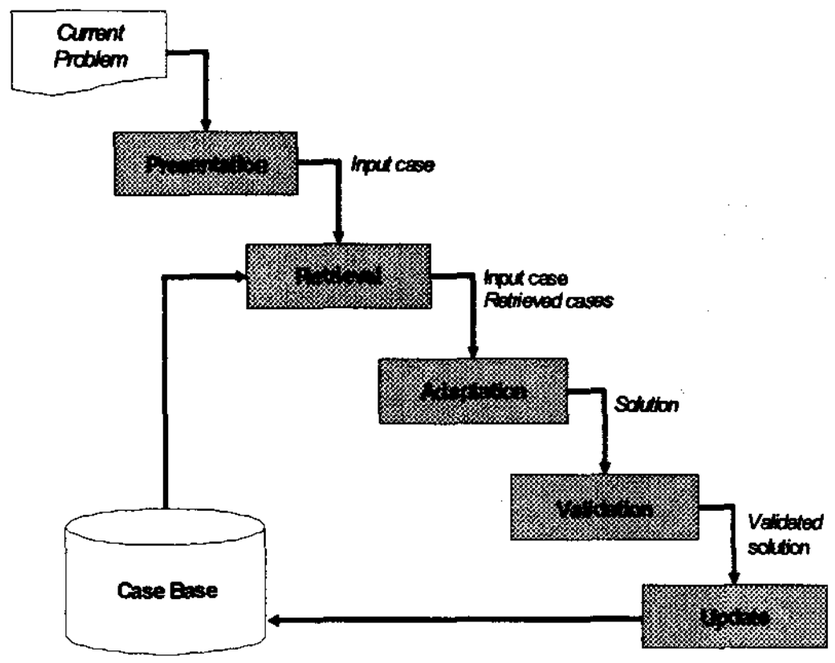

Figure 1. The general process of CBR

In this step, the CBR system retrieves the most similar cases from the case memory, which become the bases for solution of the input problem. Thus, it is crucial to determine appropriate similar cases. In particular, feature weighting (selection) and instance selection for

measuring similarity have been controversial issues in designing CBR systems. There have been many studies to determine these factors. Among many methods of instance selection and feature weighting, GAs are increasingly being used in CBR systems.

This paper proposes a new hybrid model of CBR and genetic algorithms (GAs) for feature weighting and instance selection in the context of stock market prediction. An evolutionary instance selection algorithm reduces the dimensionality of data and may eliminate noisy and irrelevant instances. In addition, this study searches the optimal feature weights for the relevant features in case retrieval process.

The rest of this paper is organized as follows: Section 2 proposes the evolutionary instance selection algorithmand describes the benefits of the proposed algorithm. Section 3 describes the application of the proposed algorithm. In the final section, conclusions and the limitations of this study are presented.

2. A Genetic Algorithms Approach to Instance Selection for CBR

As mentioned earlier, there are many studies on instance selection for the instance-based learning algorithm. However, there are few studies on instance selection for CBR. Thus, there are few relevant theories concerning instance selection for CBR. This paper proposes the GA approach to instance selection for CBR (ICBR). In this study, the GA supports the simultaneous optimization of feature weights and selection of relevant instances. The detail explanation for each phase of ICBR is presented as follows.

Step 1. For the first step, the system searchesthe space to find optimal or near-optimal parameters (feature weights and selection variables for each instance). To apply GA to search these optimal parameters, they have to be coded on a chromosome, a form of binary strings.

The value of the code for instance selection is set to '0' or '1'. '0' means the corresponding instance is not selected and '1'means selected. Because a sign for each instance selection requires just 1 bit, so n bits are required to implement instance selection by GA where n is the number of total instances.

The population (a set of seed chromosomes for finding optimal parameters) is initiated into random values before the search process. And, the encoded chromosome is searched to maximize the specific fitness function. The objective of the study is to determine appropriate the feature weights and instance selection of CBR systems, which produce the highest prediction accuracy for the test data. Thus, we set the prediction accuracy of the test data asthe fitness function for GA (Shin & Han, 1999; Kim, 2004). Mathematically, the fitness function (f_T) for the test set T can be expressed as equation (1):

$$f_T = \frac{1}{n}\sum_{i=1}^{n} CA_i$$

$$CA_i = 1 \text{ if } PO_i = AO_i \text{ for the item } I_i$$

$$CA_i = 0 \text{ if } PO_i \neq AO_i \text{ for the item } I_i \quad (1)$$

where CA_i is the classification accuracy for the ith test case, I_i, which is denoted by 1 or 0 ('correct'=1, 'incorrect'=0), PO_i is the predicted output from the model for the ith test case, AO_i is the actual output from the model for the ith test case and test set T is $\{I_1, I_2, I_3, \cdots, I_n\}$.

Step 2. In the second step, the parameters that are set in Step 1 are appliedto the CBR system and general reasoning process of CBR goes on. We use the weighted average of Euclidean distance for the each feature as a similarity measure. And, we use 1-NN(one-nearest neighbor) matching as a method of case retrieval. After adoption reasoning process for all of test cases, the values of the fitness function (f_T) for the items of test set T are

updated.

Step 3. In this step, the process of GA's evolution goes on towards the direction to maximize the value of the fitness function. It includes selection of the fittest, crossover and mutation. Step 2 and 3 are iterated again and again until the stopping conditions are satisfied.

Step 4. In the last stage, the system determines the parameters - the optimal weights of features and selection of instances whose performance for the test data is the best. And, it applies them to the hold-out data to check the generalizability of the selected parameters. Sometimes, optimized parameters by GA fit to the test data, but they don't fit to the unknown data, i.e. overfitting. Thus, this step is required to check the possibility of overfitting.

3. Application: Analysis of Stock Market Data

This section applies ICBR to stock market prediction. The efficiency and effectiveness of ICBR may be properly tested because the stock market data is very noisy and complex. Many studies on stock market prediction using artificial intelligence techniques were performed in the past decade. Some of them, however, did not produce outstanding prediction accuracy partly because of the tremendous noise and non-stationary characteristics in stock market data. If these factors are not appropriately controlled, the prediction system does not produce significant performance.

3.1 Application Data

The application data used in this study consists of technical indicators and the direction of change in the daily Korea stock price index (KOSPI). The total number of samples is 2928 trading days, from January 1989 to December 1998. This study divides the samples into eight data sets according to the trading year. Experiments are repeated eight times for each data set to reflect specific knowledge as time passes.

The direction of daily change in the stock price index is categorized as "0" or "1". "0" means that the next day's index is lower than the today's index, and "1" means that the next day's index is higher than today's index. We select twelve technical indicators as feature subsets by the review of domain experts and prior research.

3.2 Experiments

Experiments are carried outfor the following three models:

Whole training data. The whole training samples are used as the training data. This is the conventional method of data analysis.

Selected instances with ICBR. Experiments on stock market data are implemented using ICBR. The procedure of the experiment is as follows. The GA searches for optimal or near-optimal featureweights and relevant instances for CBR. As mentioned earlier, this study needs two sets of parameters: The weight codes for the relevant features and the codes for instance selection.

This study uses the following encoding for the strings: 12 input features are used. Thus, the first 12 bits represent the feature weights for the relevant features. These bits are searched from -5 to 5. The following bits are instance selection codes for the training data. The chromosome of these bits consists of n genes (where n is the number of initial training instances), each one with two possible states: 0 or 1. "1"means the associated instance is selected into the analysis and "0" means the associated instance is not chosen.

The encoded chromosomes are searched to maximize the fitness function. The fitness function is specific to applications. In this study, the objectives of the model are to approximate connection weights and to select relevant instances for the correct solutions. These objectives canbe represented by the average prediction accuracy of the selected instances within the training data. Thus, this study applies the average prediction accuracy of the selected instances within the training data to the fitness function. Mathematically, the fitness function is represented as equation (2):

$$Fitness = \frac{1}{n}\sum_{i=1}^{n} CR_i \qquad (i = 1, 2, n)$$

$$\begin{aligned} &if\ POi = AOi \quad CRi = 1 \\ &otherwise \quad CRi = 0 \end{aligned} \qquad (2)$$

where CRi is the prediction result for the i th trading day which is denoted by 0 or 1, POi is the predicted output from the model for the i th trading day, and AOi is the actual output for the i th trading day.

For the controlling parameters of the GA search, the population size is set at 100 organisms and the crossover and mutation rates are varied to prevent CBR from falling into a local minimum. The value of the crossover rate is set at 0.7 while the mutation rate is 0.1. For the crossover method, the uniform crossover method is considered better at preserving the schema, and can generate any schema from the two parents, while single-point and two-point crossover methods may bias the search with the irrelevant position of the variables. Thus, this study performs crossover using the uniform crossover routine. For the mutation method, this study generates a random number between 0 and 1 for each of the variables in the organism. If a variable gets a number that is less than or equal to the mutation rate, then that variable is mutated. As the stopping condition, only 100 generations are permitted.

3.3 Experimental Results

This study compares ICBR to the conventional CBR. ICBR uses the GA to determine the feature weights and learns the patterns of the stock market data from the selected instances through an evolutionary search process. For the conventional CBR model, about 20% of the data is used for holdout and 80% for training. The number of the training instances in the conventional CBR are 2347 and the number of the selected instances within the training instances in ICBR are 820. Table 1 describes the average prediction accuracy of each model.

Table 1. Average Prediction Performance

Year	CBR	ICBR
1989	56.1%	56.1%
1990	50.0%	58.6%
1991	51.7%	56.9%
1992	44.0%	51.7%
1993	49.2%	54.2%
1994	52.5%	54.2%
1995	58.6%	55.2%
1996	62.1%	56.9%
1997	51.7%	53.4%
1998	44.8%	51.7%
Total	52.06%	54.89%

4. Conclusions

In this paper, we use the GA for CBR in two ways. We first use the GA to determine the feature weights. In addition, we adopt the evolutionary instance selection algorithm. This directly removes irrelevant and redundant instances from the training data. We conclude that GA-based learning and the instance selection algorithm significantly outperforms the conventional CBR model in stock market prediction.

The prediction performance may be more enhanced if the GA is employed not only for instance selection but also for relevant feature selection, and this remains a very interesting topic for further study. Although instance selection is a direct method of noise and dimensionality reduction, feature selection effectively reduces the dimensions of feature space.

References

Kim, K. (2004) Toward global optimization of CBR systems for financial forecasting, Applied Intelligence, 21(3), 239-249.

Kolodner, J. (1993) Case-based Reasoning, Margan Kaufmann, San Mateo, CA.

Shin, K.S., and I. Han (1999) Case-based reasoning supported by genetic algorithms for corporate bond rating, Expert Systems with Applications, 16, 85-95.

Turban, E. and J.E. Aronson (2001) Decision support systems and intelligent systems (6th edition), Prentice-Hall: Upper Saddle River, NJ.