Hybrid Case-based Reasoning and Genetic Algorithms Approach for Customer Classification

Kyoung-jae Kim and Hyunchul Ahn, *Member, KIMICS*

Abstract—This study proposes hybrid case-based reasoning and genetic algorithms model for customer classification. In this study, vertical and horizontal dimensions of the research data are reduced through integrated feature and instance selection process using genetic algorithms. We applied the proposed model to customer classification model which utilizes customers' demographic characteristics as inputs to predict their buying behavior for the specific product. Experimental results show that the proposed model may improve the classification accuracy and outperform various optimization models of typical CBR system.

Index Terms—Case-based reasoning, genetic algorithms, feature selection, instance selection, customer classification.

I. INTRODUCTION

CBR is a problem solving technique that reuses past experiences to find a solution. It often improves the effectiveness of complex and unstructured decision making, so it has been applied to various problem-solving areas including manufacturing, finance and marketing.

However, it's not easy to obtain successful results with high classification accuracy by applying CBR because there is no mechanism to design effective systems in typical CBR. In particular, it is very important to design appropriate mechanism for case retrieval. In this aspect, the selections of the appropriate feature and instance subsets in the case retrieval step have been the most important research issues.

For these reasons, simultaneous optimization of several components in CBR attracts the interests of researchers. As a pioneering study, the approach to combine simultaneously feature and instance selection was proposed (Kuncheva & Jain, 1999; Rozsypal & Kubat, 2003). Nonetheless, there have been few attempts to apply the simultaneous model to real-world data.

This article proposes genetic algorithms (GA) approach to optimize the feature and instance selection simultaneously. In addition, we apply the proposed model to the real-world case of customer classification and present experimental results from the application.

This article is organized as follows. Section 2 provides a brief review for prior research and the next section describes our proposed model. In section 4, the research design and experiments are explained. In the fifth section, the empirical results are summarized and discussed. The final section presents contributions and the limitations of this study.

Manuscript received November 20, 2005.

Kyoung-jae Kim: Corresponding author, Department of Information Systems, Dongguk University (Tel: +82-2-2260-3324, Fax: +82-2-2260-3684, Email: kjkim@dongguk.edu)

Hyunchul Ahn: Graduate School of Management, KAIST.

II. RESEARCH BACKGROUNDS

Feature selection is the process of picking a subset of features that are relevant to the target concept and removing irrelevant or redundant features. These are important factors that determine the performance of the classification model, so they have been the most popular research issues in designing most classification models.

In addition, instance selection in CBR literature is the technique that selects an appropriate reduced subset of a case-base and applies the nearest-neighbor rule to the selected subset. It may increase the performance of CBR systems dramatically if the systems contain many noisy cases. So, it has been another popular research issue in CBR systems for a long time.

Prior study tried to integrate simultaneously above two concepts. Kuncheva and Jain (1999) first proposed a data reduction method via feature and instance selection in CBR, so there are few studies because of its short history They proposed simultaneous optimization of feature and instance selection using GA, and compared their model to sequential combining of traditional feature and instance selection algorithms. Rozsypal and Kubat (2003) also tried simultaneous optimization of feature and instance selection using GA, but they differentiated their model by applying the value encoding method and more effective design of the fitness function. They showed that their model outperforms the model by Kuncheva and Jain. Fig. 1 shows an example of simultaneous feature and instance selection methods. In Fig. 1, the original data set consists of four features and ten instances. However, the reduced data set have only two features and four instances via simultaneous feature and instance selection process.

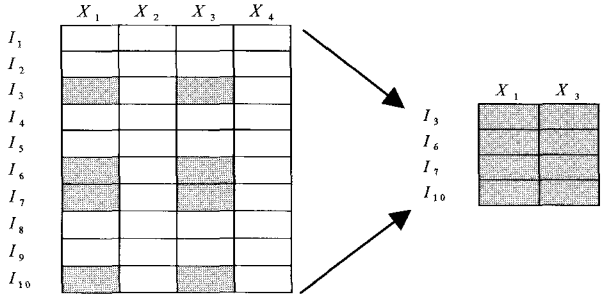

Fig. 1 Integration of feature and instance selection

III. HYBRID CASE-BASED REASONING AND GENETIC ALGORITHMS APPROACH

In order to enhance the performance of typical CBR systems, this study proposes data reduction technique via feature and instance selection process using genetic algorithms. The detail explanation for each phase of the proposed model is presented as follows.

In the first phase, the system searches the space to find optimal or near-optimal parameters (selection codes for each feature and instance). To apply GA to search for these optimal parameters, they have to be coded on a chromosome, a form of binary strings. The length of each chromosome for the proposed model is $n + m$ bits where n is the number of features and m is the number of instances. The value of the code for feature and instance selection is set to '0' or '1'. '0' means the corresponding feature and instance are not selected and '1' means they are selected. A sign for each feature and instance selection requires just 1 bit, so $n + m$ bits are required to implement simultaneous data reduction technique by GA.

The population (a set of seed chromosomes for finding optimal parameters) is initiated into random values before the search process. The population is searched to find the encoded chromosome for maximizing the specific fitness function. The objective of the study is to select relevant features and instances for CBR systems, which produce the highest classification accuracy for the test data. Thus, we set the classification accuracy of the test data as the fitness function for GA (Shin & Han, 1999; Kim, 2004).

In the second phase, the parameters that are set in the first stage are applied to the CBR system and the general reasoning process of CBR goes on. We use the weighted average of Euclidean distance for each feature as a similarity measure. We use 3-NN (three-nearest neighbor) matching as a method of case retrieval. It means the system searches for three nearest neighbors for an input case and suggests final classification result by voting for them. We set k parameter of k-NN to 3 because 3-NN showed the best prediction accuracy among 1-NN, 3-NN, 5-NN, 7-NN and 9-NN in this paper. After the adoption of the reasoning process for all of test cases, the values of the fitness function are updated.

In the third phase, the process of GA's evolution goes on towards the direction of maximizing the value of the fitness function. It includes selection of the fittest, crossover and mutation. The second and third phases are iterated repeatedly until the stopping conditions are satisfied.

In the last phase, the system determines the parameters whose performance for the test data is the best. It applies them to the hold-out data.

IV. RESEARCH DATA AND EXPERIMENTS

A. Research data

The research data is collected from an online diet portal site in Korea which contains all kinds of services for online diets such as providing information, community services and a shopping mall. The experimental data includes 980 cases that consist of the purchasing and non-purchasing customers. It contains demographic variables and the status of purchase or non-purchase for the corresponding user. The status of purchase for each user is categorized as '0' or '1' and it is used as a dependent variable. '0' means that the user has not purchased the company's products and '1' means he or she made a purchase. We collect totally 46 independent variables including demographic and other personal information. To eliminate irrelevant variables and make the reasoning process more efficient, we adapt two statistical methods, two-sample t-tests for ratio variables and chi-square tests for nominal variables. Finally, we select only 14 factors which prove to be the most influential in the purchase of the company's product.

We split the data into three groups: reference, test, and hold-out case-bases. The portion of these three case-bases is 60% (588 cases), 20% (196 cases) and 20% (196 cases) each.

B. Research design and experiments

For the controlling parameters of GA search for our experiments, we use 200 organisms in the population and set the crossover to 0.7 and mutation rate to 0.1. As a stopping condition, we use 4000 trials (20 generations).

To test the effectiveness of the proposed model, we also apply three different CBR models to the same data set. The first model, labeled CCBR (Conventional CBR), uses a conventional approach for the reasoning process of CBR. This model considers all initially available features as a feature subset. That is to say, there is no special process of feature subset selection. In addition, instance selection is not considered here, so all instances are used in this model.

The second model selects relevant features using genetic algorithms. This study names this model FCBR (Feature selection for CBR). In this model, we try to optimize feature selection by GA, but we are still unconcerned with instance selection. Siedlecki and Sklanski (1989) and Kim (2004) proposed similar models.

The third model uses GA to select a relevant instance subset. This study names the model ICBR (Instance selection for CBR). In this model, we try to optimize instance selection by GA, but we are unconcerned with feature selection. Babu and Murty (2001) proposed a similar model.

These experiments are done by our private prototype software which is developed using Microsoft Excel 2003 and Evolver Industrial Version 4.08 (Palisade Software, www.palisade.com), a commercial GA tool. The 3-NN algorithm is implemented in VBA (Visual Basic for Applications) of Microsoft Excel 2003 and the VBA codes are incorporated with Evolver to optimize the parameters by GA.

V. EXPERIMENTAL RESULTS

For CCBR, we apply k-NN algorithm by varying parameter k into odd numbers range from 1 to 9. As a result, we find that 3-NN shows the best performance. So, as indicated earlier, we apply 3-NN to all of other CBR models.

We compare the prediction performances of the proposed model and other alternative models. Table 1 describes the average prediction accuracy of each model. Among the models, the proposed model has the highest level of accuracy (64.29%) in the given hold-out data set, followed by ICBR (63.27%) and FCBR (60.20%). It means the proposed model is effective to enhance performances of other CBR models. The results also show that ICBR outperforms FCBR. It means selection of appropriate instances is more important than proper selection of the features for this data set.

Table 1 Average prediction accuracy of the models

Mode / Data	CCBR	FCBR	ICBR	Proposed model
Test	62.24%	62.76%	66.33%	70.92%
Hold-out	56.12%	60.20%	63.27%	64.29%

We also can find that ICBR and the proposed model use only a portion of the training case-base, so ICBR use 89.12% (524/588) and the proposed model use 81.29% (478/588) of total training samples. In addition, the results show that FCBR and the proposed model use only a part of the features. FCBR employs only 8 features and the proposed model uses 10 features through feature selection procedure while CCBR and ICBR use 14 features.

We use the two-sample test for proportions to examine whether the differences of predictive accuracy between the proposed model and other comparative algorithms is statistically significant. Table 2 shows Z values for the pairwise comparison of performance between models. As shown in Table 2, the proposed model outperforms CCBR at the 5% statistical significance level, but it does not outperform other two models with statistical significance. We can also find that ICBR outperforms CCBR at the 5% statistical significance level.

Table 2 Z values for the hold-out data

	FCBR	ICBR	Proposed model
CCBR	-0.8191	-1.4416*	-1.6510*
FCBR		-0.6235	-0.8335
ICBR			-0.2102

*significant at the 5% level

VI. CONCLUSIONS

In this paper, we have suggested a novel model, the data reduction method via feature and instance selection, to improve the performance of the typical CBR system. This paper uses GA as a tool to optimize the feature and instance selection simultaneously. From the results of the experiment, we show that our proposed model may outperform other comparative algorithms such as CCBR, FCBR, and ICBR in the case of customer classification.

This study has some limitations. First of all, it takes too much computational time to obtain optimal parameters for the proposed model. So, the efforts to make the proposed model more efficient should be followed in the future to apply our model to general cases in reality.

Second, there are other factors which enhance the performance of the CBR system that may be incorporated with the simultaneous optimization model. For example, k parameter of k-NN, the number of cases to combine, may be another parameter to be optimized. Feature weighting is another factor to be optimized in the CBR system. Building a universal simultaneous optimization model for CBR including feature and instance selection as well as other factors like k parameter and feature weights may improve the overall performance of the CBR system.

Finally, the results of this study may depend on the experimental data set. In particular, the research data set is relatively small and seems to be noisy, so it is difficult to distinguish the prediction ability of each model. Consequently, the generalizability of the proposed model should be tested further by applying it to other problem domains in the future.

REFERENCES

[1] T.R. Babu and M.N. Murty, "Comparison of genetic algorithm based prototype selection schemes," *Pattern Recognition*, vol. 34, pp. 523-525, 2001.

[2] K. Kim, "Toward global optimization of case-based reasoning systems for financial forecasting," *Applied Intelligence*, vol. 21, no. 3, pp. 239-249, 2004.

[3] L.I. Kuncheva and L.C. Jain, "Nearest neighbor classifier: Simultaneous editing and feature selection," *Pattern Recognition Letters*, vol. 20, pp. 1149-1156, 1999.

[4] A. Rozsypal and M. Kubat, "Selecting representative examples and attributes by a genetic algorithm," *Intelligent Data Analysis*, vol. 7, pp. 291-304, 2003.

[5] K.S. Shin and I. Han, "Case-based reasoning supported by genetic algorithms for corporate bond rating," *Expert Systems with Applications*, vol. 16, pp. 85-95, 1999.

[6] W. Siedlecki and J. Sklanski, "A note on genetic algorithms for large-scale feature selection," *Pattern Recognition Letters*, vol. 10, pp. 335-347, 1989.

Kyoung-jae Kim
Received his B.B.A. degree from Chung-Ang University and M.E. and Ph.D. degrees from Korea Advanced Institute of Science and Technology. He is currently an assistant professor of MIS in the Department of Information Systems, Dongguk University, Seoul, Korea. He published his papers in *Applied Intelligence*, *Expert Systems*, *Expert Systems with Applications*, *Intelligent Data Analysis*, *Intelligent Systems in Accounting Finance & Management*, *Neural Computing & Applications*, *Neurocomputing*, etc. His research interests include data mining, knowledge management, and intelligent agents.

Hyunchul Ahn
Received the B.S. and M.E. degrees from Korea Advanced Institute of Science and Technology (KAIST). He is currently pursuing the Ph.D. degree in management engineering at KAIST Graduate School of Management. His research interests are in the areas of data mining in marketing and finance and artificial intelligence techniques such as case-based reasoning, genetic algorithms and support vector machines