최적화 사례기반추론을 이용한 통신시장 고객관계관리

Customer Relationship Management in Telecom Market using an Optimized Case-based Reasoning

안현철[1], 김경재[2]

[1] 서울시 동대문구 청량리동 207-43 한국과학기술원 테크노경영대학원
E-mail: hcahn@kaist.ac.kr
[2] 서울시 중구 필동 3-26 동국대학교 경영대학 경영정보학과, 교신저자
E-mail: kjkim@dongguk.edu

Abstract

Most previous studies on improving the effectiveness of CBR have focused on the similarity function aspect or optimization of case features and their weights. However, according to some of the prior research, finding the optimal k parameter for the k-nearest neighbor (k-NN) is also crucial for improving the performance of the CBR system. Nonetheless, there have been few attempts to optimize the number of neighbors, especially using artificial intelligence (AI) techniques. In this study, we introduce a genetic algorithm (GA) to optimize the number of neighbors that combine, as well as the weight of each feature. The new model is applied to the real-world case of a major telecommunication company in Korea in order to build the prediction model for the customer profitability level. Experimental results show that our GA-optimized CBR approach outperforms other AI techniques for this mulriclass classification problem.

Key Words : Case-based Reasoning, Genetic Algorithm, Feature Weighting

1. Introduction

Case-based reasoning (CBR) is a problem-solving technique that is similar to the decision making process that human beings use in many real world applications. It often shows significant promise for improving the effectiveness of decision making in complex and unstructured situations. Due to its high adaptability for general purposes, it has been applied to various problem-solving areas including manufacturing, finance and marketing.

Regardless of its many advantages, there are some problems that must be solved in order to design an effective CBR system. In particular, the fact that there are no mechanisms for designing case indexing or retrieval steps in CBR systems has been the most critical restriction of CBR. In this aspect, the selections of the appropriate similarity measures, feature subsets and their weights in the case retrieval step have been popular research issues.

According to some prior studies, finding the optimal k parameter for k-NN (k-nearest neighbor) may be crucial for improving the performance of CBR systems (Lee & Park, 1999; Garrell i Guiu *et al.*, 1999; Jarmulak *et al.*, 2000). Nonetheless, there have been few attempts to optimize the number of neighbors (i.e. the k parameter).

This paper proposes genetic algorithms (GAs) to optimize not only the feature weights, but also the number of neighbors that combine in the CBR system. This study applies the proposed model to the real-world case for a telecommunication company's customer relationship management (CRM), which is a kind of multiclass classification problem.

2. Simultaneous Optimization of Case-based Reasoning using Genetic Algorithm

Most of prior studies in CBR have only focused on the optimization of feature selection or feature weights, although the k parameter of k-NN may also affect the performance of CBR systems. Thus, in this study, we propose a novel approach for CBR which applies GA as a simultaneous optimization algorithm for the k parameter, as well as feature weights in k-NN. Our study names this model GAKNN (GA-optimized k-Nearest Neighbor algorithm with feature weighting). The framework of GAKNN is shown in Figure 1.

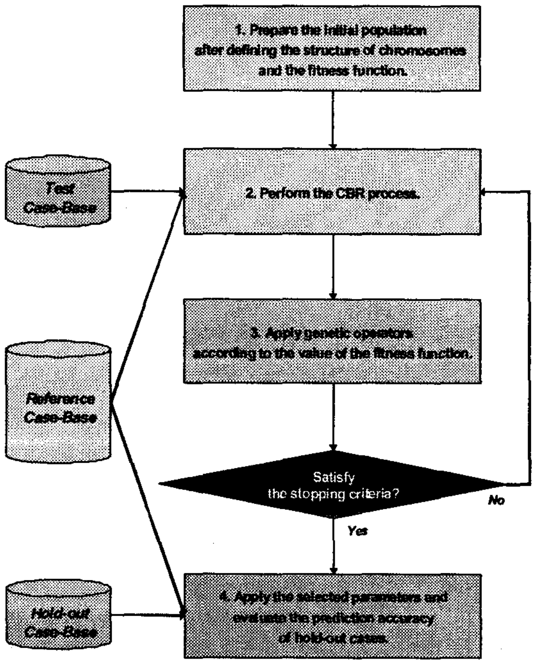
1. Prepare the initial population
after defining the structure of chromosomes
and the fitness function.
2. Perform the CBR process.
3. Apply genetic operators
according to the value of the fitness function.
Satisfy
the stopping criteria?
No
Yes
4. Apply the selected parameters and
evaluate the prediction accuracy
of hold-out cases.
Test
Case-Base
Reference
Case-Base
Hold-out
Case-Base

Figure 1. Framework of GAKNN

The process of GAKNN consists of the following four stages:

Stage 1. Initialization

Before applying GA to search for the optimal feature weights and k parameter, they have to be coded on a chromosome, a form of binary strings. In addition, the fitness function for evaluating each chromosome should be defined. Thus, in this step, the fundamental work for initiating the search process takes place.

The population is searched to maximize the specific fitness function. The objective of this paper is to determine appropriate feature weights and the k parameter of k-NN that lead to accurate prediction results. Evaluation can be performed by using the average prediction accuracy of the test data. Thus, our study applies it to the fitness function of GA.

Stage 2. The application of CBR using current parameters

In the second step, k-NN using selected parameters in Stage 1 takes place for the cases in the test case-base. After the adoption of the reasoning process for all of test cases, the values of the fitness function for the items of test set are updated.

Stage 3. Genetic operation

In the third step, a new generation of the population is produced by applying genetic operators such as selection, crossover, and mutation. According to the fitness values for each chromosome, the chromosomes whose values are high are selected and used for the basis of crossover. The mutation operator is also applied to the population with a very small mutation rate.

After the production of a new generation, Stage 2 the reasoning process with calculation of the fitness values is performed again. From this point, Stage 2 and 3 are repeated again and again until the stopping conditions are satisfied. When the stopping conditions are satisfied, the genetic search finishes and the chromosome that shows the best performance in the last population is selected as the final result.

Stage 4. The calculation of the prediction accuracy

The final stage applies the selected feature weights and k, the optimal number of cases that combine, to the hold-out case-base, and the last reasoning process of CBR goes on. In this stage, the prediction accuracy of the proposed approach is finally evaluated.

3. Research Design

3.1 Research data

Our research data is collected from a major telecommunications service provider in Korea. As explained, the competition between service providers in the Korean telecommunications market is so severe that the companies have motivation to find their valuable customers and preserve them. In this aspect, our target company also has interest in building a customer classification model that classifies its current or potential customers into several levels according to their profitability. In this study, we build a customer classification model for our target case. To improve classification accuracy, we apply the GAKNN method to the collected data.

The experimental data includes 4,000 cases that consisted of various features on customer demographic information and telecommunications call details. It consists of 38 independent variables and a dependent variable which represents the level of customer profitability. Our target company classifies their customers into 4 levels VIP, gold, silver, and bronze. We use this 4-class feature as the dependent variable of our experiment. By applying the independent samples t-test and chi-square test, we select 31 features among the given independent variables that affect the dependent variable with statistical significance at the 95% level. Then, we finally select only 14 factors which prove to be the most influential in customer profitability level. We do this by applying the stepwise selection procedure using *Wilk's lambda* based on multiple discriminant analysis. The *F*-value for stepwise entry is set at 3.84 and that for stepwise removal is set at 2.71. In our experiment, the data is split into the three groups: reference, test and hold-out case-bases.

3.2 Experiments

For the controlling parameters of GA search, the population size is set at 50 organisms, and the crossover and mutation rates are set at 0.7 and 0.1. As the stopping condition, only 1000 trials (20 generations) are permitted. These settings are determined by referencing some prior studies including Shin and Han (1999) and Shin and Han (2000), which tried to optimize the parameters in AI methods using GA.

To compare the result of GAKNN, we also applied two other algorithms to the same dataset. The compared algorithms include artificial neural networks (ANN) and conventional *k*-NN (CKNN).

ANN is designed as a three-layer network whose learning rate and momentum rate are 0.1. The hidden and output nodes use the sigmoid transfer function. We experiment using ANN models by varying the number of their hidden nodes from 7 to 28. Among them, we have chosen the model whose performance is the best. This study allows 150 learning epochs for ANN. To experiment using ANN models, we apply Ward System Group's Neuroshell 4.0.

CKNN is the *k*-NN algorithm but selects *k* as a fixed number that usually ranges from 1 to 10. In this experiment, we select the *k* for CKNN that shows the best performance in the range between 1 and 10.

4. Experimental Results

In this section, the prediction performance of GAKNN and other alternative models is compared. Table 1 describes the average prediction accuracy of each model. our proposed model shows the best performance among comparative models from the viewpoint of prediction accuracy. This may prove the usefulness of our model for customer classification.

Table 1. Average prediction accuracy

Case-base	Reference	Test	Hold-out	Remarks
ANN	98.04%	97.50%	96.75%	No. of hidden nodes = 21
CKNN			93.63%	*k*=4
GAKNN		97.25%	97.88%	*k*=6

McNemar's test is used to examine whether the predictive performance of GAKNN is significantly higher than other algorithms. This test is used with nominal data and is particularly useful with before-after measurement of the same subjects (Kim, 2004). Table 2 shows the results of the McNemar's test to compare the performance of three algorithms for the hold-out data

Table 2. McNemar values

	CKNN	GAKNN
ANN	11.755**	N/A$^{a)*}$
CKNN		24.750**

*significant at the 10% level, ** significant at the 1% level, a) These cases were inappropriate for calculating McNemar values because the patterns of the two comparative models were too similar for application. Thus, for these cases, we applied the test using binomial distribution as the substitute for McNemar's test.

As shown in Table 2, GAKNN is better than CKNN at the 1% statistical significance level. However, GAKNN outperforms ANN at the 10% statistical significance level.

5. Concluding Remarks

We have suggested a new kind of hybrid system of GA and CBR to improve the performance of the typical CBR system. This paper used GA as a tool for optimizing the feature weights as well as the number of neighbors that combine the *k* parameter in *k*-NN. From the results of the experiment, we show that GAKNN, our proposed model, outperforms all the comparative algorithms.

This study has some limitations. First, there are other factors that enhance the performance of a CBR system that may be incorporated with the simultaneous optimization model. For example, GA can be applied to relevant instance selection (see Kuncheva & Jain, 1999; Rozsypal & Kubat, 2003), which enables filtering of noisy or redundant reference cases. It may enhance the efficiency of our GAKNN model, and also refine the prediction results. Consequently, future research extending from this study maybe oriented to the multiple or simultaneous application of GA to a CBR system. Moreover, the general applicability of GAKNN should be tested further by applying it to other problem domains.

References

[1] Garrell i Guiu, J. M., E. Golobardes i Rib, E. Bernad i Mansilla and X. Llor i Fbrega, "Automatic diagnosis with genetic algorithms and case-based reasoning," Artificial Intelligence in Engineering, Vol. 13, pp. 367-372, 1999.

[2] Jarmulak, J., S. Craw and R. Rowe, "Self-optimizing CBR Retrieval," Proceedings of the 12th IEEE International Conference on Tools with Artificial Intelligence, pp. 376-383, 2000.

[3] Kim, K., "Toward global optimization of case-based reasoning systems for financial forecasting," Applied Intelligence, Vol. 21, pp. 239-249, 2004.

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

[5] Lee, H.Y. and K.N. Park, "Methods for Determining the optimal number of cases thatcombine in an effective case based forecasting system," Korean Journal of Management Research, Vol. 27, pp. 1239-1252, 1999.

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

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

[8] Shin, T. and I. Han, "Optimal signal multi-resolution by genetic algorithms to support artificial neural networks for exchange-rate forecasting," Expert Systems with Applications, Vol. 18, pp. 257-269, 2000.