PERGAMON

Expert Systems with Applications 22 (2002) 163–168

Expert Systems with Applications

www.elsevier.com/locate/eswa

A case-based customer classification approach for direct marketing

Chaochang Chiu*

Department of Information Management, Yuan Ze University, Neili 320, Taiwan, ROC

Abstract

Case-based reasoning (CBR) shows significant promise for improving the effectiveness of complex and unstructured decision making. CBR is both a paradigm for computer-based problem-solvers and a model of human cognition. However the design of appropriate case retrieval mechanisms is still challenging. This paper presents a genetic algorithm (GA)-based approach to enhance the case-matching process. A prototype GA–CBR system used to predict customer purchasing behavior is developed and tested with real cases provided by one worldwide insurance direct marketing company, Taiwan branch. The results demonstrate better prediction accuracy over the results from the regression-based CBR system. Also an optimization mechanism is integrated into the classification system to reveal those customers most likely and most unlikely customers to purchase insurance.

Keywords: Direct marketing; Case-based reasoning; Genetic algorithms; Customer classification

1. Introduction

Case-based reasoning (CBR) shows significant promise for improving the effectiveness of complex and unstructured decision making. It is a problem-solving technique that is similar to the decision making process used in many real world applications. CBR is both a paradigm for computer-based problem-solvers and a model of human cognition. The reasoning mechanism in the CBR system is based on the synergy of various case features. Therefore this method differs from a rule-based system because of its inductive nature. That is CBR systems reason using analogy concepts rather than the pure decision tree (or IF-THEN rules) usually adopted in rule-based systems.

Basically the CBR core steps are (1) retrieving past cases that resemble the current problem; (2) adapting past solutions to the current situation; (3) applying these adapted solutions and evaluating the results; and (4) updating the case base. Basically CBR systems make inferences using analogy to obtain similar experiences for solving problems. Similarity measurements between pairs of features play a central role in CBR (Kolodner, 1992). However the design of an appropriate case-matching process in the retrieval step is still challenging. Some CBR systems represent cases using features and employ a similarity function to measure the similarities between new and prior cases (Shin & Han, 1999). Several approaches have been presented to improve the case retrieval effectiveness. These include the parallel approach (Kolodner, 1988), goal-oriented model (Seifert, 1988), decision trees induction approach (Quinlan, 1986; Utgoff, 1989), domain semantics approach (Pazzani and Silverstein, 1991), instance-based learning algorithms (Aha, 1992), fuzzy logic method (Jeng & Liang, 1995), etc. These methods have been demonstrated effective in retrieval processes. However, most of these research works focused on the similarity function aspect rather than synergizing the matching results from individual case features. In essence when developing a CBR system, determining useful case features that are able to differentiate one case from others must be resolved first. Furthermore the weighting values used to determine the relevance of each selected feature has to be assigned before proceeding with the case matching process. Rather than being precisely or optimally constructed, the weighting values are usually determined using subjective judgment or a trial and errors basis. To provide an alternative solution this article presents a genetic algorithm (GA)-based approach to automatically construct the weights by learning the historical data. A prototype CBR system used to predict which customers are most likely to buy life insurance products is developed. The data provided by one worldwide insurance direct marketing subsidiary in Taiwan was used for constructing this model. The results show that the GA-based design of CBR system generates more accurate and consistent decisions than the regression-based CBR system.

* Fax: +886-3-435-2077.
E-mail address: imchiu@im.yzu.edu.tw (C. Chiu).

0957-4174/02/$ - see front matter
PII: S0957-4174(01)00052-5

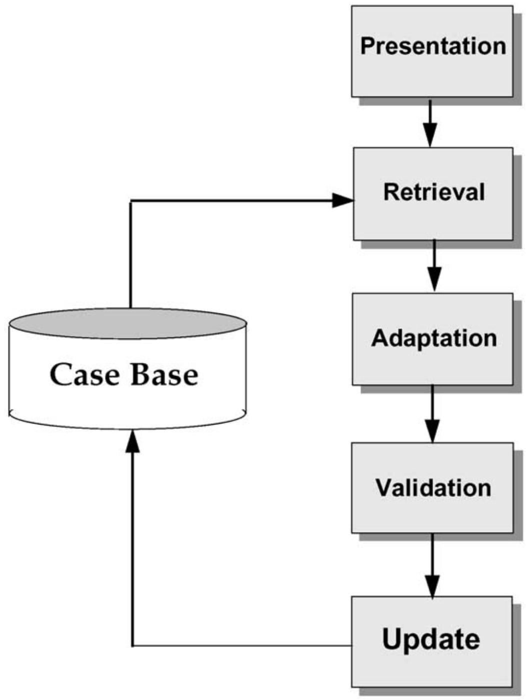

Fig. 1. The general CBR process.

2. An overview of CBR

Analogy, one way of human reasoning, is the inference that a certain resemblance implies further similarity. The CBR is a similar machine reasoning that adapts previous similar cases to solve new problems. It can be considered as a five-step reasoning process shown in Fig. 1 (Bradley, 1994).

Presentation: a description of the current problem is input into the system.
Retrieval: the system retrieves the closest-matching cases stored in a case base (i.e. a database of cases).
Adaptation: the system uses the current problem and closest-matching cases to generate a solution to the current problem.
Validation: the solution is validated through feedback from the user or the environment.
Update: if appropriate, the validated solution is added to the case base for use in future problem-solving.

Case retrieval searches the case base to select existing cases sharing significant features with the new case. Through the retrieval step, similar cases that are potentially useful to the current problem are retrieved from the case base. That is, previous experience can be recalled or adapted for the solution(s) to the current problem and mistakes made previously can be avoided.

The computing of the degree of similarity between the input case and the target case can usually be calculated using various similarity functions among which the nearest-neighbor matching is one of the frequently used methods.

2.1. Nearest-neighbor matching

Nearest-neighbor matching is a quite direct method that uses a numerical function to compute the degree of similarity. Usually, cases with higher degree of similarities are retrieved. A typical numerical function (Eq. (1)) is shown in the following formula (Kolodner, 1993).

$$\frac{\sum_{i=1}^{n} W_i * \text{sim}(f_i^{\text{I}}, f_i^{\text{R}})}{\sum_{i=1}^{n} W_i} \tag{1}$$

where W_i is the weight of the ith feature, f_i^{I} is the value of the ith feature for the input case, f_i^{R} is the value of the ith feature for the retrieved case, and sim() is the similarity function for f_i^{I} and f_i^{R}.

The implicit meaning of nearest-neighbor matching is that each feature of a case is a dimension in the search space. A new case can be added into the same space according to its feature values and the relative importance of the features. The nearest-neighbors identified can then be presented as similar cases. However, the matching process can only be executed with prior known weighting values as well as clearly defined similarity functions. Most of times weighting values are determined using human judgement, and thereby the retrieved solution(s) cannot always be guaranteed. Though Brieman et al. argued that the nearest-neighbor algorithms are sensitive to the similarity functions (Brieman, Friedman, Olshen, & Stone, 1984), the additional effects from weighting synergy could leverage the potential uncertainty. Wettschereck, Aha, and Mohri, (1997) organized feature weighting methods and summarized that feature weighting methods have a substantially higher learning rate than k-nearest-neighbor. Kohavi, Langley, and Yun, (1995) described the evidence that feature weighting methods lead to superior performance as compared to feature selection methods for tasks where some features are useful but less important than others. Though Kelly and Davis (1991) proposed a GA-based, weighted K-NN approach to attain lower error rates than the standard K-NN approach, seldom has other study focused on the non-linear feature value distance relationship between an old case and an input case. To overcome this shortcoming in the traditional case retrieval process, this study presents a GA approach to support the determination of the most appropriate weighting values for each case feature.

3. The genetic algorithm approach

GA is an optimization technique inspired by biological evolution (Holland, 1975). Its procedure can improve the search results by constantly trying various possible solutions