Applying Case-Based Reasoning in Supporting Strategy Decisions

S. M. Seyedhosseini, A. Makui, M. Ghadami

Abstract—Globalization and therefore increasing tight competition among companies, have resulted to increase the importance of making well-timed decision. Devising and employing effective strategies, that are flexible and adaptive to changing market, stand a greater chance of being effective in the long-term. In other side, a clear focus on managing the entire product lifecycle has emerged as critical areas for investment. Therefore, applying well-organized tools to employ past experience in new case, helps to make proper and managerial decisions. Case based reasoning (CBR) is based on a means of solving a new problem by using or adapting solutions to old problems. In this paper, an adapted CBR model with k-nearest neighbor (K-NN) is employed to provide suggestions for better decision making which are adopted for a given product in the middle of life phase. The set of solutions are weighted by CBR in the principle of group decision making. Wrapper approach of genetic algorithm is employed to generate optimal feature subsets. The dataset of the department store, including various products which are collected among two years, have been used. K-fold approach is used to evaluate the classification accuracy rate. Empirical results are collected among two years, have been used. K-fold approach is used to evaluate the classification accuracy rate. Empirical results are collected among two years, have been used. K-fold approach is used.

Keywords—Case based reasoning, Genetic algorithm, Group decision making, Product management.

I. INTRODUCTION

In the actual globally changing business environment, companies are seeking new ways of providing additional value to customers and gain a competitive edge over their competitors. Product design and a clear focus on managing the entire product lifecycle have emerged as critical areas for investment. Companies are focusing on total management of product lifecycles because today’s worldwide economic conditions demand they make process changes to remain competitive [1]. Product life cycle management (PLM) vision represents in fact the need to manage a large amount of product data that are generated in the various phases of life cycle, for supporting efficiency, flexibility and efficacy in the business processes applying innovative approaches [2].

Case Based Reasoning (CBR) is a reasoning methodology that reuses past cases to find a solution to the new problem and is preferred for ill-structured managerial decisions [7]. A case based reasoner uses remembered cases to find a solution to the new problem. It focuses on solving a new problem, to suggest how to adapt a solution that doesn’t work quickly, to warn of possible failures, to interpret a new situation, to critique a solution in progress, or to focus attention on some part of a situation or problem [8]. As CBR just refers to specific knowledge of previously experienced situations, it fits with complex and unstructured problems, and it is easy and convenient to update the knowledge base [9]. For these reasons, CBR has been popularly applied to management and engineering areas. CBR systems have been used in a wide variety of fields and applications [10].

S. M. Seyedhosseini, Industrial Engineering Department, Iran University of Science and Technology, Tehran, Iran(seyedhosseini@iust.ac.ir)
A. Makui, Industrial Engineering Department, Iran University of Science and Technology, Tehran, Iran (amakui@iust.ac.ir)
M. Ghadami, Industrial Engineering Department, Iran University of Science and Technology, Tehran, Iran (Corresponding author, phone: +982188900131; ghadami_maryam@mid.iust.ac.ir).
retrieval, statistics/robotics, data structures, software engineering, and process planning”. Furthermore, “Reference [12] suggested a classification method of CBR applications”. According to this classification scheme, CBR applications can be classified into two main categories, Classification and Synthesis tasks. In classification tasks, a new case is assigned to a specific class in the case-base from which a solution can be derived. Maintenance systems [13], engineering applications such as detecting locomotive faults [14], Legal and medical knowledge management and diagnosis [15], Product recommendation in e-commerce [16] and efficient helpdesks and customer support systems [17] are some examples of this category. Synthesis tasks, such as configuration, planning, and designing, attempt to get a new solution by combining previous solutions [12]. There are a variety of constraints during this process. Comparatively, they are harder to implement. CBR systems that perform synthesis tasks must make use of adaptation and are usually hybrid systems [18].

However, so far there has been no thorough research on applying case-based reasoning to support strategy decision making for product management. There has also been no reliable empirical research conducted to verify this approach and assess how useful it actually is in business practice [4]. At the same time global consulting companies have been building systems of databases containing information on the projects they have carried out with a view to adapting them for new clients. First applications of this class of systems in management include ORCA—a system supporting company restructuring processes during acquisitions and mergers, and ESAS—a system supporting business strategy planning processes. Both approaches mentioned above were promising prototypes [4].

For these reasons, the development of a strategy solution provider where it is possible to retrieve, use, manage data and information to obtain knowledge useful for supporting and taking decisions along the product life cycle is one important issue. Therefore case-based reasoning has been chosen as a suitable decision making paradigm to weight and rank cases in this step. This method is called nearest neighbor (NN) retrieval, i.e. nearest neighbor retrieval, and inductive retrieval, which the former is more widely used [21]. The system matches a new problem against cases in the case base using a specific retrieval method, and finds the most similar cases in this step. This method is called nearest neighbor (NN) matching [22]. The matching is realized with a similarity function. In NN matching, similar cases that are found affect the quality of the solution significantly, thus it is very important to design an effective retrieval method. The similarity between an input case and stored cases can be determined in many ways. It depends on the type of feature values. When cases are represented as feature vectors, calculating the weighted sum of feature distances is a common approach. The typical numerical function for NN matching is shown in (1) [22]:

II. BASIC CONCEPT

A. Case-Based Reasoning

CBR is an important paradigm of artificial intelligence mainly used for problem-solving [18]. It tends to apply efficient methods to define descriptive patterns and explanations within an enormous amount of data. The basic idea behind CBR is to solve a new problem by remembering and reusing information from a previous similar experience. When given a new case to be classified, a case-based reasoner searches from case base for similar cases and composes predictive result on the basis of class labels of similar cases to the current case. Numerous CBR model exists, however one of the more used is the cyclic process [20], known as the R4 model but it can be extended to the R5 model (if the preliminary step, case representation is included) (Fig.1)

The preliminary step consists in representing the past experiences contained in cases for the reasoning purpose. Many ways for case representations are possible, but the more used is a vector of feature–value pairs for the problem and solution descriptions. The initial step before applying the CBR process is to find relevant features for the problem and solution descriptions. After the filling of the target problem features, the next step in the cycle consists in retrieving the case or a subset of cases, stored in the case base, that are relevant to solve the target problem. Among well known methods, there are two chief methods to carry out case retrieval, i.e. nearest neighbor retrieval, and inductive approach, which the former is more widely used [21]. The system matches a new problem against cases in the case base using a specific retrieval method, and finds the most similar cases in this step. This method is called nearest neighbor (NN) matching [22]. The matching is realized with a similarity function. In NN matching, similar cases that are found affect the quality of the solution significantly, thus it is very important to design an effective retrieval method. The similarity between an input case and stored cases can be determined in many ways. It depends on the type of feature values. When cases are represented as feature vectors, calculating the weighted sum of feature distances is a common approach. The typical numerical function for NN matching is shown in (1) [22]:

![Fig. 1 The CBR process cycle](image-url)
attribute subsets. Some researchers have concluded that if the purpose of the model is to minimize the classifier error rate, and the measurement cost for all the features is equal, then the classifier’s predictive accuracy is the most important factor. In other words, the classifier should be constructed to achieve the highest classification accuracy. The features adopted by the classifier are then chosen as the optimal features. In the wrapper model, the meta-heuristic approaches are commonly employed to help in looking for the best feature subset. Although meta-heuristic approaches are slow, they obtain the (near) best feature subset [19].

III. PROPOSED CBR

This study applies an adapted CBR model which performs the principle of the group decision making to propose weighs of solutions in order of preference in the target case. Genetic algorithm is used as a wrapper approach to select a proper subset of features. The information encoded about the past experiences, depends on the domain of application as well as on the goal for which the cases are used. For our purpose, the cases are composed of two parts: the problem and the solution one. This problem is expressed by a limited set of features, \( P = \{p_1, p_2, ..., p_n\} \). \( P \) is a n-dimensional vector space, which means a decision making problem and \( p_i \) \( (i=1,2, ..., n) \) means an evaluation criterion of problem solution. \( S = \{s_1, s_2, ..., s_m\} \) is also a limited set, representing the solution set of case. \( S \) is an m-dimensional vector space which \( s_j \) \( (j=1,2, ..., m) \in [0,1] \) and \( \sum_{j=1}^{m} s_j = 1 \). For each problem, which is stored in the case base, solutions are compared and weighted in order the suitability for the given problem. For a new problem, the best cases will be retrieved from the case base. In this step, CBR use the k-nearest neighbor algorithm for total holdout data set. Generally, the technique of the nearest neighbor uses Euclidean distance, as follows:

\[
DIS_{xy} = \sqrt{\sum_{i=1}^{n} w_i (p_{xi} - p_{yi})^2} \quad i=1,2,...,k
\]

Where \( DIS_{xy} \) is a distance between \( x \) and \( y \), \( p_{xi} \) and \( p_{yi} \) are the values of case \( x \) and \( y \) on the \( i \)th feature, \( n \) is the number of features and \( w_i \) is the importance weighting of features \( x \), \( y \). By using the weighted distance defined in (3), a similarity measure between two cases, \( SM_{xy} \), can be defined as follows:

\[
SM_{xy} = \frac{1}{1 + DIS_{xy}}
\]

After calculating the similarity of each candidate case with the new case, \( k \) cases with the highest similarity are selected and returned to the user.

As previously mentioned, the goal of case reuse is to propose a solution to the target case, derived from solution(s) of the retrieved case(s), various methods to adapt a case exist. But in our study, source solutions of the \( k \) most similar problems are used to build and propose a solution to the target
problem. The adaption method consists in taking a weighting majority vote to select the most likely solution to the target problem, as follows [28]:

\[
\text{class} = \arg \max_v \sum_{j=1}^{|E|} w'_j \times I(v, c'_j) \tag{4}
\]

Where \( v \) is one of the class labels, \( |E| \) is the set of closest cases to the target case (z), \( c'_j \) is the class label of one of the nearest neighbors, \( I(p,q) \) is a function that returns the value 1 if \( p=q \) and 0 otherwise, and \( w'_i \), which is used to penalize neighbors that are located far away from the target case, is defined as follows:

\[
w' = \frac{1}{\text{DIS}^2_x} \quad ; \quad (x \in |E|) \tag{5}
\]

Thus \( w' \), weighted vector of the k most similar cases to the target case is expressed as follows:

\[
w' = \begin{bmatrix} w'_1 \\ w'_2 \\ \vdots \\ w'_k \end{bmatrix} \tag{6}
\]

Each of the k most similar problems (cases) has a set of solutions, which must be reused to the target problem. Different values have been assigned to every solution in the solution sets of the k similar problems. Therefore, for the target problem, there are k similar cases and each of them has m solutions’ value, which has been determined by the expert.

The weighted solution matrix is defined as follows:

\[
\begin{bmatrix}
w_{11} & w_{12} & \cdots & w_{1m} \\
w_{21} & w_{22} & \cdots & w_{2m} \\
\cdots & \cdots & \cdots & \cdots \\
w_{k1} & w_{k2} & \cdots & w_{km}
\end{bmatrix} \tag{7}
\]

Where \( w_{ij} \) (i=1,2, ..., k; j=1,2, ..., m, 0 < \( w_{ij} < 1 \)) is the value of solution \( i \)th in the case \( j \)th among the k most cases to the target case and \( \sum_{j=1}^{m} w_{ij} = 1 \). In this step the principles of group decision making method and weighting majority vote are applied in order to evaluate and adapt the solutions developed in selected cases. Therefore, ith solution’s value of the target problem, \( W_i \), is a group integrated value of solution \( i \) among k similar cases, which has a general form:

\[
W_i = \sum_{j=1}^{k} w'_j \times w_{ji} \quad (i = 1,2,\ldots,m) \tag{8}
\]

Consequently, the set of proposed solution to the target problem, which is acquired by group integrated value of solutions of the k-most similar cases is:

\[
W = \begin{bmatrix} W_1 \\ W_2 \\ \cdots \\ W_m \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{k} w'_{j1} \times w_{j1} \\ \sum_{j=1}^{k} w'_{j2} \times w_{j2} \\ \cdots \\ \sum_{j=1}^{k} w'_{jm} \times w_{jm} \end{bmatrix} \tag{9}
\]

The solution generated is then evaluated for validity. Finally, the completed new case is retained in the case base. The flowchart of the model is shown in Fig. 2. The detailed explanation for each step is presented as follows.

Step 1: The collected data is first input to the system and scaling is applied to prevent feature values in greater numeric ranges and to prevent numerical difficulties in the calculation. The scaled data is divided into two disjoint sets, known as the training set and test set, respectively.

Step 2: Initial population, which individually is comprised selected features, is generated.

Step 3: For each individual population, the training data is randomly split into the train and validation using cross validation, k-fold cross validation is conducted on the training set. Then the average validation accuracy of the k-fold cross validation, the fitness value, is calculated.

Step 4: if the termination criteria are satisfied, the process reruns CBR on the largest set to measure classification accuracy on the test set with selected features. Otherwise, the next iteration, forming a new population by genetic operation, occurs.
IV. EXPERIMENTAL RESULT

The data set is derived from the data of real department store in Iran over two years and has 95 records. Each record, representing information of a product, is composed of 24 features. The features fall into the three categories; customer information, product information and purchasing information. Customer information includes gender, marital status, children, yearly income, education, occupation and house status. Characteristics of the product itself, such as weight, volume, shelf width and height are included in the product information category. Also price, cost, units sales, frequency of sales and customers are associated to the category of purchasing information. Based on critical factors that are considered in each product, various strategies should be defined to manage products in the middle of life phase. Three strategies including more advertisement, changing prices and discontinuing of product selling, are selected to be considered as a solution set of each problem. The values of solutions, which are determined by the expert, reflect the importance of them in past problems. Therefore, there are 24 input variables, that describe product situation and 3 output variables, which are weighted to express the preference of each strategy in special product. The range of each feature value is scaled to the range of [-1, +1]. The k-fold approach is used to evaluate the classification accuracy rate [25]. This study set k as 4; that is, the data is divided into four portions. The data from each portion is formed based on the ratio of each category (classification) in the original data. Three portions of data are retrieved as training data and left one portion for testing each time. Since each portion is used as testing data once, four accuracy rates can be obtained. The final accuracy rate is the average of the four accuracy rates. The results obtained by the model are shown in Table. 1, that the accuracies of the model under 1NN-11NN are listed.

![Table I: Predictive Accuracy of CBR-GA in the Search of NN](image)

It is found that the accuracy the model does best in 5NN, which nine features are selected as input features, and the corresponding predictive accuracy is 0.86655 %.

In order to verify the performance of the model, 3 models are experimented for the same data set. The first model, CBR, does not have any special process of feature selection. CBR-PCA adopted the filter approach feature selection based on Principal Component Analysis (PCA). And the last one is Artificial Neural Network (ANN), a Multi-Layer Perceptron (MLP) neural network. Table II shows the performance of mentioned models. The paired t-test is used to examine whether the differences of predictive accuracy between CBR-GA and other models are statistically significant.

The results show that there are significantly different predictive performances between CBR-GA and two CBR algorithms (CBR, CBR-PCA) at least at the level of 1%. It is also found that there is no significant difference in predictive accuracy between CBR-GA and ANN. It can be interpreted that CBR-GA may improve the prediction accuracy of conventional CBR up to the accuracy of ANN.

![Table II: Predictive Performance of Various Models](image)

V. CONCLUSIONS

Business today is becoming more of a knowledge-based (KB) activity. PLM is widely recognized as a business necessity in current time. PLM enables companies to leverage their investments in product related intellectual and physical assets and is the vehicles to reduce cost, provide solid return on investment, and enable product and process innovation. Many of these activities involve making decisions by using past experience to select the appropriate choice from a set of possibilities. However, businesses would not automate their decision processes unless they could have enough confidence in the correctness of the solution produced. Case Based Reasoning (CBR) is a reasoning methodology that solves new problems by adapting previously successful solutions to similar problems. This paper applies CBR to provide suggestions for managing strategies of a given product for the middle of life phase. The set of solutions, as available strategies, are weighted by CBR in the principle of group decision making and reflect the importance of implementing each strategy in the product. The dataset of a department store is used to experiment the model. 24 features under 3 categories are selected to evaluate 95 products. The K-fold cross-validation is utilized to assess model and genetic algorithm is used to select features. Compared to other models such as Basic CBR and CBR-PCA, CBR-GA has the highest prediction accuracy with the dataset.

ACKNOWLEDGMENT

M. Ghadami gratefully thanks Mr. Mehdi Kariminasab for his useful comments.

REFERENCES


