A Hybrid Approach Using Case-based Reasoning and Genetic Algorithms for Corporate Bond Rating

April 1998

Kyung-shik Shin and Ingoo Han

Graduate School of Management

Korea Advanced Institute of Science and Technology

The author who is responsible for correspondences :

Professor, Ingoo Han

Graduate School of Management

Korea Advanced Institute of Science & Technology

207-43 Cheongryangri-Dong, Dongdaemoon-Gu

Seoul, 130-012, Korea

Tel : 02-958-3613, 958-3131, Fax : 02-958-3604

E-mail : ingoohan@msd.kaist.ac.kr

A Hybrid Approach Using Case-based Reasoning and Genetic Algorithms for Corporate Bond Rating

ABSTRACT

A critical issue in case-based reasoning (CBR) is to retrieve not just a similar past case but a usefully similar case to the problem. For this reason, the integration of domain knowledge into case indexing and retrieving process is highly recommended in building CBR system. However, this task is difficult to carry out since such knowledge often cannot be successfully and exhaustively captured and represented. This paper utilizes a hybrid approach using genetic algorithms (GAs) to case-based retrieval process in an attempt to increase the overall classification accuracy. We propose a machine learning approach using GAs to find an optimal or near optimal weight vector for the attributes of cases in the case indexing and retrieving. We apply this weight vector to the matching and ranking procedure of CBR. This GA-CBR integration reaps the benefits of both systems. Case-based reasoning technique provides analogical reasoning structures for experience-rich domain while GAs provide case-based reasoning with knowledge through machine learning. The proposed approach is demonstrated by applications to corporate bond rating.

Key Words :

Hybrid System, Case-based reasoning, Genetic algorithms, Corporate bond rating

I. INTRODUCTION

Corporate bond rating informs the public of the likelihood of an investor receiving the promised principal and interest payments associated with the bond issues. Bond ratings characterize the risk for the investments and affect the costs of borrowing for the issuer. Bonds are rated by independent rating agencies such as Moody's Investor Service and Standard & Poor. Numerous bond rating studies have traditionally used statistical techniques such as multiple discriminant analysis (Baran, Lakonishok & Ofer, 1980; Belkaoui, 1980; Pinches & Mingo, 1975), regression (Horrigan, 1996; Pogue & Soldofsky 1969; West, 1970), probit (Kaplan & Urwitz, 1979) and logit (Ederington, 1985) models.

Recently, however, a number of studies have demonstrated that artificial intelligence approaches such as neural networks (Dutta & Shekhar, 1988; Kwon, Han & Lee, 1997; Maher & Tarun, 1997; Singleton & Surkan, 1995), rule-based system (Kim & Lee, 1995) and case-based reasoning (Buta, 1994; Shin, Shin & Han, 1997) can be alternative methodologies for business classification problems.

Case-based reasoning (CBR) is a problem solving technique by re-using past cases and experiences to find the solution to the problems. While other major artificial intelligence techniques rely on making associations along generalized relationships between problem descriptors and conclusions, CBR is able to benefit from utilizing the specific knowledge of previously experienced, concrete problem situations. The central tasks that case-based reasoning methods have to deal with are to identify the current problem situation, find a past case similar to the new one, use that case to suggest a solution to the current problem, evaluate the proposed solution and update the system by learning from this experience (Kolodner, 1991,1993; Riesbeck & Schank, 1989; Slade, 1991).

CBR is preferred over rule based systems if rules are inadequate to express the richness of the domain knowledge. Brown and Gupta (1994) states that CBR seems best suited for domains that are experience-rich such as legal litigation, design, planning, and diagnosis. Wide range of applications of CBR have been reported (Brown & Gupta, 1994; Chi, Chen & Kiang, 1993; Hansen, Meservy & Wood, 1995; Morris, 1994; Mechitov, Moshkovich, Olson & Killingsworhth, 1995; O'Roarty, Patterson, McGreal & Adair, 1997; Riesbeck & Schank, 1989), including business classification such as bond rating and bankruptcy prediction (Bryant,1997; Buta,1994; Shin, Shin & Han, 1997).

The success of CBR system largely depends on effective retrieval of useful prior case for the problem (Hansen, Meservy & Wood, 1995; Jeng & Liang, 1995). The retrieval of cases that are just similar to the new situation along just any dimension is not enough. A good matching and retrieving function should take into account which features of a case are more important and scores cases for usefulness according to given tasks. A case that matches on important features but not on less important ones may be a better match than one that matches on less important features but does not match on important ones. For this reason, the integration of domain knowledge into case matching and retrieving process is highly recommended in building successful CBR system. However, this task is difficult to carry out since such knowledge often cannot be successfully and exhaustively captured and represented. In assignment of domain knowledge such as the importance weight of each attribute, Kolodner (1993) suggests several ways such as knowledge of human experts and statistical evaluation. However, even for experts, it is difficult to tell *a priori* which set of weights could be the most effective to solve a specific problem.

In this paper, we propose a hybrid approach using genetic algorithms (GAs) as an alternative methodology to assign the importance for case-based retrieving. We use GAs to extract knowledge that can guide effective retrieval of useful cases. Our particular interest lies in the assignment of importance values to each dimension of case features in the problem of corporate bond rating.

This paper is organized as follows. The following section provides a brief description of GAs. Section 3 describes the characteristics of indexing and retrieving methods of CBR. Section 4 describes our hybrid approach with GAs and CBR. Section 5 and 6 reports the empirical results of corporate bond rating application. Final section discusses the conclusions and future research issues.

II. GENETIC ALGORITHMS

GAs are stochastic search techniques that can search large and complicated spaces on the ideas from natural genetics and evolutionary principle (Davis,1991; Holland, 1975; Goldberg, 1989). They have been demonstrated to be effective and robust in searching very large spaces in a wide range of applications (Colin, 1994; Fogel, 1993; Han, Jo & Shin, 1997; Klimasauskas, 1992; Koza, 1993). GAs are particularly suitable for multi-parameter

optimization problems with an objective function subject to numerous hard and soft constraints. GAs perform the search process in four stage: initialization, selection, crossover, and mutation (Davis, 1991; Wong & Tan, 1994). Figure 1 shows the basic steps of genetic algorithms.

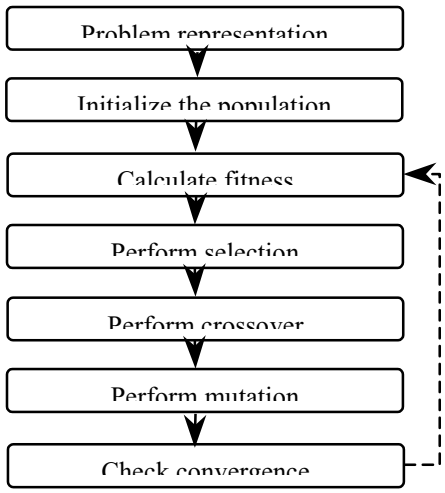

Figure 1. Basic steps of genetic algorithms

In the initialization stage, a population of genetic structures (called chromosomes) that are randomly distributed in the solution space, is selected as the starting point of the search. After the initialization stage, each chromosome is evaluated using a user-defined fitness function. The goal of the fitness function is to numerically encode the performance of the chromosome. For real-world applications of optimization methods such as GAs, the choice of the fitness function is the most critical step.

The mating convention for reproduction is such that only the high scoring members will preserve and propagate their worthy characteristics from generations to generation and thereby help in continuing the search for an optimal solution. The chromosomes with high performance may be chosen for replication several times whereas poor-performing structures may not be chosen at all. Such a selective process causes the best-performing chromosomes in the population to occupy an increasingly larger proportion of the population over time.

Crossover causes to form a new offspring between two randomly selected 'good parents'. Crossover operates by swapping corresponding segments of a string representation of the parents and extends the search for new solution in far-reaching direction. The crossover occurs only with some probability (the crossover rate). There are many different types of crossover that can be performed: the one-point, the two-point, and the uniform type

(Syswerda,1989).

Mutation is a GA mechanism where we randomly choose a member of the population and change one randomly chosen bit in its bit string representation. Although the reproduction and crossover produce many new strings, they do not introduce any new information into the population at the bit level. If the mutant member is feasible, it replaces the member which was mutated in the population. The presence of mutation ensures that the probability of reaching any point in the search space is never zero.

III. CASE INDEXING AND RETRIEVING

Case indexing involves assigning indices to cases to facilitate their retrieval. Indices organize and label cases so that appropriate cases can be found when needed. In building case-based reasoning systems, CBR community proposes several guidelines for choosing indexes for particular cases: (1) indexes should be predictiveness, (2) indexes should be abstract enough to make a case useful in a variety of future situation, (3) indexes should be concrete enough to be recognizable in future cases, and (4) prediction should be useful (Kolodner, 1991; 1993). Both manual and automated methods have been used to select indices. Choosing indices manually involves deciding the purpose of a case with respect to the aims of the reasoner and deciding under what circumstances the case may be useful.

The second issue of indexing cases is how to structure the indices so that the search through case library can be done efficiently and accurately. An index used to retrieve cases from memory may fail even if there is a relevant case in memory (Kolodner, 1991). This happens when the index does not correspond to the one used to index the case. The ‘indexing problem’ (Kolodner, 1991) refers to the task of storing cases for the effective and efficient retrieval.

There are three approaches to case indexing: nearest-neighbor, inductive, and knowledge-guided (Barletta, 1991). The nearest-neighbor approach let the user retrieve cases based on a weighted sum of features in the input cases that match the cases in memory. Every feature in the input cases is matched to its corresponding feature in the stored or old cases and the degree of match of each pair is computed based on the importance assigned to each dimension. This approach is preferred to use if the retrieval goal is not well-defined or if few cases are available. One of the major problems of the nearest-neighbor indexing approach is

that it is difficult to converge on a set of global-feature weights that could accurately retrieve cases in all situations (Barletta, 1991, Kolodner, 1993).

Induction is a technology that automatically extracts knowledge from training samples. Induction algorithms, such as ID3 and CART (Classification And Regression Trees), determine which features do the best job in discriminating cases, and generate a tree type structure to organize the cases in memory. An induction tree is built upon a database of training cases. The partitioning procedure of ID3 uses a preference criterion based on the information gain. At each node in the induction tree, the information gain is evaluated for all the attributes that are relevant and the one which yields the highest increase is selected. This approach is useful when a single case feature is required as a solution and where that case feature is dependent upon others.

Knowledge-guided indexing applies existing domain and experimental knowledge to locate relevant cases. Although this method is conceptually superior to the other two, knowledge-guided indexing is difficult to carry out since such knowledge often cannot be successfully and exhaustively captured and represented. Therefore, many systems use knowledge-guided indexing in conjunction with other indexing techniques (Barletta, 1991; Brown & Gupta, 1994).

IV. HYBRID APPROACH

4.1 Prior Research on the Hybrid Approach Using GAs

Hybridization of techniques can produce better systems if it ensures synergistic combination. GAs have been increasingly applied in conjunction with other AI techniques such as neural networks, rule-based system, fuzzy theory, and CBR. The integration of GAs and neural networks is a rapidly expanding area. The common problems faced by researchers and developers in using neural network techniques are optimization of input selection, network design and learning condition. Various problems of neural network design can be optimized using GAs (Wong & Tan, 1994). Examples include selecting relevant input variables, determining the optimal number of hidden layers, nodes and connectivity, and tuning the learning parameters (Bishop, Bushnell, Usher & Westland, 1993; Harp & Samad, 1991; Schffer, Whitley & Eshelman, 1992). Another approach of combining neural networks

and GAs is genetic training. GAs have been used to search the weight space of a neural network without the use of any gradient information (Montana & Davis, 1989; Ichikawa & Ishii, 1993; Deboeck, 1995).

GAs are also used in conjunction with fuzzy logic systems to provide an appropriate set of fuzzy IF-THEN rules for classification problem (Ishibuchi, Nozaki & Yamamoto, 1993) and to improve fuzzy logic controller (Karr, 1991, Park, Kandel & Langholz, 1994).

Few studies have dealt with hybridization of genetic and case-based reasoning, though there exist a great potential for useful applications in this area. Wang and Ishii (1997) applied GAs to the method of similarity metrics based on the cases being represented by structured representations.

4.2 Hybrid Systems with Case-based Reasoning and Genetic Algorithms

Integration of domain knowledge into case indexing and retrieving process is important in building a useful case-based system. The central idea of combination of GAs and case-based system is that CBR transfers the burden of knowledge assignment of the indexing and retrieving process to the searching and learning capabilities of evolutionary algorithms. In this subsection, we propose a hybrid approach using GAs to case indexing and retrieving process in an attempt to increase the overall effectiveness of the system.

4.2.1 Case indexing

In the development of CBR system, the choice of indexing and retrieving method is a very important step. As we described in the previous section, prior studies suggest some guidelines on choosing indexing technique. Barletta (1991) and Buta (1994) state that inductive indexing is more appropriate when the retrieval goal is well-defined while nearest-neighbor is preferred when the retrieval goal is subjective. In this aspect, inductive approaches are suitable for business classification tasks, such as corporate bond rating and bankruptcy prediction. However, inductive indexing requires large number of cases to form induction trees (Barletta, 1991). In addition, finding and maintaining an optimal induction tree for case-based retrieval is an expensive task (Kolodner, 1993).

We employ nearest-neighbor approach for the current study. The nearest-neighbor approach, however, has difficulty in deciding a set of feature weights that could accurately

retrieve cases in a given situation (Barletta, 1991). Since the feature weights for most problem domains are context dependent, each case should have its own set of feature weights for determining the relevance of that case to a new problem. So the accuracy of the system largely depends on the matching functions which include the weight vectors specified by developer.

4.2.2 Nearest-neighbor matching

Kolodner (1993) states that the importance of a dimension in judging similarity and degree of match should be considered in building matching functions. Matching and ranking is the process of comparing two cases with each other and determining their degree of match and ordering cases according to the goodness of match or the usefulness for the application (Kolodner, 1993). A good matching function takes into account which features of a case are more important and scores cases for the usefulness according to those criteria. A case that matches on important features but not on less important ones needs to be judged a better match than one that matches on less important features but does not match on important ones.

One of the most obvious measures of similarity between two cases is the distance. This study uses a numeric evaluation function measuring the distance taking into account of importance features to compute the degree of match in retrieval. A matching function of nearest-neighbor method is as follows:

$$DIS_{ab} = \sqrt{\sum_{i=1}^{n} w_i \times (f_{ai} - f_{bi})^2}$$

where *DIS* is the matching function using Euclidean distance between cases, n is the number of features, and w_i is the importance weighting of a feature i . Basic steps of nearest-neighbor retrieval algorithms are quite simple and straightforward. Every feature in the input case is matched to its corresponding feature in the stored case, and the degree of match of each pair is computed using matching function. Based on the importance assigned to each dimension, an aggregate match score is computed. Ranking procedures order cases according to their scores, and higher scoring cases are used before lower scoring ones. Figure 1 represents the nearest-neighbor matching algorithm.

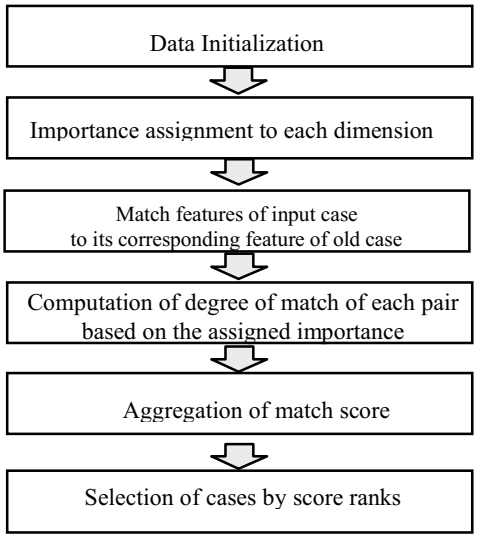

Figure 1. Nearest-neighbor matching algorithm

4.2.3 Assigning Importance Values to Case Attributes

The importance associated with each field tells us how much attention to pay to the match. Although Kolodner (1993) suggests several ways of assigning the importance values such as knowledge of human experts and statistical evaluation, it is difficult to tell *a priori* which set of weight could be the most effective to solve a specific problem. Considering a function that computes the degree of match can only be as good as the knowledge it has of the importance of dimensions, it is important an task to find an optimal set of weights.

As Kolodner's suggests, one way to assign importance values is to have a human expert assign them as the case library is being built. The expert is expected to have knowledge and experience for deciding which dimensions make good predictors.

Another way to assign importance values is to do a statistical evaluation of known cases to determine which dimensions predict the solutions best. The more significant predictors in the statistical evaluations are considered to be of importance for matching. For example, the magnitude of the correlation coefficient between each input and the output in the reference set can be used to weigh each input when computing the distance measure for a new example.

As an alternative approach, we introduce the notion of machine learning to learn the optimal weights from historical cases using evolutionary search technique. By evaluating the fitness of different weight vectors, we may find good solutions for the system. As we described in the above section, GAs apply crossover and mutation to generate a new population of problem solutions and select the best solution for the problem.

4.2.4 Hybrid Structure of GA-CBR System

Overall structure of hybrid approach is shown in Figure 2.

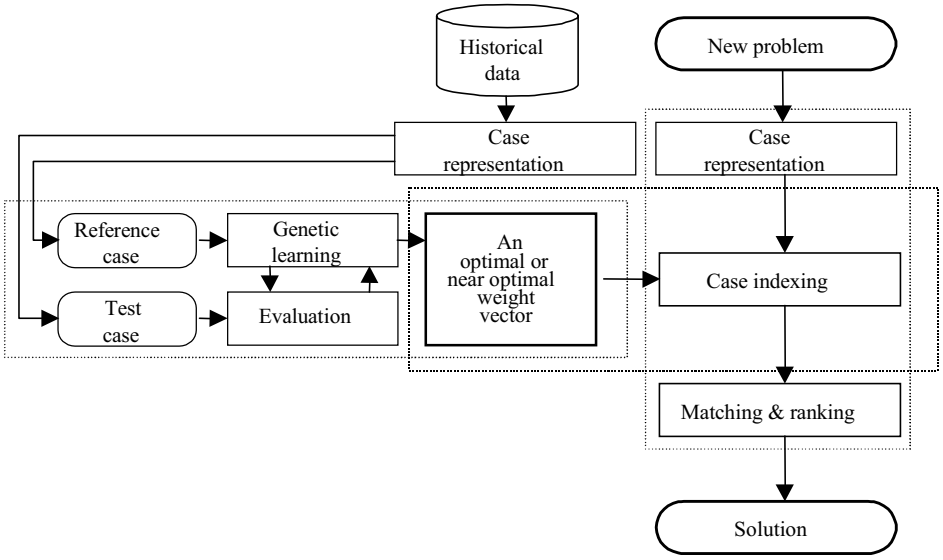

Figure 2. Hybrid structure of GA-CBR system

Phase 1

For the first step, we search an optimal or near optimal weight vector with precedent cases for which the classification outcome has been determined. In setting up the genetic optimization problem, we need:

1. The parameters that have to be coded for the problem
2. An objective or fitness function to evaluate the performance of each string
3. The ranges for the adjustable parameters
4. Potential constraints to be met

The parameters that are coded are the weight vectors for nearest-neighbor matching. These weight vectors assign importance values to each feature and are inserted into the matching formulae. We set the range of the weights between 0 to 1 and do not apply particular constraints for this search.

The task of defining a fitness function is always application specific. In this case, the objective of the system is to retrieve more relevant cases that can lead to the correct solutions. The ability of case-based system to achieve these objectives can be represented by the fitness function that specifies how well the matching function increase the classification accuracy. We apply the classification accuracy rate of test set to the fitness function for this study. The test set consists of known cases of which the classification outcome has been determined and

is used to evaluate fitness of different set of feature weights. Mathematically, this fitness function is expressed as:

$$\text{Maximize} \qquad CR = \frac{1}{n}\sum_{i=1}^{n} CA_i$$

$$\text{s.t.} \qquad CA_i = \begin{cases} 1 & \text{if } O(T_i) = O(S_{j*(i)}) \\ 0 & \text{otherwise} \end{cases}$$

$$S_{j*(i)} = \underset{j \in R}{\text{Min}} \left(\sqrt{\sum_{k=1}^{l} w_k (T_{ik} - R_{jk})^2} \right)$$

for given i ($i=1,2,\ldots\ldots,n$)

where

CR = the classification accuracy rate of test set

CA_i = the classification accuracy of i th case of test set denoted by 1 and 0 ('correct' = 1, 'incorrect' = 0). For example, the bond rating of i th case of test set and the closest case in the reference set are 'A1' and 'A1', respectively, CA_i is 1, otherwise 0)

$O(T_i)$ = the target output of i th case of test set

$O(S_{j*(i)})$ = the output of of j th case of reference set that has the minimum distance with i th case of test set

$S_{j(i)}$ = the distance between i th case of test set and the j the case in reference set

T_{ik} = the k th feature of the i th case of test set (T)

R_{jk} = the k th feature of the j th case of reference set (R)

w_k = the importance weight of the k th feature of case

l = the number of features

n = the number of test cases

For the controlling parameters for experiment, we should also specify:

1. Population size
2. Crossover rate
3. Mutation rate
4. The criteria for stopping the process

There has been much debate regarding the optimal population size for the problem. Generally, they determine the population size depending on the size of the problem (bigger population for larger problem). The common view is that a larger population takes longer to settle on a solution, but is more likely to find a global optimum because of its more diverse gene pool (*Evolver* manual, 1995). We use 50 organisms in the population for this study. The

crossover and mutation rates are changed to prevent the output from falling into the local optima. The crossover rate ranges 0.5 - 0.7 and the mutation rate ranges 0.06 - 0.1 for this experiment. As a stopping condition, we use 2,500 trials. These processes are done by the genetic algorithms software package Evolver 3.5, called from an Excel macro.

Phase 2

The second step is to apply derived weight vector of *phase 1* to case indexing scheme for case-based retrieval process and evaluate the resulting model with additional validation case for which the outcome is also known. A weight vector is used in nearest-neighbor matching function to rank and retrieve useful cases. Since the validation cases are not used for parameter optimization, the prediction performance tested by these cases would be the closest to the current or future cases. If the project is successful, this leads to the production.

Phase 3

Phase 3 is a production phase. In production phase, new (unclassified) data is presented to the model to solve the problem.

V. DATA AND VARIABLES

The research data consists of 168 financial ratios and corresponding bond ratings of Korean companies. The ratings have been performed by National Information and Credit Evaluation, Inc., one of the most prominent bond rating agencies in Korea. The total sample available includes 3,886 companies whose commercial papers have been rated from 1991 to 1995. Credit grades are defined as outputs and classified as 5 grade groups (A1, A2, A3, B, C) according to credit levels. Table 1 shows the organization of data set.

Table 1. Number of companies in each rating

Ratings	Number of cases	%
A1	260	6.7
A2	833	21.4
A3	1,314	33.8
B	1,406	36.2
C	73	1.9
Sum	3,886	100.0

Table 2. Definition of variables

Variable	Definitions
X1	Firm classification by group (conglomerate) types
X2	Firm types
X3	Total assets
X4	Stockholders' equity
X5	Sales
X6	Year after founded
X7	Gross profit to sales
X8	Net cash flow to total assets
X9	Financial expenses to sales
X10	Total liability to total asset
X11	Depreciation to total expenses
X12	Working capital turnover

We apply two stages of input variable selection process. At the first stage, we select 27 variables (23 quantitative / 4 qualitative) by factor analysis, 1-way ANOVA (between input variable and credit grade as output variable) and Kruskal-Wallis test (for qualitative variables). In the second stage, we select 12 financial variables (10 quantitative / 2 qualitative) using stepwise method to reduce the dimensionality. The aim of input variable selection approach is to select the input variables satisfying the univariate test first, and then select significant variables by stepwise method for refinement. In choosing qualitative variables, the four variables have been initially selected. However, audit opinion and audit firm are excluded by the expert's opinion. Two selected qualitative variables are firm classification by group (conglomerate) types and firm types. We classify conglomerates into five categories include top-ten conglomerates, top-twenty conglomerates, top-thirty conglomerates, top-forty conglomerates and non-conglomerates. Four types of firms are listed, registered, externally audited, and others. Table 2 illustrates selected variables for this study.

Each data set is split into three subsets, a reference set, a test set and a validation (holdout) set of 80, 10, 10 percent of data, respectively. The reference data are used to learn an optimal weight vector in genetic learning and also used as a case base for retrieval. The test data are used to evaluate the weight vectors, verifying how well the indexing of CBR system is working. The validation data are used to test the results with the data which have not been used to develop the system.

VI. RESULTS

As mentioned above, we have two ways to assign importance values on attributes: one way is to have a human expert assign them, the other is to do a statistical evaluation of a known cases to determine the importance weights.

For this experiment, we have experts designate the importance of an attribute by assigning 5 levels of qualitative values by interview. We have selected 5 experts, three from bond rating department of a credit rating agency and two from credit analysis department of a commercial bank in Korea. The work experience of selected experts related to credit analysis ranged from 2 to 8 and half years and the average of experience is 4 years and 2 months. Five qualitative values are "most important," "very important," "important," "less important," and "ignored" which are associated with numbers for computation as 1.0, 0.8, 0.6, 0.4, and 0.2,

respectively. Table 3 shows the assigned importance to each attribute by experts. The importance values range from 0.4 to 0.88.

Table 3. Importance weights assigned by experts

Variable	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12
Average assigned value by experts	0.72	0.40	0.80	0.88	0.80	0.72	0.68	0.72	0.76	0.80	0.48	0.68

As another way of importance assignment method, we have done statistical evaluations. Since there is no formal statistical method to assign importance for effective case retrieval, our experiment should be explorative. As importance measures, we have derived three values using reference data set: a magnitude of the Pearson's correlation coefficient between the input and output in the reference set, the absolute value of coefficient estimate obtained by linear regression, and the partial F values of independent variables derived by multiple discriminant analysis (MDA).

Considering that ratings are not continuous values, we cannot apply correlation and regression analysis for the context. However, since ratings are ordinal in nature, ratings are considered as continuous values ranged from 1 to 5. Another limitation of these analyses is that these weight vectors do not consider the correlation among the features.

MDA is a representative of statistical classification methods, and the linear discriminant function is as follow.

$$Z = W_1 X_1 + W_2 X_2 + \Lambda + W_n X_n$$

where Z is a discriminant score, W_i is a discriminant weight for variable i, X_i is an independent variable i. The interpretation of MDA involves examining the discriminant functions to determine the relative importance of each independent variable in discriminating between the groups.

Three methods of determining the relative importance have been proposed: standardized discriminant weights, discriminant loadings (structure correlation) and partial F values (Hair, Anderson, Tatham & Black, 1995). Independent variables with relatively larger weights contribute more to the discriminating power of the function than do variables with smaller weights. However, we do not use discriminant weights for two reasons. First, since

the multiple discriminant analysis generates multiple discriminant functions, additional problems of interpretation are required, that is, we should represent the impact of each variable across the functions. Another reason is that discriminant weights are subject to considerable instability. Discriminant loadings measure the simple linear correlation between each independent variable and the discriminant function. Although discriminant loadings are considered more valid than weights as a means of interpreting the discriminating power of independent variables because of their correlational nature, these are subject to instability like weights (Hair, Anderson, Tatham & Black, 1995).

We use the partial F values for this study as a mean of interpreting the relative discriminating power of the independent variables. This is accomplished by examining the absolute sizes of the significant F values and ranking them. Large F values indicate greater discriminating power. The values obtained from statistic analysis are linearly transformed to 0 to 1 scale using min-max values. Table 4 shows the assigned importance to each attribute by statistical evaluations.

Table 4. Importance values assigned by statistical evaluations

Variables	Pearson's coefficient	Regression coefficient	Partial F values / MDA
X1	0.99	0.99	0.18
X2	0.95	0.65	0.68
X3	0.69	0.16	0.01
X4	0.89	0.01	0.61
X5	0.46	0.07	0.09
X6	0.20	0.20	0.63
X7	0.40	0.28	0.25
X8	0.01	0.09	0.29
X9	0.91	0.66	0.69
X10	0.23	0.01	0.99
X11	0.51	0.56	0.52
X12	0.60	0.59	0.05

Among variables, firm classification by conglomerate types shows higher importance than other variables in respect to correlation and regression analysis. Total liability to total asset is selected as the most important variable in discriminating ratings in the view of F values.

To study the effectiveness of hybrid approach using GAs and CBR for the corporate bond rating problem, predictive performance of case-based retrieval using weight vectors obtained by various methods including expert opinions and statistical evaluations are compared. Experimental results are summarized in Table 5.

Table 5. Classification accuracy rates (%)

Methods		A1	A2	A3	B	C	Total
MDA		57.7	69.8	58.3	55.0	77.8	60.0
Induction		65.4	55.8	47.5	72.9	33.3	59.0
CBR							
Pure		65.4	66.3	58.3	66.4	0.0	62.0
Expert		69.2	65.1	58.3	63.6	0.0	61.0
Statistical evaluations	Pearson's coefficient	69.2	65.1	59.7	68.6	11.1	63.5
	Regression coefficient	57.7	69.8	52.5	68.6	22.2	61.5
	Partial F values / MDA	61.5	55.8	51.1	63.6	0.0	56.0
GA-CBR		76.9	76.7	74.8	78.6	22.2	75.5

Among the techniques of importance assignment, the hybrid approach has the highest level of accuracy (75.5%) in the given data sets, followed by statistical evaluations using Pearson's coefficients (63.5%), pure retrieval with equal weights (62.0%), and expert's opinion (61.0%).

We have expected that nearest-neighbor retrieval using weight vectors derived by experts' opinion and statistical evaluations would show the better performance than the pure one that has equal weights among variables. Among statistical evaluation methods, we have expected that multiple discriminant analysis using partial F values would show the best results since the technique considers both relationships between independent variable and dependent variable and among independent variables.

However, the classification accuracies using weight vectors assigned by experts' opinion, regression coefficients and partial F values derived by MDA are inferior to that of the pure CBR model. The results support the argument that the weight vector which has equal weights may be appropriate when no relevant information or tools are available. It may be safe not to assign weights inappropriately, i.e., assign heavy weights on less important attributes (Lee, 1992).

The methods used to extract importance weights from experts and statistical evaluations and assign them to the features may be inappropriate for the context. These methods may not be an optimal one, and perhaps, there are better methods both in capturing expert knowledge and in statistical evaluations.

It is very difficult to find an optimal or near optimal weight vector using experts' opinion or statistical evaluations. Obtained knowledge using those two methods can hardly fit case-based system, because assumptions, algorithms and objective functions applied to those

methods fundamentally differ from nearest-neighbor matching algorithm. This is further supported by the results that the model using partial F values of multiple discriminant analysis show the lowest classification accuracies among the methods. In the statistical perspective, these values are more appropriate to apply since the other two values do not consider multicollinearity problem.

Based on the empirical results, we conclude that the method to assign weights derived from genetic search process can be the most effective one since GAs find optimal or near optimal solution for the specified objective function.

The results of other classification techniques, such as MDA and ID3, are also presented in Table 5 as benchmarks to verify the applicability to the context. As we mentioned above, MDA is a very popular classification method, based on a bayesian method using classification rules to minimize the expected misclassification cost. ID3 extracts knowledge from training samples, generating a tree type structure to organize the cases in memory (Han, Chandler & Liang, 1996). In comparison with CBR, induction technique makes direct use of past experience at the problem solving stage while CBR only uses an abstraction of the cases. Among the classification techniques, the GA-CBR hybrid approach also outperforms MDA (60.0%) and ID3 (59.0%), ensuring the practical applicability to the domain.

We use the McNemar tests to examine whether the predictive performance of hybrid approach is significantly higher than that of other techniques. McNemar test is a nonparametric test of the hypothesis that two related dichotomous variables have the same means. This test is useful for detecting changes in responses due to experimental intervention in 'before and after' designs using the chi-square distribution. Since we are interested in the correct prediction of cases, the measure for testing is the classification accuracy rate (the number of correct classification from the number of whole holdout samples). Table 6 shows the results of McNemar tests to compare the classification ability between benchmark models and a hybrid model using GAs for holdout samples.

As shown in the table 6, GA-CBR hybrid model has the higher prediction accuracy than the individual method of MDA, ID3, and CBR models with different importance measures. The result shows the GA-CBR model performs significantly better than every benchmark model proposed for this study at a 1% level. Also, the table 6 shows that MDA-assisted model is less useful than other statistical evaluation methods to assign importance.

Table 6. McNemar values for the pairwise comparison of performance between models

	ID3	CBR-Pure	CBR-Expert	CPR-PC	CBR-RC	CBR-F/M	GA-CBR
MDA	0.0500[d]	0.2988	0.0529	1.1419	0.1563	1.2784 *	22.1488 ***
ID3	-	0.7857	0.3182	1.7622	0.5473	0.7469	27.4351 ***
CBR-Pure	-	-	0.1000	0.2717	0.0079	5.0865 **	25.5364 ***
CBR-Expert	-	-	-	0.6639	0.0068	2.7769 *	27.0750 ***
CBR-PC[a]	-	-	-	-	0.6282	6.5703 **	21.6569 ***
CBR-RC[b]	-	-	-	-	-	3.2910 *	26.5351 ***
CBR-F/M[c]	-	-	-	-	-	-	39.5267 ***

[a.] Nearest-neighbor retrieval using importance weights assigned by Pearson's coefficient
[b.] Nearest-neighbor retrieval using importance weights assigned by regression coefficient
[c.] Nearest-neighbor retrieval using importance weights assigned by partial F-values of MDA
[d.] Chi-square value
* significant at 10% / ** significant at 5% / *** significant at 1%

We compared our results of experiment with those of past studies which used Korean bond rating data. Table 7 illustrates the results of previous research.

Table 7. The results of previous Korean bond rating studies compared with this study

	Year of data	Number of data	Method	Accuracy (%)
Kwon *et al*. (1997)	1991 - 1993	3,085	MDA	58.4 – 61.6
			CNN[a]	66.5 – 67.3
			OPP[b]	70.9 – 73.3
Shin *et al*. (1997)	1991 - 1995	2,651	MDA	58.6
			ID3	53 – 54
			CBR-Pure[c]	57 – 63
			Induction-CBR	54 - 63
Shin *et al*. (1998)	1991 - 1995	3,886	MDA	60.0
			ID3	59.0
			CBR-Pure	62.0
			CBR-Expert	61.0
			CBR-Statistics	56.0 - 63.5
			GA-CBR	75.5

[a.] conventional back-propagation neural networks
[b.] ordinal pair-wise partitioning method applied to back-propagation neural networks training
[c.] nearest-neighbor retrieval using equal weights among attributes

Kwon *et al.* (1997) used ordinal pairwise partitioning (OPP) approaches to back-propagation neural networks training for corporate bond rating prediction. The main idea of the OPP approach is to partition the data set in the ordinal and pairwise manner into the

output classes. Experimental results show that the OPP approach has the highest level of accuracy (71%-73%), followed by conventional neural networks (66%-67%) and multiple discriminant analysis (MDA) (58%-61%).

Another bond rating research performed by Shin *et al*. (1997) used inductive approach to case indexing for CBR system. Despite the optimistic hope that case indexing using induction technique might improve the effectiveness of case reasoning resulting higher classification accuracy, the experimental results were rather disappointing. The performance of CBR models with inductive indexing (55%-63%) were lower than that of the pure CBR model (57%-63%) which used the nearest-neighbor indexing with equal weights among attributes. They have concluded that the reason of failure is that the induction using certain partitioning criteria does not ensure optimal decision trees to retrieve useful prior cases that are the best for the job.

VII. CONCLDING REMARKS

Both CBR and GAs are artificial intelligent approaches that have received significant attention in recent years. In this paper, we have proposed a hybrid approach using CBR and GAs to the problem of corporate bond rating. In this approach, the genetic search technique is used to assign relative importance of feature weights for case indexing and retrieving. We have shown that this approach supports effective retrieval of cases and increases overall classification accuracy rate significantly.

These results support the following findings. First, the knowledge acquired by problem domain supports the retrieval of usefully similar case to solve the problem. Since the task of GAs to define a fitness function is always domain specific, this hybrid approach utilizes domain knowledge and case specific knowledge simultaneously. Second, GAs are effective method of knowledge extraction for case-based retrieval. Using GAs, we can obtain the near optimal weights representing the importance of each feature.

This study has some limitations. First, our statistical benchmarks are not optimal ones for the problem. This is, however, partly due to the fact that there is no formal statistical method to assign importance for effective case retrieval, The second limitation is on the appropriateness of importance assignment method using experts' opinion. We found that capturing experts' knowledge and assigning to the model are by no means an easy task.

Finally, this study only focuses on the feature weights to optimize for CBR system. However, a GA approach could potentially be used to optimize other specific points of the case-based reasoning process. We believe the potential is great for further research with hybrid approaches using GAs and also different intelligent techniques as ways to improve the performance for the applications.

REFERENCES

1) Baran, A., Lakonishok, J. & Ofer, A. R., "The value of general price level adjusted data to bond rating," *Journal of Business Finance and Accounting*, Vol.7, No.1 (1980), 135-149.
2) Barletta, R., "An introduction to case-based reasoning," *AI EXPERT,* Vol.6, No.8 (1991), 42-49.
3) Belkaoui, A., "Industrial bond ratings: a new look," *Financial Management,* Vol.9, No.3 (1980), 44-51.
4) Bishop, J.M., Bushnell, M.J., Usher, A. & Westland, S., "Genetic optimization of neural network architectures for color recipe prediction," *Artificial Neural Networks and Genetic Algorithms*, Springer-Verlag, New York, (1993), 719-725.
5) Brown, C.E. & Gupta, U.G., "Applying case-based reasoning to the accounting domain," *Intelligent Systems in Accounting, Finance and Management,* Vol.3 (1994), 205-221.
6) Bryant, S.M., "A case-based reasoning approach to bankruptcy prediction modeling," *Intelligent Systems in Accounting, Finance and Management,* Vol. 6 (1997), 195-214.
7) Buta, P., "Mining for Financial Knowledge with CBR," *AI EXPERT*, Vol.9, No.2 (1994), 34-41.
8) Chi, R.T., Chen, M. & Kiang, M.Y., "Generalized case-based reasoning system for portfolio management," *Expert Systems with Applications,* Vol.6, No.1 (1993), 67-76.
9) Colin, A.M., "Genetic algorithms for financial modeling," In Deboeck, G.J. (Eds.), *Trading On The Edge.* New York: John Wiley, 1994, 148-173.
10) Davis, L., *Handbook of genetic algorithms.* Van Nostrand Reinhold, NY., 1991.
11) Deboeck, G.J., "Using GAs to optimize a trading system," In Deboeck, G.J (Eds.), *Trading On The Edge.* New York: John Wiley, 1995, 174-188.
12) Dutta, S. & Shekhar, S., "Bond rating: a non-conservative application of neural networks," *Proc. of IEEE International Conference on Neural Networks,* Vol.2. San Diego, CA, 1988.
13) Ederington, H.L., "Classification models and bond ratings," *Financial Review*, Vol.20, No.4 (1985).
14) *Evolver*^{TM} Manual, 1995.
15) Fogel, D.B., "Applying evolutionary programming to selected traveling salesman problems," *Cybernetics and Systems*, Vol.24, No.1 (1993).
16) Goldberg, D.E., *Genetic Algorithms in Search, Optimization and Machine Learning.* MA : Addison-Wesley, 1989.
17) Hair, J.F., Anderson, R.E., Tatham, R.L. & Black, W.C., *Multivariate Data Analysis with Readings*, 4th ed. NJ : Prentice Hall, 1995.
18) Han, I., Chandler, J.S. & Liang, T.P., "The impact of measurement scale and correlation structure on classification performance of inductive learning and statistical methods," *Expert System With Applications*, Vol.10, No.2 (1996), 209-221.
19) Han, I., Jo, H. & Shin, K.S., "The hybrid systems for credit rating," *Journal of the Korean Operations Research and Management Science Society*, Vol.22, No.3 (1997), 163-173.
20) Hansen, J., Meservy, R.D. & Wood, L.E., "Case-based reasoning: application techniques for decision support," *Intelligent Systems in Accounting, Finance and Management,* Vol.4 (1995), 137-146.
21) Harp, S.A. & Samad, T., "The genetic synthesis of neural network architectures," In Davis, L. (Eds.), *Handbook of Genetic Algorithms.* Van Nostrand Reinhold, New York, 1991, 202-221.
22) Holland, J.H., *Adaptation in natural and artificial systems.* Ann Arbor: The University of Michigan Press, 1975.
23) Horrigan, J.O., "The determination of long term credit standing with financial ratios," *Journal of Accounting Research,* supplement (1996), 44-62.
24) Ichikawa, Y. & Ishii, Y., "Retaining diversity of genetic algorithms for multivariable optimization and neural network learning," *Proc. of IEEE International Conference on Neural Networks,* Vol.2. San Francisco, 1993, 1110-1114.

25) Ishibuchi, H., Nozaki, K. & Yamamoto, N., "Selecting fuzzy rules by genetic algorithm for classification," *Proc. of IEEE International Conference on Fuzzy Systems,* Vol.2. San Francisco, 1993, 1119-1124.
26) Jeng B.C. & Liang, T.P., "Fuzzy indexing and retrieval in case-based systems," *Expert Systems with Applications,* Vol.8, No.1 (1995), 135-142.
27) Kaplan, R.S. & Urwitz, G., "Statistical models of bond ratings: a methodological inquiry," *Journal of Business,* Vol.52, No.2 (1979), 231-262.
28) Karr, C., "Genetic algorithms for fuzzy controllers," *AI Expert*, Vol.6, No.2 (1991), 26-33.
29) Klimasauskas, C.C., "Hybrid neuro-genetic approach to trading algorithms," *Advanced Technology for Developers*, Vol.1, No.7 (1992).
30) Kim, B.O. & Lee, S.M., "A bond rating expert system for industrial companies," *Expert Systems with Applications,* Vol.9, No.1 (1995), 63-70.
31) Kolodner, J., "Improving human decision making through case-based decision aiding," *AI Magazine,* Vol.12, No 2 (1991), 52-68.
32) Kolodner, J., *Case-Based Reasoning,* Morgan Kaufmann, San Mateo, CA, 1993.
33) Koza, J., *Genetic programming*, Cambridge: The MIT Press, 1993.
34) Kwon, Y.S., Han, I.G. & Lee, K.C., "Ordinal pairwise partitioning (OPP) approach to neural networks training in bond rating," *Intelligent Systems in Accounting, Finance and Management,* Vol. 6 . (1997), 23-40.
35) Lee, H.Y., *Predictive insights through analogical reasoning: application to screening new financial service concept.* Ph.D. Thesis, The Wharton School, University of Pennsylvania, 1992.
36) Maher J.J. & Tarun, K.S., "Predicting bond ratings using neural networks: a comparison with logistic regression," *Intelligent Systems in Accounting, Finance and Management,* Vol.6 (1997), 59-72.
37) Mechitov, A.I., Moshkovich, H.M., Olson, D.L. & Killingsworth, B., "Knowledge acquisition tool for case-based reasoning system," *Expert Systems with Applications,* Vol.9, No.2 (1995), 201-212.
38) Montana, D.J. & Davis, L., *Training feedforward networks using genetic algorithms.* Proc. of 11th Int. Joint Conference on Artificial Intelligence. Detroit, 1989, 762-767.
39) Morris, B.W., "SCAN: a case-based reasoning model for generating information system control recommendations," *Intelligent Systems in Accounting, Finance and Management,* Vol. 3 (1994), 47-63.
40) O'Roarty, B., Patterson, D., McGreal, S. & Adair, A., "A case-based reasoning approach to the selection of comparable evidence of retail rent determination," *Expert Systems with Applications,* Vol.12, No.4 (1997), 417-428.
41) Park, D., Kandel, A. & Langholz, G., "Genetic-based new fuzzy reasoning models with application to fuzzy control," *IEEE Transactions on Systems, Man and Cybernetics*, Vol.24, No.1 (1994), 39-47.
42) Pinches, G.E. & Mingo, K.A., "A multivariate analysis of industrial bond ratings," *Journal of Finance, Vol.28,* No.1 (1973), 1-18.
43) Pinches, G.E. & Mingo, K.A., "The role of subordination and industrial bond ratings," *Journal of Finance*, Vol.30, No.1 (1975), 201-206.
44) Pogue, T.F. & Soldofsky, R.M., "What's in a bond rating?," *Journal of Financial and Quantitative Analysis, Vol.*4, No.2 (1969), 201-28.
45) Riesbeck, C.K. & Schank, R.C., *Inside Case-Based Reasoning*. Lawrence Erlbaum Associates, Hillsdale, NJ., 1989.
46) Schffer, J.D., Whitley, D. & Eshelman, L.J., "Combinations of genetic algorithms and neural networks: a survey of the state of the art," *Proc. of the International Workshop on Combinations of Genetic Algorithms and Neural Networks.* Baltimore, 1992, 1-37.
47) Shin, K.S., Shin, T.S. & Han, I., "Using induction technique to support case-based reasoning: a case of corporate bond rating," *Proc. of MS/OR Society Conference.* Seoul, Korea, 1997, 199-202.
48) Singleton, J.C. & Surkan, A.J., "Bond rating with neural networks," In Refenes, A.N. (Eds.), *Neural Networks in the Capital Markets.* London Business School, England: John Wiley, 1995,.301-307.
49) Slade, S., "Case-based reasoning: A research paradigm," *AI Magazine,* Vol.12, No.1 (1991), 42-55.
50) Syswerda, G., "Uniform crossover in genetic algorithms," In Schaffer, J.D. (Eds.), *Proc. 3rd Int.Conf. Genetic Algorithms.* San Maeto, CA: Morgan Kaufmann, 1989.
51) Wang, Y. & Ishii, N., "A method of similarity metrics for structured representations," *Expert Systems with Applications,* Vol.12, No.1 (1997), 89-100.
52) West, R.R., "An alternative approach to predicting corporate bond ratings," *Journal of Accounting Research*, Vol.8, No.1 (1970), 118-125.
53) Wong, F. & Tan, C., "Hybrid neural, genetic and fuzzy systems," In Deboeck, G.J. (Eds.), *Trading On The Edge.* New York: John Wiley, 1994, 245-247.