A Hybrid Approach Using Case-based Reasoning and Fuzzy Logic for Corporate Bond Rating

Hyun-jung Kim [a], Kyung-shik Shin [b]

Ewha Womans University, College of Business Administration
11-1 Daehyun-Dong, Seodaemun-Gu , Seoul 120-750, Korea
[a] *Tel: +82-2-3277-3767, Fax: +82-2-3277-2776, E-mail: charitas@empal.com*
[b] *Tel: +82-2-3277-2799, Fax: +82-2-3277-2776, E-mail: ksshin@ewha.ac.kr*

Abstract

A number of studies for corporate bond rating classification problems have demonstrated that artificial intelligence approaches such as Case-based reasoning (CBR) can be alternative methodologies to statistical techniques.

CBR is a problem solving technique in that the case specific knowledge of past experience is utilized to find a most similar solution to the new problems. To build a successful CBR system to deal with human information processing, the representation of knowledge of each attribute is an important key factor. We propose a hybrid approach of using fuzzy sets that describe the approximate phenomena of the real world because it handles inexact knowledge represented by common linguistic terms in a similar way as human reasoning compared to the other existing techniques. Integration of fuzzy sets with CBR is important to develop effective methods for dealing with vague and incomplete knowledge to statistically represent using membership value of fuzzy sets in CBR.

Keywords:

Fuzzy sets, Case-based reasoning, Corporate bond rating, Knowledge representation

Introduction

The rating process generally involves a review of the financial statements on the basis of quantitative factors as well as the human judgment on the basis of qualitative factors to a significant extent. Furthermore, the final bond rating is taken as a task where a decision made after the review and discussion by a group of analysts. In spite of these complex interactions through corporate bond rating process, an attempt to construct the model are facing challenging to achieve performance comparable to that of highly trained human expertise in bond rating domain. For financial institutions, predicting the bond ratings based on a model is helpful to be able to assess the credit status of the firms independently, since rating agencies do not provide credit ratings for all firms.

The early studies of bond rating applications have tended to use statistical techniques such as multiple discriminant analysis (MDA) models, which is most common means for classifying bonds into their rating categories. However, studies using only traditional statistical methods for prediction reach their limitations in applications with the violation of multivariate normality assumptions for independent variables frequently occurring with financial data. Recently, however, a number of studies have demonstrated that artificial intelligence approaches such as case-based reasoning (Buta, 1994; Kim and Han, 2001; Shin and Han, 1999; 2001) can be alternative methodology for corporate bond rating classification problems.

Case-based reasoning (CBR) is a problem solving technique in that the case specific knowledge of past experience is utilized to find a most similar solution to the new problems. The more basic principle underlying CBR is that a model articulates and restructures the knowledge acquisition framework of domain expertise, which uses analogical reasoning to solve complex problems and to learn from problem-solving experiences. Since, even for experts, to carry out such knowledge often cannot be successfully and exhaustively captured and represented, a CBR system should have to be considered to select an appropriate knowledge representation form based on the analogy of human information processing, which a human can easily describe one's knowledge and the described knowledge is comprehensible.

Furthermore, Building a successful CBR system largely depends on a good indexing and retrieving function, and for the effective indexing and retrieving tasks to given problem, the knowledge representation of each attribute is also generally regarded as one of the most important issues and is crucial to the success of a CBR system to work intelligently. Therefore, knowledge engineers have been desired to deal with human information processing that is on the basis of epistemic inference and imperfect decision-making process and investigated the realization of knowledge based system to interact between the human and the model.

We propose a hybrid approach of using fuzzy sets as an alternative methodology for handling vague and incomplete knowledge to statistically represent using membership values in CBR. Integration of fuzzy sets with CBR is also

importan' to develop effective methods for leading to avoid the mutually exclusivity of the case-based indexing and retrievin̲. It includes the question about whether fuzzy set concepts can be successfully integrated with a CBR system.

Fuzzy se s is used to describe the approximate phenomena of the re al world because it handles inexact knowledge represented by common linguistic terms in a similar way as human reasoning compared to the other existing techniques that defir e data using a crisp approach; in reality, however, it is almost impossible to define the data with certainty. Fuzziness itself is an effective means for the representation of such ambiguous concept, hence fuzzy sets are employed in this s udy. Our proposed approach is demonstrated by applications to corporate bond rating.

The rese arch questions are investigated as follows; the first asks whether fuzzy sets theory can be very attractive as a knowled ge representation technique in a successful CBR system. The second research question asks whether integration of fuzzy sets with CBR can be used to predict more accurate than the benchmark model for corporate bond rating classification task.

The remainder of this paper is organized as follows. In section 2, prior studies related to bond rating applications are reviewed. Next section contains the methodologies used in this study and a hybrid structure of CBR using fuzzy sets. The specific information about the data and experiments is described in the research development section. In the result & analysis section, empirical results are summarized and analyzed. The final section discusses the conclusion and future research issues.

Bond Rating Applications

Numerous Bond rating studies have traditionally tended to use statistical techniques such as ordinal least squares (OLS) (Horrigan, 1966; Pogue and Soldofsky, 1969; West, 1970), multiple discriminant analysis (MDA) (Pinches and Mingo, 1973; Baran et al., 1980; Belkaoui, 1980), probit (Kaplan and Urwitz, 1979; Gentry et al., 1988; Jackson and Boyd, 1988; Reiter and Emery, 1991; Iskandar-Datta and Emery, 1994) and logit (Ederington, 1985) models. Among the statistical techniques, the most common means for classifying bonds into their rating categories is MDA, which yields a liner discriminant function relating a set of independent variables to a dependent variable. However, studies using only statistical techniques for prediction reach their limitations in applications with the violation of multivariate normality assumptions for independent variables frequently occurring with financial data (Deakin, 1976).

While traditional statistical methods assume certain data distributions and focus on optimizing the likelihood of correct classification (Liang, Chandler and Han, 1990), inductive learning is a technology that automatically extracts knowledge from training samples, in which induction algorithms such as ID3 (Quinlan, 1986) and CART (Classification and Regression Trees) generate a tree type structure to organize cases in memory. Thus, the difference between a statistical approach and an inductive learning approach is that different assumptions and algorithms are used to generate knowledge structures (Shin and Han, 2001).

Shaw and Gentry (1990) applied inductive learning methods to risk classification applications and found that inductive learning's classification performance was better than probit or logit analysis. They have concluded that this result can be attributed to the fact that inductive learning is free from parametric and structural assumptions that underlie statistical methods.

Artificial neural networks have been found to be successful predictors for modeling a wide variety of business classification, clustering and pattern-recognition problems (Kwon et al, 1997; Moody and Utans, 1995; Salchenberger, Cinar and Lash, 1992). While basically an information processing technology, neural networks fundamentally differ from parametric statistical models. Parametric statistical models require the developer to specify the nature of the functional relationship such as linear or logistic between the dependent and independent variables. Once an assumption is made about the functional form, optimization techniques are used to determine a set of parameters that minimizes the measure of error. In contrast, neural networks with at least one hidden layer use data to develop an internal representation of the relationship between variables so that a priori assumptions about underlying parameter distributions are not required. As a consequence, better results might be expected with neural networks when the relationship between the variables does not fit the assumed model (Salchenberger, Cinar and Lash, 1992).

Dutta and Shekhar (1988) were the first to investigate the ability of neural networks to bond rating. They obtained a very high accuracy of 83.3% in discerning AA from non-AA rated bonds. However, they distinguished only one category of bonds, and the study was not clearly comparable with earlier research, which predicted a wide range of rating categories. They used both 6 and 10 financial variables that are used in prior bond rating studies. Since only 30 patterns are used for training neural networks, it is hard to conclude based on their study that the developed models can be generalized.

Singleton and Surkan (1990) also investigated the bond rating abilities of neural networks and linear models. They used multiple discriminant analysis, and found that neural networks outperformed the linear model for bond rating application. Another study by Singleton and Surkan (1995) showed that neural networks could predict the direction of a bond rating better than multiple discriminant analysis did.

Kim et al. (1993) compared neural networks model with regression, ID3, discriminant analysis and logistic analysis for bond rating with six categories of ratings. The results showed that the neural network model was the best among the above techniques in terms of classification accuracy.

Another study in bond rating prediction using neural networks was conducted by Moody and Utans (1995). They obtained 63.8% and 85.2% rates of accuracy when five and three classes were considered, respectively.

The recent study of bond rating done by Maher and Sen

(1997) compared the performance of neural networks with that of logistic regression. The results indicate that neural networks model performed better than a traditional logistic regression model. The best performance of the model was 70% (42 out of 60 samples).

Kwon et al. (1997) developed a corporate bond rating model using Korean bond rating data. They used ordinal pair-wise 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 an ordinal and pair-wise 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 (58%-61%)(Shin and Han, 2001).

Although numerous experimental studies reported the usefulness of neural networks in classification studies, there is a major drawback in building and using a model in which the user cannot readily comprehend the final rules that neural network models acquire. Case-based reasoning (CBR), in contrast, utilizes the most natural form of knowledge; a memory of stored cases recording specific prior episode. The basic principle underlying CBR is that human experts use analogical reasoning to solve complex problems and to learn from problem-solving experiences (Shin and Han, 2001).

Few studies have applied case-based reasoning for bond rating. Buta (1994) developed a CBR model that predicts corporate bond rating using financial data and ratings information of 1,000 companies from 1991 to 1992 in the S&P's Compustat database. Although performance of the system varied considerably based on the specific rating class of the company, using an inductive indexing scheme, the system matched the S&P recommended ratings for unseen cases (100 cases) 90.4% of the time.

Shin and Han (1999) proposed a case-based approach using genetic algorithms to case-based retrieval process in an attempt to increase the overall classification accuracy to predict bond rating of firms. They utilized a machine learning approach using genetic algorithms to find an optimal or near optimal importance weight vector for the attributes of cases in case indexing and retrieving. They applied the obtained importance weights of attributes to the matching and ranking procedure of CBR. Experimental results show that the GA-CBR hybrid model has the higher prediction accuracy (75.5%) than the individual method of MDA, ID3, and CBR models with different importance measures.

The recent study of Shin and Han (2001) developed a corporate bond rating model using Korean bond rating data. They applied case-based reasoning using an inductive indexing method to case indexing process. The total sample used was 3,886 companies whose commercial papers had been rated from 1991 to 1995. Experimental results show that inductive indexing methods can improve the effectiveness of case reasoning compared to the pure nearest-neighbor method resulting in higher classification accuracy (70%). That is, specifically, the success of the case-based reasoning system largely depends on the appropriateness of the indexing approach. In case of using induction trees as inductive indexing method, optimizing trees is the central tasks that represent an optimal combination level between the general domain knowledge and case-specific knowledge.

Kim and Han (2001) presented a case-based reasoning using the clustering methods to case indexing process to improve classification accuracy for the prediction of corporate bond rating. They utilized competitive artificial neural networks such as self-organizing map and learning vector quantization to generate the centroid value of clusters of bond rating cases, which these clustering techniques show better effective clusters than statistical clustering algorithms to the indexing and retrieving procedure of CBR. Experimental results show that the cluster-indexing CBR model has the higher prediction accuracy (67.1%-69.1%) than the individual method of MDA, ID3, and inductive learning indexing CBR models.

Research Methodology

Case-based Reasoning

Case-based reasoning (CBR) is a problem solving technique in that the case specific knowledge of past experience is utilized to find a solution most similar to the new problems; that is, unlike other generalized techniques, the past cases themselves are used as the basis for coping with a new situation.

Providing a solution to the new problem using CBR is a two-step process. The first step is checking against the case base and identifying similar case to solve the new case, then applying the new case from resulting in a solution for the given problem.

Case Representation

A case is a contextualized piece of knowledge representing an experience. It contains a past lesson that is the content of the case and a context in which the lesson can be used. Typically a case comprises of: (1) the problem that describes the state of the world when the case occurred, (2) the solution, which states the derived solution to that problem, and (3) the outcome, which describes the state of the world after the case occurred (Kolodner, 1991).

Cases can be represented in a variety of forms using the full range of AI representational formalisms, including frames, objects, predicates, semantic nets, and rules (Kolodner, 1993; Riesbeck and Schank, 1989).

Case Indexing and Retrieving

Case indexing involves assigning indexes to cases to facilitate their retrieval. The Indexes organize and label cases so that appropriate cases can be found when needed. In building case-based reasoning systems, the CBR community proposes several guidelines for choosing indexes for particular cases: (1) indexes should be predictive, (2) indexes should be abstract enough to make a case useful in a variety of future situations, (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 indexes. Choosing indexes manually involves deciding the purpose of the 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 indexes so that the search through case library can be done efficiently and accurately. Given a description of a problem, a retrieval algorithm, which uses the indexes in a case-memory, should retrieve the most similar cases to the current problem or situation. The retrieval algorithm relies on the organization of the memory to direct the search to potentially useful cases.

The indexes can either index case features independently for strictly associative retrieval or arrange cases from the most general to the most specific for hierarchical retrieval. There are three approaches to case indexing: nearest neighbor, inductive, and knowledge-guided (Brown and Gupta, 1994; Bryant, 1997; Shin and Han, 1999). 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. One of the most obvious measures of similarity between two cases is the distance. A matching function of the nearest-neighbor method using Euclidean distance between cases is as follows:

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

where 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 the matching function. Based on the importance assigned to each dimension, an aggregate match score is then computed. Ranking procedures order cases according to their scores where higher scoring cases are used before lower scoring ones.

Inductive indexing methods generally look for similarities over a series of instances and then form categories based on those similarities. Induction algorithms, such as ID3 and CART, determine which features best discriminate cases, and generate a tree type structure to organize the cases in memory. An induction tree is then built upon a database of training cases. 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 (Brown and Gupta, 1994).

Adaptation

Adaptation is the process of adjusting the retrieved cases to fit the current case. Once a matching case is retrieved, a CBR system should adapt the solution stored in the retrieved case to the needs of the current case. Adaptation looks for prominent differences between the retrieved case and the current case and then applies formulae or rules that take those differences into account when suggesting a solution (Shin and Han, 2001).

Fuzzy Logic

Since fuzzy set theory was introduced by Zadeh in 1965 as a generalization of the conventional set theory (Zadeh, 1965), the fuzzy set theory has been widely used in many fields of application, such as pattern recognition, data analysis, system control, and so on (Cannon et al., 1986; Driankov et al., 1993; Kruse et al., 1994; Klir and Yuan, 1995; Theodoridis and Koutroumbas, 1999; Xiong et al., 2001). The primary objectives of fuzzy set theory are to represent the structured knowledge based on the way the brain deals with inexact information, and improve the intelligence of systems working in an uncertain, imprecise and noisy environment.

Crisp Set

An object either belongs or does not belong to a given set, which the membership of a crisp set, so-called the characteristic function, is defined as to dichotomize the object into a binary class. For example, a person who is classified as "young" cannot be considered "not young" at the same time. A person who is classified as a young man cannot be an old man at the same time. This is called a classical set or a crisp set.

Let U be a universe, the characteristic function $m_s(x)$ that represents whether an object x belongs to a crisp set S in U takes its value in 0 or 1. That is, the characteristic function $m_s(x)$ is defined as follows:

$$m_s(x) = \begin{cases} 1 & \text{if } x \in S \\ 0 & \text{if } x \notin S \end{cases} \qquad (2)$$

Fuzzy Set

Classical sets practically have a limitation, since they are not suitably used to represent the slight difference of the object feature. For example, if we define young as someone whose age is 20 or younger, a 21-year-young person cannot be categorized as a young person under crisp set concepts as described above session.

However, human beings often adopt a more flexible approach by means of assigning the different degrees of possibility to a person who can be young and old; a 50-year-old person may be considered old and young of different degrees at the same time. A set that allows partial membership is called a fuzzy set.

Let U denote a universe space of objects, then a fuzzy set F in the universe of U can be defined as a following set of ordered pairs:

$$F=\{(x, m_f(x))|x \in U\} \qquad (3)$$

where $m_t(x)$ is the grade of membership of x in F, which indicates the degree that x belongs to a set F. The range of membership function for fuzzy sets generally maps to the unit interval [0,1].

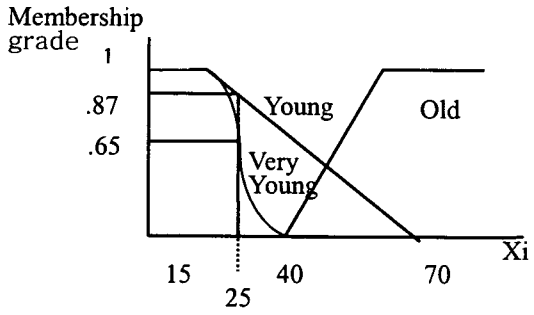

Figure 1 - The membership functions of old and young

As Figure 1 shows the membership function of old and young persons, we can see that the membership function of old person who is older than 70 and young person who is younger than 15 is 1.0, whereas the membership grade of a 25-year young person is 0.87.

In addition to fuzzy terms such as old or young, modifiers such as very, somewhat, and more or less are frequently used to represent an object in human languages. An adopted approach to handle these modifiers in fuzzy sets is only to add the simple operations to the membership function of the fuzzy term, which is another advantage of fuzzy sets. Given the membership function of young person $m_{young}(x)$ in Figure 1, the membership function of very young person can be defined as $[m_{young}(x)]^2$. In accordance, the membership grade of a 25-year young person is 0.87, whereas the membership grade of a 25-year very young person is 0.65 (Jeng and Liang, 1995).

It is essential in fuzzy sets that the boundaries of a class are not exactly divided in which the transition from membership to non-membership is gradual, since the membership value of fuzzy sets is presented by possibilities rather than binary; that is, compared with classical sets, the fuzzy set representation allows a range of gray area to define set membership for classification instead of using a single threshold value. Although fuzzy sets make it possible to express the uncertain, vague, and inexact information, the representation process using fuzzy sets is extremely logical and mathematical in describing the properties of objects that is not completely known to us.

Fuzzy Operations

Since fuzzy sets are regarded as an extension case of a crisp sets, most basic operations defined on crisp sets can also be applied to fuzzy sets and some operations are unique to the fuzzy sets.

In brief, the application of fuzzy sets theory normally includes three procedures, that is fuzzification, logic decision and defuzzification. Fuzzification includes constructing the membership function after the identification of the input and output variables and the division variables into different partition by a given problem domain. Logic decision involves the design of the IF-THEN inference rules, the calculation of the degree of applicability of each IF-THEN rule, and the determination of the output fuzzy set. Defuzzification involves the determination of the crisp output from the fuzzy outputs of the IF-THEN inference system (Xiong et al., 2001).

Fuzzy systems have been successfully applied to a variety of knowledge-based model to improve their intelligence. At first, they begin at finding formalized information about the structured categories in an environment and then articulate fuzzy if-then rules as a sort of expert knowledge.

Hybrid Structure using CBR and Fuzzy logic

Integration of fuzzy sets with CBR is important to develop effective methods for handling vague and incomplete knowledge to statistically represent using membership values of fuzzy sets in CBR. Fuzzy sets is used to describe the approximate phenomena of the real world because it has similarity to human reasoning and natural languages compared to the other existing techniques.

Adaptation is the process of adjusting the retrieved cases to fit the current case. However, for classification tasks, adaptation is not applicable and therefore is not used in this study.

In the design of a CBR system using fuzzy sets, the first step is to convert inputs into fuzzy representation based on membership functions defined in fuzzifier, and output are recommended as a result of the indexing and retrieving process in CBR.

Fuzzy Representation

For the case representation in the hybrid system, fuzzification using the membership function is the first process that converts to the fuzzy linguistic terms and calculates the similarity of a single feature of a case with the corresponding attribute of the target (Watson, 1999). Thus, constructing the fuzzy membership functions that transform the continuous input attributes into the membership value of fuzzy preference and linguistic terms is considered to have a critical impact on the performance of the proposed hybrid model.

Previous studies have proposed numerous methods to determine the number of membership functions and to find optimal parameters of membership functions by identifying the class prototype with the smallest distance or highest similarity that distinguish the distinctions of given patterns such as Kohonen's learning vector quantization algorithm, fuzzy c-means clustering algorithm and subtractive clustering method and so on (Bezdek, 1987; Chiu, 1994; Chen and Wang, 1999).

In this study, we used an approach using k-means clustering algorithm suggested by Klimasauskas (1992) to formulate the membership functions, which finds an effective cluster in a crisp set, and converts into the fuzzy membership value associated with each class.

The Formula used to transform an input value X_i in the set following boundary [a,b] to calculate the degree of membership in fuzzy set, $F_i(X_i)$, is shown as follows;

$$F_i(X_i) = \max(0, 1 - K \times |X_i - C|) \quad (4)$$

where K: the scale factor = $2 \times (1-M)/(b-a)$, C: the center

between the boundaries

The fuzzification process for fuzzy indexing and retrieving is summarized as the following steps:

a. The membership function of each class based on clustering algorithms is determined.
b. Numerical values of each case are converted into proper classes.
c. Attributes are presented into fuzzy terms and membership values based on membership functions defined in step a.

Fuzzy Indexing and Retrieving

The major advantage of fuzzy indexing and retrieving is that they allow multiple class memberships to be defined on a single attribute; that is, a person may be classified as old and young at the same time with different membership grades, which would make a case qualify for the retrieving criteria of old and young (Jeng and Liang, 1995).

Cases are indexed based on the fuzzy terms of each attributes processed by fuzzifier in fuzzy representation step, before being stored in the case base. Table 1 shows the example of attribute X_i represented by fuzzy terms and membership values. Figure 2 illustrate the indexing result of the cases in Table 1 by their classes.

Table 1 - The example of fuzzy representation

Firm	Rating	X_i	Fuzzy X_i / membership value
Company A	A1	0.5	low / 0.7
Company B	A1	2.2	low / 0.56, middle / 0.55
Company C	A2	3.5	middle / 0.67, high / 0.5
Company D	B	3.8	middle / 0.56, high / 0.54
Company E	B	5.0	high / 0.71
...	...	...	...

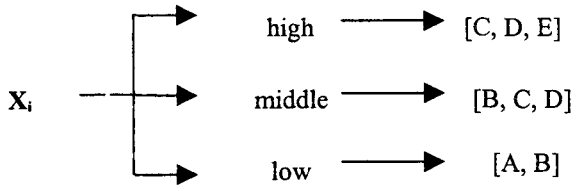

Figure 2 - The example of fuzzy indexing
(Adapted from Jeng and Liang, 1995)

Once cases are indexed and stored in the case base, they can be used for problem solving. When a new case is encountered, the CBR searches the case base to retrieve similar case. The fuzzy indexing and retrieving process is summarized as the following steps:

a. The fuzzy terms resulting from fuzzifier are used to search the candidate cases that matched with new case in the case base.
b. The one that has highest similarity among the candidate cases in the prior step is selected to construct a solution to the new case.

There are several ways of finding the most similar case. The most straightforward way is to count the number of cases showing a particular result. Another approach is to use functions to use the distance function between the new and candidate cases and choose the one having the shortest distance as the most similar case (Jeng and Liang, 1995).

The distance can be measured by the difference of the original attribute value or the converted fuzzy membership grades between candidate cases and the new case. If we use the original value, the distance function can be defined as follows:

$$DIS_{ab} = \sqrt{\sum_{i=1}^{n} w_i \times (x_{ai} - x_{bi})^2} \quad (5)$$

where x_{ai} and x_{bi} are the original value of attribute i for candidate a and new case b, respectively and n is the number of attribute, and w_i is the weight of a attribute i.

If we use the converted fuzzy grades, the distance function can be defined as follows:

$$DIS_{ab} = \sum_{i=1}^{n}\sum_{j=1}^{m} | x_{aij} - x_{bij} | \quad (6)$$

where x_{aij} and x_{bij} are the grades of attribute i, class j for candidate a and new case b, respectively, n is the number of attribute and m is the number of class.

Research Development

The research data consists of 297 financial ratios and the corresponding bond ratings of Korean companies. The ratings we employ in this study covers all bonds rated by National Information and Credit Evaluation, Inc., one of the most prominent bond rating agencies in Korea. Our total sample includes 1,816 companies whose commercial papers have been rated in the period 1997-2000. Credit grades are defined as outputs and classified as 5 coarser rating categories (A1, A2, A3, B, C) according to credit levels. Table 2 shows the organization of the data set.

Table 2 - Number of companies in each rating

Ratings	Number of cases	%
A1	58	3.2
A2	242	13.3
A3	586	32.3
B	780	43.0
C	150	8.3
Total	1,816	100.0

We apply two stages of input variable selection process. At the first stage, we select 106 variables by 1-way ANOVA between each financial ratio as input variable and credit grade as output variable. In the second stage, we select 9 variables using MDA stepwise method to reduce the dimensionality. We select input variables satisfying the univariate test first, and then select significant variables by stepwise method for refinement. The selected variables for this research are shown in table 3.

The data set is split into two subsets; about 90% of the data is used for a reference set and 10% for a holdout set. The reference data is used to construct a case base for fuzzy indexing and retrieving. The holdout data is used to test the results with the data that is not utilized to develop the model. The number of the reference cases and the holdout

cases are 1,635 and 181, respectively.

Table 3 - Definition of variables

Variables	Description
X1	Net income to total asset
X2	Net interest coverage ratio
X3	Times interest earned
X4	Net income to capital stock
X5	Equity to total asset
X6	Fixed assets to total asset
X7	Current liabilities to total asset
X8	Transition of ordinary profit
X9	Transition of operating activities cash flows

For each continuous input variable, the fuzzy preprocessing procedure of three fuzzy sets is designed based on the input data analysis. In this study, experimenting the bond rating classification, three fuzzy sets (high, middle, low) were assumed for each of the 9 input variables as shown in Figure 3.

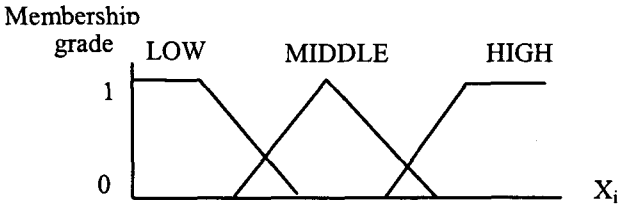

Figure 3 - The membership functions of fuzzy terms

According to formulas of computing the degree of membership in fuzzy set, as suggested in the previous session of fuzzy representation, the shape of the fuzzy set is controlled by the centroid, upper and lower boundary of each cluster. The value of fuzzy membership function at the boundary (M) is set at 0.5 in this experiment, which means that we will consider fuzzy terms whose membership values are higher than 0.5 relatively reliable.

For example, in the case of X_1 variable, three membership functions that use values generated by k-means clustering algorithm are constructed as the following equations:

$$m_{high}(x) = \begin{cases} \max(0, 1 - 11.73 \times |X - 0.10|) & \text{if } x \le 0.10 \\ 1 & \text{if } x > 0.10 \end{cases}$$

$$m_{middle}(x) = \max(0, 1 - 11.36 \times |X - (-0.02)|)$$

$$m_{low}(x) = \begin{cases} \max(0, 1 - 8.94 \times |X - (-0.17)|) & \text{if } x \ge 0.17 \\ 1 & \text{if } x < 0.17 \end{cases}$$

In the equation shown above, x is the input value of a financial ratio X_1, and $m_{high}(x)$, $m_{middle}(x)$, and $m_{low}(x)$ are the fuzzy sets corresponding to the linguistic terms high, middle, and low, respectively. Accordingly the same procedure is applied to the other input variables.

We utilized statistical clustering algorithm such as k-means clustering algorithm and competitive artificial neural networks such as Kohonen self-organizing maps to determine the centroid, lower and upper boundary value of clusters of the input variable for generating the fuzzy membership functions corresponding to the categorized linguistic terms.

Result & Analysis

To investigate the effectiveness of the integrated approach for case representation and indexing using fuzzy sets in the context of the corporate bond rating problem, the results obtained are compared with those of conventional techniques such as CBR-pure model and MDA. The CBR-pure model uses a nearest-neighbor algorithm that weights obtained by Pearson's correlation analysis are assigned among attributes; correlation coefficient acquired from Pearson's correlation analysis is transformed to the weighted value of each attributes.

Among the several retrieving methods as mentioned above, we apply three approaches; to choose the majority of cases, to measure the distance of the original attribute value, and the converted fuzzy membership grades between the new and candidate cases. Fuzzy-CBR[a] and Fuzzy-CBR[b] model applies the nearest-neighbor retrieval using membership grade of each converted class and original quantitative value among candidate cases obtained by indexing process, respectively. Fuzzy-CBR[c] follows the procedure of the majority rule retrieving approach.

Table 4 - Classification accuracies (%)

Methods		A1	A2	A3	B	C	Avera
MDA		60.00	20.83	50.00	42.86	53.85	43.3
Pure-CBR		20.00	29.17	60.78	52.11	41.67	49.6
Fuzzy-CBR[a]	K-means	40.00	41.67	50.98	57.75	50.00	52.1
	Kohonen	40.00	50.00	55.56	71.43	30.77	59.0
Fuzzy-CBR[b]	K-means	60.00	45.83	50.98	69.01	41.67	57.6
	Kohonen	40.00	41.67	66.67	75.71	46.15	64.4
Fuzzy-CBR[c]	K-means	40.00	41.67	62.75	74.65	25.00	61.3
	Kohonen	0.00	58.33	62.96	80.00	15.38	63.8

a. Nearest-neighbor retrieval using original quantitative value of each converted class

b. Nearest-neighbor retrieval using membership grade of each converted class

c. Majority rule retrieval

Table 4 shows the comparison of the results of the classification techniques applied for this study. Each cell contains the accuracy of the various classification techniques by classes. The results of statistical classification techniques, MDA are also presented as benchmark to verify the applicability of the proposed model to the domain. Among the techniques, the fuzzy-CBR integrated models using majority rule retrieving approach have the highest level of accuracies in the given data sets. And between two clustering techniques to obtain the information of clusters of bond rating cases, Kohonen self-organizing maps show better effective clusters than k-means clustering algorithm to the indexing and retrieving procedure of CBR.

Since bond-rating prediction applied domain in this study is

the multiple classification problem, it is difficult to design a well-performed model that can obtain the high classification accuracy as shown in Table 4. Therefore, to show that the fuzzy CBR system can be considered as a successful bond ratings estimator, the modified classification accuracies based on the assumption that one difference of the rating is matched are suggested in Table 5.

Table 5 - Modified classification accuracies (%)

Methods		A1	A2	A3	B	C	Average
MDA		60.00	95.83	83.33	91.43	92.31	88.55
Pure-CBR		80.00	79.17	96.30	88.57	92.31	89.76
Fuzzy-CBR[a]	K-means	100.00	75.00	98.04	91.55	100.00	92.02
	Kohonen	80.00	75.00	96.30	97.14	84.62	92.17
Fuzzy-CBR[b]	K-means	100.00	83.33	98.04	91.55	75.00	91.41
	Kohonen	100.00	87.50	98.15	95.71	92.31	95.18
Fuzzy-CBR[c]	K-means	100.00	79.17	100.00	94.37	91.67	93.87
	Kohonen	80.00	75.00	100.00	97.14	84.62	93.37

a. Nearest-neighbor retrieval using original quantitative value of each converted class

b. Nearest-neighbor retrieval using membership grade of each converted class

c. Majority rule retrieval

McNemar test results for the comparison of the predictive performance between the comparative models and the fuzzy CBR models for the holdout cases are summarized in Table 6.

Table 6 - McNemar values for the comparison of performance between models

(Chi-square value)

		MDA	Pure-CBR	Fuzzy-CBR[a]	Fuzzy-CBR[b]
Pure-CBR		0.141			
Fuzzy-CBR[a]	K-means	0.068 *	0.672		
	Kohonen	0.002 ***	0.041 **		
Fuzzy-CBR[b]	K-means	0.001 ***	0.111	0.157	
	Kohonen	0.000 ***	0.002 ***	0.243	
Fuzzy-CBR[c]	K-means	0.000 ***	0.023 **	0.091 *	0.488
	Kohonen	0.000 ***	0.002 ***	0.200	1.000

* significant at 10%
** significant at 5%
*** significant at 1%

The results of McNemar tests support that the fuzzy CBR model has higher prediction accuracy than all the comparative models with significant levels and the fuzzy CBR model with a clustering technique of Kohonen self-organizing maps performs significantly better than that of k-means clustering algorithm. In addition, it appears that Fuzzy-CBR[c] with a clustering technique of k-means algorithm performs better than Fuzzy-CBR[a] at 10% significance level; however, the other integration types of fuzzy-CBR are significantly indifferent.

The overall result shows the integrated models with majority rule retrieving approach performs better than MDA and conventional CBR. Based on the results, we conclude that the integrated approach proposed for this study is effective, enhancing the classification accuracy of the CBR for the corporate bond rating application domain.

Conclusion

In this study, we have proposed a hybrid approach of using fuzzy sets as an alternative methodology to represent for case-based indexing and retrieving to the problem of corporate bond rating. Integration of fuzzy sets with CBR is important to develop effective methods for handling vague and incomplete knowledge to statistically represent using membership values of fuzzy sets in CBR. The preliminary results show that the integrated models are effective, enhancing the classification accuracy of CBR for the bond rating application domain. We also show that the proposed approach increases the flexibility of indexing and retrieving process in CBR.

Our study has the following limitations that need further research. First, the determination of classes and membership functions has a critical impact on the performance of the resulting system. The second issue for future research relates to the information bias or information loss due to data conversion using fuzzy sets. In addition to the above issues, we also need to examine the integration of the fuzzy approach with existing other types of techniques.

References

[1] Baran, A., Lakonishok, J., and Ofer, A. R. (1980). "The Value of General Price Level Adjusted Data to Bond Rating," *Journal of Business Finance and Accounting*, Vol. 7, pp. 135-149.

[2] Belkaoui, A. (1980). "Industrial Bond Ratings: A New Look," *Financial Management*, Vol. 9, pp. 44-51.

[3] Bezdek, J.C. (1987). *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York.

[4] Brown, C.E., and Gupta, U.G. (1994). "Applying Case-based Reasoning to the Accounting Domain," *Intelligent Systems in Accounting, Finance and Management*, Vol. 3, pp. 205-221.

[5] Bryant, S.M. (1997). "A Case-based Reasoning Approach to Bankruptcy Prediction Modeling," *Intelligent Systems in Accounting, Finance and Management*, Vol. 6, pp. 195-214.

[6] Buta, P. (1994). "Mining for Financial Knowledge with CBR," *AI EXPERT*, Vol. 9, pp. 34-41.

[7] Cannon, R.L., Dave, J.V., and Bezdek, J.C. (1986). "Efficient Implementation of the Fuzzy C-means Clustering

Algorithms," *IEEE Trans, -PAMI*, Vol. 8, pp. 248-555.

[8] Chen, M.S., and Wang, S.W. (1999). "Fuzzy Clustering analysis for Optimizing Fuzzy Membership functions," *Fuzzy Sets and Systems*, Vol. 103, pp. 239-254.

[9] Chiu, S. (1994). "Fuzzy Model Identification Based On Cluster Estimation," *J. Intell. Fuzzy Systems*, Vol. 2, pp. 267-278.

[10] Deakin, E.B. (1976). "Discriminant Analysis of Predictors of Business Failure," *Journal of Accounting Research*, pp.167-179.

[11] Driankov, D., Hellendoorn, H., and Reinframk, M. (1993). *An Introduction to Fuzzy Control*, Springer, Berlin.

[12] Dutta, S., and Shekhar, S. (1988). "Bond Rating: A Non-conservative Application of Neural Networks," *Proceedings of IEEE International Conference on Neural Networks*, pp. 443-450.

[13] Ederington, H.L. (1985). "Classification Models and Bond Ratings," *Financial Review*, Vol. 20, pp. 237-262.

[14] Gentry, J.A., Whitford, D.T., and Newbold, P. (1988). "Predicting Industrial Bond Ratings with a Probit Model and Funds Flow Components," *Financial Review*, Vol. 23, pp. 267-286.

[15] Horrigan, J.O. (1966). "The Determination of Long Term Credit Standing with Financial Ratios," *Journal of Accounting Research, supplement*, pp. 44-62.

[16] Iskandar-Datta, M.E., and Emery, D.R. (1994). "An Empirical Investigation of the Role of Indenture Provisions in Determining Bond Ratings," *Journal of Banking and Finance*, Vol. 18, pp. 93-111.

[17] Jackson, J.D., and Boyd, J.W. (1988). "A Statistical Approach to Modeling the Behavior of Bond Raters," *The Journal of Behavioral Economics*, Vol. 17, pp. 173-193.

[18] Jeng, B.C., and Liang, T.P. (1995). "Fuzzy Indexing and Retrieval in Case-Based Systems," *Expert Systems with Applications*, Vol. 8, pp. 135-142.

[19] Kaplan, R.S., and Urwitz, G. (1979). "Statistical Models of Bond Ratings: A Methodological Inquiry," *Journal of Business*, Vol. 52, pp. 231-262.

[20] Kim, B.O., and Lee, S.M. (1995). "A Bond Rating Expert System for Industrial Companies," *Expert Systems with Applications*, Vol. 9, pp. 63-70.

[21] Kim, J., Weistroffer, H.R., and Redmond, R.T. (1993). "Expert Systems for Bond Rating: A Comparative Analysis of Statistical, Rule-based and Neural network Systems," *Expert Systems*, Vol. 10, pp. 167-172.

[22] Kim, K.S., and Han, I. (2001). "The Clustering-indexing Method for Case-based Reasoning Using Self-organizing Maps and Learning Vector Quantization for Bond Rating Cases," *Expert Systems with Applications*, Vol. 12, pp. 147-156.

[23] Klimasauskas, C.C. (1992). "Hybrid Fuzzy Encoding for Improved Backpropagation Performance," *Advanced Technology for Developers*, Vol. 1, pp. 13-16.

[24] Klir, G.J., and Yuan, B. (1995). *Fuzzy Sets and Fuzzy Logic-Theory and Applications*, Prentice-Hall, London.

[25] Kolodner, J. (1991). "Improving Human Decision Making through Case-based Decision Aiding," *AI Magazine*, Vol. 12, pp. 52-68.

[26] Kolodner, J. (1993). *Case-Based Reasoning*, Morgan Kaufmann, San Mateo, CA.

[27] Kruse, R., Gebhardt, J., and Klawonn, F. (1994). *Foundations of Fuzzy Systems*, Wiley, New York.

[28] Kwon, Y.S., Han, I.G., and Lee, K.C. (1997). "Ordinal Pairwise Partitioning (OPP) Approach to Neural Networks Training in Bond Rating," *Intelligent Systems in Accounting, Finance and Management*, Vol. 6, pp. 23-40.

[29] Liang, T.P., Chandler, J.S., and Han, I. (1990). "Integrating Statistical and Inductive Learning Methods for Knowledge Acquisition," *Expert Systems with Applications* Vol. 1, pp. 391-401.

[30] Maher, J.J., and Sen, T.K. (1997). "Predicting Bond Ratings Using Neural Networks: A Comparison with Logistic Regression," *Intelligent Systems in Accounting, Finance and Management*, Vol. 6, pp. 59-72.

[31] Moody, J., and Utans, J. (1995). "Architecture Selection Strategies for Neural Networks Application to Corporate Bond Rating," in: Refenes, A. (Eds.), *Neural Networks in the Capital Markets*, John Wiley.

[32] Pinches, G.E., and Mingo, K.A. (1973). "A Multivariate Analysis of Industrial Bond Ratings," *Journal of Finance*, Vol. 28, pp. 1-18.

[33] Pogue, T.F. and Soldofsky, R.M., (1969). "What's in a Bond Rating?" *Journal of Financial and Quantitative Analysis*, Vol. 4, pp. 201-228.

[34] Quinlan, J.R. (1986). "Induction of Decision Trees," *Machine Learning*, Vol. 1, pp. 81-106.

[35] Reiter, S.A., and Emery, D.R. (1991). "Estimation Issues in Bond-rating Models," *Advances in Quantitative Analysis of Finance and Account*, Vol. 1, pp. 147-163.

[36] Riesbeck, C.K., and Schank, R.C. (1989). *Inside Case-Based Reasoning*, Lawrence Erbium Associates, Hillsdale, NJ.

[37] Salchenberger, L.M., Cinar, E.M., and Lash, N.A. (1992). "Neural Networks: A New Tool for Predicting Thrift Failures," in: Trippi, R., and Turban, E. (Eds.), *Neural Networks in Finance and Investing*, Probus Publishing Company.

[38] Shaw, M., and Gentry, J. (1990). "Inductive Learning for Risk Classification," *IEEE Expert*, pp. 47-53.

[39] Shin, K.S. (1998). "The Hybrid Modeling of Case-Based Reasoning for Corporate Bond Rating," Ph.D. Dissertation, Korea Advanced Institute of Science & Technology.

[40] Shin, K.S., and Han, I. (1999). "Case-based Reasoning Supported by Genetic Algorithms for Corporate Bond Rating," *Expert Systems with Applications*, Vol. 16, pp. 85-95.

[41] Shin, K.S., and Han, I. (2001). "A Case-based Approach Using Inductive Indexing for Corporate Bond Rating," *Decision Support Systems*, Vol. 32, pp. 41-52.

[42] Singleton, J.C., and Surkan, A.J. (1990). "Neural Networks for Bond Rating Improved by Multiple Hidden

Layers," *Proceedings of the IEEE International Conference on Neural Networks*, pp. 163-168.

[43] Singleton, J.C., and Surkan, A.J. (1995). "Bond Rating with Neural Networks," in: Refenes, A. (Eds.), *Neural Networks in the Capital Markets*, John Wiley.

[44] Slade, S. (1991). "Case-based Reasoning: A Research Paradigm," *AI Magazine*, Vol. 12, pp. 42-55.

[45] Theodoridis, S., and Koutroumbas, K. (1999). *Pattern Recognition. Academic Press*, New York.

[46] Watson, I. (1999). "Case-based Reasoning is a Methodology Not a Technology," *Knowledge Based System*, Vol. 12, pp. 303-308.

[47] West, R.R. (1970). "An Alternative Approach to Predicting Corporate Bond Ratings," *Journal of Accounting Research*, Vol. 8, pp. 118-125.

[48] Wilson, R.L., and Sharda, R. (1994). "Bankruptcy Prediction Using Neural Networks," *Decision Support Systems*, Vol. 11, pp. 545-557.

[49] Xiong, L., Shamseldin, A. Y., and O'Connor, K. M. (2001). "A Non-linear Combination of the Forecasts of Rainfall-runoff Models by the First-order Takagi-Sugeno Fuzzy System," *Journal of Hydrology*, Vol. 245, pp. 196-217.

[50] Zadeh, L.A. (1965). "Fuzzy Sets," *Information Control*, Vol. 8, pp. 338-353.