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

Hyun-jung Kim
College of Business Administration, Ewha Womans University
(charitas@empal.com)

Kyung-shik Shin
College of Business Administration, Ewha Womans University
(ksshin@ewha.ac.kr)

This study investigates the effectiveness of a hybrid approach using fuzzy sets that describe approximate phenomena of the real world. Compared to the other existing techniques, the approach handles inexact knowledge in common linguistic terms as human reasoning does it. Integration of fuzzy sets with case-based reasoning (CBR) is important in that it helps to develop a successful system for dealing with vague and incomplete knowledge which statistically uses membership value of fuzzy sets in CBR. The preliminary results show that the accuracy of the integrated fuzzy-CBR approach proposed for this study is higher that that of conventional techniques. Our proposed approach is applied to corporate bond rating of Korean companies.

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

Received: June 2004 Accepted: October 2004 Corresponding Author: Kyung-shik Shin

1. Introduction

The rating process generally involves reviews of the financial statements on the basis of quantitative factors as well as human judgment on the basis of qualitative factors to a significant extent. Furthermore, the final bond rating is made by reviews and discussions of a group of analysts. To cope with these complex interactions of corporate bond rating process, an attempt is made to construct a model, which can achieve performance comparable to that of highly trained human expertise. For financial institutions, the prediction of bond rating based on the model is helpful to assess the credit status of firms independently, since rating agencies do not provide credit rating for every firm.

The early studies of bond rating applications 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 [6][21][39][40] 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 experiences is utilized to find the most similar solution to a new problem. 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. Because it is difficult even for experts to capture and represent such knowledge, a CBR system should be considered to select an appropriate knowledge representation form based on the analogy of human information processing. For the realization of knowledge-based system to interact between the human and the model, knowledge engineers have investigated a way to deal with human information processing that is on the basis of epistemic inference and imperfect decision-making process. In addition, building a successful CBR system largely depends on good indexing and retrieving function. For the effective indexing and retrieving tasks to given problem, the knowledge representation of each attribute is also regarded as one of the most important issues. Choosing a suitable method to represent knowledge is crucial for a CBR system to work intelligently.

We propose a hybrid approach using fuzzy sets as alternative methodology for handling vague and incomplete knowledge which statistically uses membership values in CBR. Compared to other existing techniques which define data using a crisp approach, fuzzy sets handle inexact knowledge in common linguistic terms as human reasoning describes approximate phenomena of the real world; moreover, it is almost impossible to define the data with certainty in reality. Fuzziness itself is an effective means for the representation of such an ambiguous concept, hence fuzzy sets are employed in this study. Integration of fuzzy sets with CBR is also important, because it helps to develop effective methods to avoid the mutual exclusivity of case-based indexing and retrieving. It includes the question about whether fuzzy set concepts can be successfully integrated with a CBR system. Our proposed approach is demonstrated by applications to corporate bond rating.

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

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 data and experiments is described in the research development section. In the result and analysis section, empirical results are summarized and analyzed. The final section includes a conclusion and future research issues.

2. Bond Rating Applications

Numerous bond rating studies have traditionally tended to use statistical techniques such as ordinal least squares (OLS) [15][32][45], multiple discriminant analysis (MDA) [1][2][31], probit [14][16][17][19][34] and logit [13] models. Among those 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, statistical techniques have some limitations in applications due to violation of multivariate normality assumptions for independent variables, which is frequently occurred in financial data [10].

While traditional statistical methods assume certain data distributions and focus on optimizing the likelihood of correct classification [28], inductive learning is a technology that automatically extracts knowledge from training samples, in which induction algorithms such as ID3 [33] and CART (Classification and Regression Trees) generate a tree type structure to organize cases in memory. Thus, the difference between a statistical and an inductive learning approach is that different assumptions and algorithms are used to generate knowledge structures.

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 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 [27][30][36]. However, neural networks fundamentally differ from parametric statistical models. Parametric statistical models require a 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 errors. In contrast, neural networks with at least one hidden layer use data to develop an internal representation of relationship between variables, so that prior assumptions about underlying parameter distributions are not required. As a consequence, better results can be expected with neural networks when the relationship between the variables does not fit the assumed model [36].

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 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 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 indicated 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 was to partition the data set in an ordinal and pair-wise manner into the output classes. Experimental results showed that the OPP approach had the highest level of accuracy (71%-73%), followed by conventional neural networks (66%-67%) and multiple discriminant analysis (58%-61%).

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 a 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 which recorded specific prior episodes. The basic principle underlying CBR is that human experts use analogical reasoning to solve complex problems and to learn from problem-solving experiences [38].

Few studies had applied case-based reasoning for bond rating. Buta (1994) developed a CBR model that predicted 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's recommended rating for unseen cases (100 cases) 90.4% of the time.

Shin and Han (1999) proposed a case-based approach using genetic algorithms to a 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 weight of attributes to the matching and ranking procedure of CBR. Experimental results showed that the GA-CBR hybrid model had 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 a case indexing process. The total samples used were 3,886 companies whose commercial papers had been rated from 1991 to 1995. Experimental results showed that inductive indexing methods could 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 an inductive indexing method, optimizing trees is the central tasks that represent an optimal combination level between general domain knowledge and case-specific knowledge.

Kim and Han (2001) presented a case-based reasoning using the clustering methods for 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. These clustering techniques show more effective clusters than statistical clustering algorithms to the indexing and retrieving procedure of CBR. Experimental results showed that the cluster-indexing CBR model had the higher prediction accuracy (67.1%-69.1%) than the individual method of MDA, ID3, and inductive learning indexing CBR models.

3. Research Methodology

3.1 Case-based Reasoning

Case-based reasoning (CBR) is a problem-solving technique in that the case specific knowledge of past experiences is utilized to find the most similar solution to a new problem; 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 to check the case base and to identify similar cases to solve a new case. Then the second step is to apply the new case to come up with a solution to the given problem.

3.1.1 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 occurs [24].

Cases can be represented in a variety of forms using the full range of AI representational formalism, including frames, objects, predicates, semantic nets, and rules [25][35].

3.1.2 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 [24][25]. 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 searching 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 [4][5][39]. 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 discriminate the best case, 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 because such knowledge often cannot be successfully captured and represented. Therefore, many systems use knowledge-guided indexing in conjunction with other indexing techniques [4].

3.1.3 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 [39].

3.2 Fuzzy Sets

Since the fuzzy set theory was introduced by Zadeh in 1965 as a generalization of the conventional set theory [47], the fuzzy set theory has been widely used in many fields of application, such as pattern recognition, data analysis, system control, and so on [7][11][23][26][43][46]. The primary objectives of the fuzzy set theory are to represent the structured knowledge based on the way human deals with inexact information, and to improve the intelligence of systems working in an uncertain, imprecise and noisy environment.

3.2.1 Crisp Set

The membership of a crisp set, so-called the characteristic function, is defined as to dichotomize an object into one of binary classes. 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 $m_s(x)$ represents whether an object x belongs to a crisp set S in U or not, and it takes value 0 or 1. That is, the characteristic function $m_s(x)$ is defined as follows:

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

3.2.2 Fuzzy Set

Classical sets practically have a limitation 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 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 at the same time with different degrees. 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:

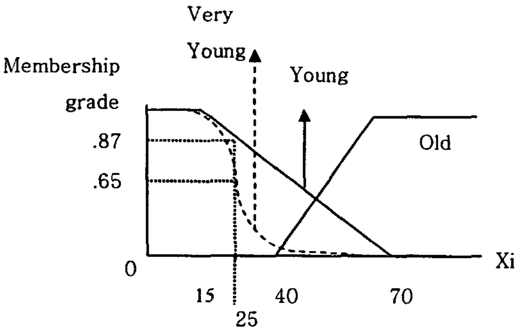
$$F=\{(x,\ m_t(x))|x \in U\} \qquad (3)$$

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

As [Fig. 1] shows the membership function of old and young persons, we can see that the membership function of an old person who is older than 70 and a 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 and 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 [Fig. 1], the membership function of a 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.

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 binaries; 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 are not completely known to us.



[Fig. 1] The membership functions of old and young

3.2.3 Fuzzy Operations

Since fuzzy sets are regarded as an extension case of crisp sets, most basic operations defined in 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—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 sets. Defuzzification involves the determination of the crisp output from the fuzzy outputs of the IF-THEN inference system [46].

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

3.3 Hybrid Structure of CBR Using Fuzzy Sets

Integration of fuzzy sets with CBR is important in that it helps to develop an effective method for handling vague and incomplete knowledge which statistically uses membership value of fuzzy sets in CBR. Fuzzy sets are used to describe the approximate phenomena of the real world because there is a similarity between human reasoning and natural languages compared to other existing techniques.

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.

3.3.1 Fuzzy Representation

For the case representation in the hybrid system, fuzzification using the membership function is the first process that converts inputs into the fuzzy linguistic terms and calculates the similarity of a single feature of a case with the corresponding attribute of the target [44]. Thus, constructing the fuzzy membership functions that transform the continuous input attributes into linguistic terms and the membership value for fuzzy preference 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 such as Kohonen's learning vector quantization algorithm, fuzzy c-means clustering algorithm and subtractive clustering method and so on [3][8][9]. Optimal parameters are identified by the class prototype with the shortest distance or the closest similarity that distinguishes the distinctions of given patterns.

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

The formula used to transform an input value Xi in the set [a,b] to the degree of membership in fuzzy sets, Fi (Xi), is shown as follows;

$$Fi\ (Xi) = \max(0,\ 1 - K\times|Xi-C|) \qquad (4)$$

where K: the scale factor = 2×(1-M)/(b-a)

C: the center between the boundaries

M: the value of fuzzy membership function at the boundary

The fuzzification process for fuzzy indexing and retrieving is summarized as 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 in fuzzy terms and membership values based on membership functions defined in step a.

3.3.2 Fuzzy Indexing and Retrieving

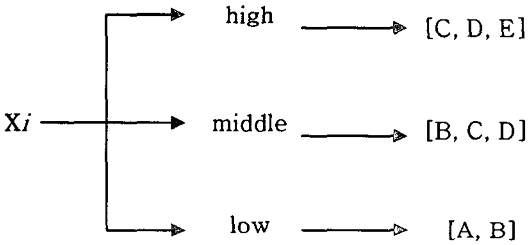
The major advantage of fuzzy indexing and retrieving is that they allow multiple class memberships to be defined on a single attribute; that is, a person may be classified as old and young at the same time with different membership grades, which would make a case qualify for the retrieving criteria of oldness and youngness [18].

Cases are indexed based on the fuzzy terms of each attributes processed by fuzzifier in a fuzzy representation step, before being stored in the case base. <Table 1> shows examples of attribute Xi represented by fuzzy terms and membership values. [Fig. 2] illustrates the indexing result of the cases in <Table 1> by their classes.

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

<Table 1> The example of fuzzy representation

Firm	Rating	Xi	Fuzzy Xi / 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
...	...	...	...



[Fig. 2] The example of fuzzy indexing

(Adapted from Jeng and Liang, 1995)

summarized as the following steps:

a. The fuzzy terms resulting from fuzzifier are used to search the candidate cases that match with a new case in the case base.

b. The one that has the closest 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 the distance function between the new and candidate cases, choose the one which has the shortest distance, and regards it as the most similar case [18].

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:

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

where x_{ai} and x_{bi} are the original value of attribute *i* for the candidate *a* and the new case *b*, *n* is the number of attributes, 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:

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

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

4. Research Development

The research data consists of 297 financial ratios and the corresponding bond rating of 1,816 Korean companies whose commercial papers have been rated from 1997 to 2000. The bond rating we employ is provided by National Information and Credit Evaluation, Inc., one of the most prominent bond rating agencies in Korea. Credit grades are classified as 5 coarser rating categories (A1, A2, A3, B, C) according to credit levels. <Table 2> shows the organization of the data set.

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.

We apply two stages of the input variable

<Table 2> Number of companies in each rating

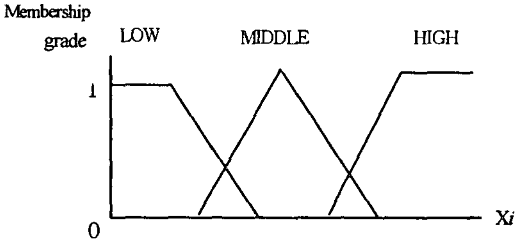
Ratings	Number of cases	%
A1	58	3.2
A2	242	13.3
A3	586	32.3
B	780	43.0
C	150	8.3
Total	1,816	100.0

<Table 3> Definition of variables

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

selection process. At the first stage, we select 106 variables by 1-way ANOVA for the numeric type and Kruskal-Wallis test for the symbolic type between each financial ratio as an input variable and credit grade as an output variable. In the second stage, we select 9 variables using a MDA stepwise method to reduce dimensionality. We select input variables satisfying the univariate test first, and then select significant variables by the stepwise method for refinement. The selected variables for this research are shown in <Table 3>.

Two qualitative variables are coded into five types according to the yearly sign transition of ordinary profit (X8) and operating activities cash flows (X9) respectively for three-year period. Each of the quantitative variables (X1-X7) is transformed into three fuzzy sets corresponding to the linguistic terms: high, middle, and low, respectively. For each numeric input variable, the fuzzy preprocessing procedure of three fuzzy sets is designed based on input data analysis. In this study in which we experiment the bond rating classification, three fuzzy sets corresponding to the linguistic terms high, middle, and low, respectively assumed for each of the input variables are shown in [Fig. 3].



[Fig. 3] The membership functions of fuzzy terms

According to the formula of computing the degree of membership in fuzzy sets, 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 consider fuzzy terms whose membership values are higher than 0.5 relatively reliable.

For example, in the case of X1 variable, three membership functions that use values generated by k-means clustering algorithm are constructed as following equations:

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

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

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

In the equations shown above, x is the input value of a financial ratio X1, 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 utilize statistical clustering algorithm such as k-means clustering algorithm and competitive artificial neural networks like Kohonen self-organizing maps to formulate the fuzzy membership functions corresponding to the categorized linguistic terms by determining the centroid, lower and upper boundary value of clusters of the input variable.

5. 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 classification problem, the results obtained are compared with those of conventional techniques such as pure-CBR model and MDA. The pure-CBR model uses a nearest-neighbor algorithm in which 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 attribute.

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 to measure the converted fuzzy membership grades between the new and candidate cases. Fuzzy-CBR[a] and Fuzzy-CBR[b] model apply 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> shows the comparison of the results of the classification techniques applied for this study. Each cell contains the accuracy of the 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 which are

<Table 4> Classification accuracies (%)

Method \ Class		A1	A2	A3	B	C	Average
MDA		60.00	20.83	50.00	42.86	53.85	43.37
Pure-CBR		20.00	29.17	60.78	52.11	41.67	49.69
Fuzzy-CBR[a]	K-means	40.00	41.67	50.98	57.75	50.00	52.15
	Kohonen	40.00	50.00	55.56	71.43	30.77	59.04
Fuzzy-CBR[b]	K-means	60.00	45.83	50.98	69.01	41.67	57.67
	Kohonen	40.00	41.67	66.67	75.71	46.15	64.46
Fuzzy-CBR[c]	K-means	40.00	41.67	62.75	74.65	25.00	61.35
	Kohonen	0.00	58.33	62.96	80.00	15.38	63.86

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

used 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.

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 5. The results of McNemar tests support that the fuzzy CBR model has higher classification accuracy than all the other comparative models with significant levels. 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 have no significant difference.

The overall result shows that the integrated models with the majority rule retrieving approach performs better than MDA and pure-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.

<Table 5> McNemar values for the comparison of performance between models

(Significance level)

		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%

6. Conclusion

In this study, we propose a hybrid approach of using fuzzy sets as 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 in that it helps to develop an effective method for handling vague and incomplete knowledge which statistically uses membership value 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 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. & Ofer, A. R *The Value of General Price Level Adjusted Data to Bond Rating*. Journal of Business Finance and Accounting, 1980, 7(1), 135-149.

[2] Belkaoui, A *Industrial Bond Ratings: A New Look* Financial Management, 1980, 9(3), 44-51.

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

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

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

[6] Buta, P *Mining for Financial Knowledge with CBR*. AI EXPERT, 1994, 9(2), 34-41.

[7] Cannon, R.L., Dave, J.V., & Bezdek, J.C *Efficient Implementation of the Fuzzy C-means Clustering Algorithms*. IEEE Trans.-PAMI, 1986, 8(2), 248-555.

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

[9] Chiu, S *Fuzzy Model Identification Based On Cluster Estimation*. J. Intell. Fuzzy Systems, 1994, 2(3), 267-278.

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

[11] Driankov, D., Hellendoorn, H., & Reinframk, M. An Introduction to Fuzzy Control. Berlin: Springer, 1993.

[12] Dutta, S., & Shekhar, S *Bond Rating: A Non-conservative Application of Neural Networks*. in: Proceedings of IEEE International Conference on Neural Networks, San Diego, CA, 1988, 443-450.

[13] Ederington, H.L *Classification Models and Bond Ratings*. Financial Review, 1985, 20(4), 237-262.

[14] Gentry, J.A., Whitford, D.T., & Newbold, P *Predicting Industrial Bond Ratings with a Probit Model and Funds Flow Components.* Financial Review, 1988, 23(3), 267-286.

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

[16] Iskandar-Datta, M.E., & Emery, D.R *An Empirical Investigation of the Role of Indenture Provisions in Determining Bond Ratings.* Journal of Banking and Finance, 1994, 18(1), 93-111.

[17] Jackson, J.D., & Boyd, J.W *A Statistical Approach to Modeling the Behavior of Bond Raters.* The Journal of Behavioral Economics, 1988, 17(3), 173-193.

[18] Jeng, B.C., & Liang, T.P *Fuzzy Indexing and Retrieval in Case-Based Systems.* Expert Systems with Applications, 1995, 8(1), 135-142.

[19] Kaplan, R.S., & Urwitz, G. *Statistical Models of Bond Ratings: A Methodological Inquiry.* Journal of Business, 1979, 52(2), 231-262.

[20] Kim, J., Weistroffer, H.R., & Redmond, R.T *Expert Systems for Bond Rating: A Comparative Analysis of Statistical, Rule-based and Neural network Systems.* Expert Systems, 1993, 10(3), 167-172.

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

[22] Klimasauskas, C.C *Hybrid Fuzzy Encoding for Improved Backpropagation Performance.* Advanced Technology for Developers, 1992, 1, 13-16.

[23] Klir, G.J., & Yuan, B. Fuzzy Sets and Fuzzy Logic-Theory and Applications. London: Prentice-Hall, 1995.

[24] Kolodner, J *Improving Human Decision Making through Case-based Decision Aiding.* AI Magazine, 1991, 12(2), 52-68.

[25] Kolodner, J. Case-Based Reasoning. San Mateo, CA: Morgan Kaufmann, 1993.

[26] Kruse, R., Gebhardt, J., & Klawonn, F. Foundations of Fuzzy Systems. New York: Wiley, 1994.

[27] 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, 1997, 6, 23-40.

[28] Liang, T.P., Chandler, J.S. & Han, I *Integrating Statistical and Inductive Learning Methods for Knowledge Acquisition.* Expert Systems with Applications, 1990, 1, 391-401.

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

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

[31] Pinches, G.E., & Mingo, K.A *A Multivariate Analysis of Industrial Bond Ratings.* Journal of Finance, 1973, 28(1), 1-18.

[32] Pogue, T.F. & Soldofsky, R.M *What's in a Bond Rating?* Journal of Financial and Quantitative Analysis, 1969, 4(2), 201-228.

[33] Quinlan, J.R *Induction of Decision Trees.* Machine Learning, 1986, 1, 81-106.

[34] Reiter, S.A., & Emery, D.R *Estimation Issues in Bond-rating Models.* Advances in Quantitative Analysis of Finance and Account, 1991, 1, 147-163.

[35] Riesbeck, C.K., & Schank, R.C. Inside Case-Based Reasoning. Hillsdale, NJ: Lawrence Erbium Associates, 1989.

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

[37] Shaw, M., & Gentry, J *Inductive Learning for Risk Classification.* IEEE Expert, 1990, 47-53.

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

[39] Shin, K.S., & Han, I *Case-based Reasoning Supported by Genetic Algorithms for Corporate Bond Rating.* Expert Systems with Applications, 1999, 16(2), 85-95.

[40] Shin, K.S., & Han, I *A Case-based Approach Using Inductive Indexing for Corporate Bond Rating.* Decision Support Systems, 2001, 32(1), 41-52.

[41] Singleton, J.C., & Surkan, A.J *Neural Networks for Bond Rating Improved by Multiple Hidden Layers.* in: Proceedings of the IEEE International Conference on Neural Networks, 1990, 2, 163-168.

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

[43] Theodoridis, S., & Koutroumbas, K. Pattern Recognition. New York: Academic Press, 1999.

[44] Watson, I *Case-based Reasoning is a Methodology Not a Technology.* Knowledge Based System, 1999, 12, 303-308.

[45] West, R.R *An Alternative Approach to Predicting Corporate Bond Ratings.* Journal of Accounting Research, 1970, 8(1), 118-125.

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

[47] Zadeh, L.A *Fuzzy Sets.* Information Control, 1965, 8, 338-353.

요약

퍼지집합이론과 사례기반추론을 활용한 채권등급예측모형의 구축

김현정 · 신경식*

최근 채권의 상환 및 이자의 확실성 정도를 측정하고 연관된 상대적인 위험의 정도를 나타내는 채권등급 평가의 중요성이 대두되고 있다. 초기의 대다수 선행 연구들에서는 기업의 채권등급예측을 위하여 통계적 기법이 많이 사용되었으나, 많은 연구들에 의해 그 우수성이 보고되고 있는 사례기반 추론 등 인공지능 기법들이 통계모형의 대안으로 제시되어지고 있다. 사례기반 추론에서는 과거의 사례들이 지식으로 표현되고 해결 방법으로 사용된다. 유용한 사례기반 시스템을 구축하기 위해서 시스템의 지식베이스를 구축할 사례들을 인간의 정보처리 과정과 유사한 방법으로 표현하는 것이 중요하다. 본 논문은 실제 세계의 애매모호한 사례들을 다루는데 적절한 퍼지집합개념을 사례기반 추론과 결합하는 통합 방법론을 제시하고자 한다. 퍼지집합이론은 인간이 의사결정시 사용하는 유사한 자연스러운 언어를 수학적으로 변환할 수 있게 해주는 인공지능 기법이다.

* 이화여자대학교 경영학과