Fuzzy Indexing and Retrieval in CBR with Weight Optimization Learning for Credit Evaluation

Cheol-Soo Park*, Ingoo Han**

(*Myongji University, **Graduate School of Management, KAIST)

Abstract

Case-based reasoning is emerging as a leading methodology for the application of artificial intelligence. CBR is a reasoning methodology that exploits similar experienced solutions, in the form of past cases, to solve new problems. Hybrid model achieves some convergence of the wide proliferation of credit evaluation modeling. As a result, Hybrid model showed that proposed methodology classify more accurately than any of techniques individually do. It is confirmed that proposed methodology predicts significantly better than individual techniques and the other combining methodologies. The objective of the proposed approach is to determines a set of weighting values that can best formalize the match between the input case and the previously stored cases and integrates fuzzy set concepts into the case indexing and retrieval process. The GA is used to search for the best set of weighting values that are able to promote the association consistency among the cases. The fitness value in this study is defined as the number of old cases whose solutions match the input cases solution. In order to obtain the fitness value, many procedures have to be executed beforehand. Also this study tries to transform financial values into category ones using fuzzy logic approach for performance of credit evaluation. Fuzzy set theory allows numerical features to be converted into fuzzy terms to simplify the matching process, and allows greater flexibility in the retrieval of candidate cases. Our proposed model is to apply an intelligent system for bankruptcy prediction.

Key words: Case-Based Reasoning, Fuzzy Set, Genetic Algorithms, Feature Transformation, Credit Evaluation

1. Introduction

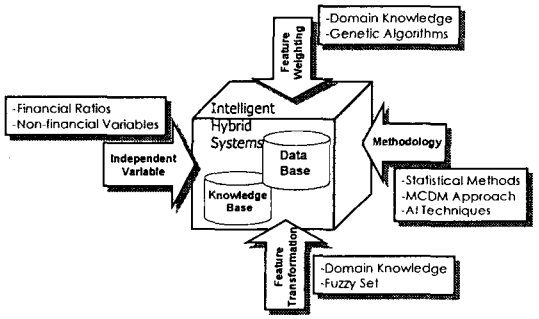
Decision-making problems in credit evaluation and its risk measurement are very important and difficult tasks for commercial banks and financial institutions due to the high level of risk associated with wrong decisions. Among these, the important risks to deal with have been a worldwide structural increase in the number of bankruptcies, more competitive margins on loans, and an increasing cost associated with monitoring solvency in order to control the risks (Wolf, 1995; Altman & Saunders, 1998).

With the growth of credit evaluation and large loan portfolios, the banking industry is actively developing more accurate credit evaluation models. These models have progressed from statistical methods to the Artificial Intelligence (AI) approach. A number of statistical methods such as multiple regression (Meyer & Pifer 1970), discriminant analysis (Altman, 1968), and logistic regression (Dimitras et al, 1996; Martin, 1997), have been typically used for financial applications including bankruptcy prediction. Recent studies in the AI approach, such as inductive learning (Han et al, 1996; Shaw & Gentry, 1998), artificial neural networks (Boritz et al, 1995; Jo & Han, 1996; Zhang et al., 1999; Coakley & Brown, 2000), and case-based reasoning (Buta, 1994; Bryant, 1997) have also been successfully applied to accounting and financial problems.

Among the challenges in successfully implementing

intelligent systems are developing effective modeling methodologies and finding ways to represent related knowledge in a domain. The modeling of credit evaluation is a complex process involving multiple strategy approach. Figure 1 gives an overview of this study. This study performs the multiple strategy modeling for credit evaluation by intelligent hybrid methodology. Intelligent combining of several good learning algorithms and their synergistic use may lead to improving predictive ability. The first approach is the selection of data for analysis. Normally, historical data is used. The data set retrieved from a single source, such a banking database. The selected data set then undergoes cleaning and preprocessing. Also, some model analysis requires data to be preprocessed to improve its quality. So, we perform the data transformation from one scale to another and identification of predictive attributes in the data set. The data set is analyzed next to identify patterns, i.e. models that represent relationships among data. There are many machine learning techniques and statistical analysis. Each technique has its own strengths and weakness. Understanding these in the context of business data mining is very useful in developing integrated methodology for credit analysis. The third approach is selection of optimal variables using statistic methodology. The last approach is to find the relative importance for each dependent variables using data driven or knowledge driven approach.

The aim of this study is to propose intelligent hybrid, and multiple strategy methodology for Case Based Reasoning(CBR) Modeling. This study employs four strategies for case based classification in the credit evaluation; feature weighting, feature transformation, development of hybrid methodology, and selection of variables. The proposed methodologies are as followed: The first, this study adopts the Genetic Algorithms (GA) to assign the importance of attributes for CBR that are considered important in the classification of credit evaluation. GA model is effective and systematic framework to obtain the feature weights and the search techniques from large and complicated spaces in a wide rang of application. The second, this study perform the feature transformation using domain knowledge and fuzzy set theory. The discretization of real value attributes is an important task in data mining, particularly for the classification problem. The third, this study propose hybrid model approach to achieve some convergence of the wide proliferation of credit evaluation modeling. As a result, this study showed that proposed methodology classify more accurately than any of techniques individually do. It is confirmed that proposed methodology predicts significantly better than individual techniques and the other combining methodologies.



<Figure 1> Multi-Dimensional Approach for Credit Evaluation

2. Research Background

2.1Classification Techniques for Case-Based Reasoning

CBR is a reasoning methodology that exploits similar experienced solutions, in the form of past cases, to solve new problems (Kolodner, 1993). When faced with a new problem, CBR will retrieve a case that is similar from a case base, and, if necessary, adapt it to provide the desired solution. A new solution is generated by retrieving and adapting an old one which approximately matches the given situation.

CBR involves: (1) accepting a new problem

representation, (2) retrieving relevant cases from a case base, (3) adapting retrieved cases to fit the problem at hand and generating the solution for it, and (4) evaluating the solution (Jeng & Liang, 1995). The key issues in the CBR process are indexing and retrieving similar cases in the case base, measuring case similarity to match the best case, and adapting a similar solution to fit the new problem. Therefore, the measurement of success of a CBR system depends on its ability to index cases and retrieve the most relevant ones in support of the solution to a new case.

As noted, successful performance of the retrieval mechanism depends on good representation, indexing and the similarity metric. The process of case retrieval for new cases is performed in three steps: (i) retrieving only candidate cases that match the important attributes of the new case; (ii) calculating an aggregate match score for the comparable cases; and (iii) retrieving those comparable cases with higher aggregate match scores.

2.1.1 Weighted k-NN Model

Among these indexing and retrieval methods, the NN matching function has been widely used in CBR for model management. The NN matching function is a non-parametric classification algorithm based on assumptions of the independence of attributes in previous cases and the availability of rules and procedures for matching. The NN techniques provide a measure of how similar a previous case is to a given problem. A primary weakness of the traditional NN function is that it is sensitive to the presence of irrelevant features in the case representation. This is because its similarity function, the Euclidean distance function, assumes that all features are equally relevant. That is, each feature has equal impact on similarity computations. The feature weighting algorithms alleviate this problem. The most relevant features are assigned the highest weights. This assigning method achieves an important improvement in classification accuracy. The overall similarity determined by a weighed NN matching function is mathematically represented as follows (Kolodner, 1993):

$$Similarity\,(T, S) = \sqrt{\sum_{i=1}^{F} w_i \times (T_i - S_i)^2}$$

where w_i is the weight of feature i, T is the target case, S is the source case, F is the number of attributes in each case, and i is an individual feature from 1 to F.

2.1.2 Review of Previous Feature Weighting Methods

Quite a few researchers have investigated empirical work on the weight setting of k-NN algorithms. Many researchers suggest that the weight of all features be acquired by domain knowledge from experts (Kolodner 1993), by machine learning techniques such as genetic algorithms (Shin & Han, 1999) and induction, by statistical methods such as multiple discriminant analysis and regression, or by AHP methodology(Park & Han, 2002).

Kibler & Aha (1987) presented a simple approach of combining a decision tree algorithm and k-NN. Their method used the presence and absence of attributes in the decision tree built by C4.5 on the same set of training examples to determine the weights in the similarity function. Wettschereck & Dietterich (1995) presented an approach of assigning continuous weights to all attributes in k-NN algorithms simply by their information gain values. Stanfill & Waltz (1986) used statistical information from the stored data to compute weights and applied the method to several tasks. Mohri & Tanaka (1994) proposed a statistical technique for calculating attribute weights, and showed that such weights are optimal in the sense that they maximize the ratio of variance between classes to variance of all cases. Wettschereck & Aha (1995) explored several weight setting methods in their comparative empirical study.

Several studies showed that one can set feature weights using another learning algorithm.

2.1.3 Review of Previous Feature Transformation

The Feature Transformation of real value attributes is an important task in data mining, particularly for the classification problem. Empirical results are showing that the quality of classification methods depends on the feature transformation algorithm used in preprocessing step. In general, feature transformation is a process of searching for partition of attribute domains into interval and unifying the value over each interval. Hence feature transformation problem can be defined as a problem of searching for a suitable set of cuts on attribute domains.

There are two reasons why category formation is a useful step in exploring a dataset: as an end in itself and as part of some other method for discovering regularities. Existing study on feature transformation in data mining has concentrated in forming categories as a means to an end rather than as an end in itself. Most of it has been motivated by the desire to extend classification tree induction methods to handle numerical variables. Recent research on feature transformation of numeric features for classification learning procedures has been reported by Catlett (1991), and Quinlan (1986) includes a useful systematic overview of this study.

There are two basic approaches - domain knowledge and data driven methodology - for feature transformation in data mining. Especially, data driven methodologies are decision tree, fuzzy logic and etc. There is a fundamental distinction between procedures that depend only on the values of the variable to be partitioned and those that also use information about the corresponding values of one or more other variables. It is term the former endogenous and the letter exogenous methods. Dougherty et al (1995) use the terms unsupervised and supervised to draw a similar but not identical distinction. Endogenous category formation procedures use only information concerning the distribution of values of the variables to be partitioned. There are two approaches: percentile methods and clustering methods. The exogenous methods transform an independent variable to maximize its association with the values of dependent and other independent variables.

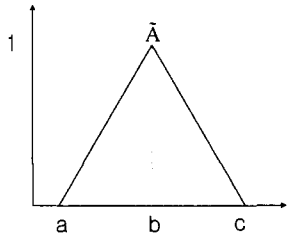
2.2 Fuzzy Logic

Fuzzy logic (Zadeh, 1965) is designed to handle imprecise 'linguistic' concepts such as *small, big, young, old, high* or *low*. Systems based on fuzzy logic exhibit an inherent flexibility and have proven to be successful in a variety of industrial control and pattern-recognition tasks ranging from handwriting recognition to credit evaluation. Central to the flexibility that fuzzy logic provides is the notion of a fuzzy set. In conventional set theory an item has a clear boundary or demarcation. A fuzzy membership diagram can be defined by a triplet (a_1, a_2, a_3) shown in <Figure 2>. Data that have been converted into fuzzy membership functions are referred to as having been 'fuzzified'. Fuzzy numbers are a fuzzy subset of real numbers, and they represent the expansion of the idea of confidence interval. According to the definition made by Dubois and Prade (1978), those numbers that can satisfy these three requirements will then be called fuzzy numbers, and the followings is the explanation for the features and calculation of the Triangular Fuzzy Numbers(TFN).

Comparing with the traditional investigative research, the importance degree for the serving attribute used 5-points of Likert Scale, applying TFN that the utilization of linguistic variables is rather widespread at the present time, and the linguistic values found in this study are primarily used to assess the linguistic rating given by the evaluators. According to the nature of TFN and the extension principle put forward by Zadeh (1965), the algebraic calculation of the TFN. The membership function is

defined as:

$$\mu_{\tilde{A}}(x)=\begin{cases}0,\\ \dfrac{x-a_1}{a_2-a_1}, x\langle a_1,\\ \dfrac{a_3-x}{a_3-a_2}, a_1\le x\le a_2, a_2\le x\le a_3, x\ge a_3.\\ 0,\end{cases}$$



<Figure 2> A triangle fuzzy numbers Ã

Alternatively, by defining the interval of confidence at level α, we can characterize the TFN as (Cheng & Mon, 1994; Kaufman & Gupta, 1991):

$\forall_\alpha \in [0,1]$,

$\tilde{A}_\alpha = \left[a_1^\alpha, a_3^\alpha\right]$

$= \left[(a_2-a_1)\alpha + a_1, -(a_3-a_2)\alpha + a_3\right]$

Some main operations for positive fuzzy numbers and described by the interval of confidence (Cheng & Mon, 1994; Kaufman & Gupta, 1991) are

$\forall a_L, a_R, b_L, b_R \in R^+$. $\tilde{A}_\alpha = [a_L^\alpha, a_R^\alpha]$,
$B_\alpha = [b_L^\alpha, b_R^\alpha]$, $\alpha \in [0,1]$
$\tilde{A} \oplus B = [a_L^\alpha + b_L^\alpha,\ a_R^\alpha + b_R^\alpha]$,
$\tilde{A} \ominus B = [a_L^\alpha - b_L^\alpha,\ a_R^\alpha - b_R^\alpha]$,
$\tilde{A} \otimes B = [a_L^\alpha\ b_L^\alpha,\ a_R^\alpha\ b_R^\alpha]$,
$\tilde{A} \oslash B = [a_L^\alpha / b_L^\alpha,\ a_R^\alpha / b_R^\alpha]$,

Where $\tilde{A}\alpha$ and α are crisp values, ⊕, ⊖, ⊗ and ⊘ denote the addition, subtraction, multiplication and division operator of two intervals of confidence, respectively. In this study, the computational technique is based on the following fuzzy numbers, which are defined in Table 1.

<Table 1> Fuzzy number and Characteristics function

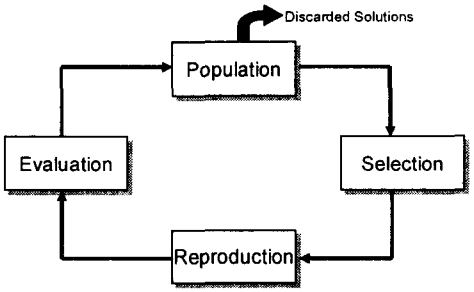
Fuzzy Number	Characteristics (or membership) function
1	(1, 1, 3)
x	(x-2, x, x+2) for x=3, 5, 7
9	(7, 9, 9)

2.3 Genetic Algorithms

Genetic Algorithms (GAs) (Holland, 1987; Goldberg, 1989) are efficient problem-solving mechanisms that are inspired by the mechanisms of biological evolution. They reward candidate solutions that contribute towards solving a problem at hand and penalize solutions that appear unsuccessful. GAs have produced very good solutions for complex optimization problems that have large numbers of parameters. Areas where these have been applied include electronic circuit layout, gas pipeline control, job shop scheduling (Davis, 1991), and credit evaluation (Kingdon & Feldman, 1995).

The main idea of a genetic algorithm is to start with a population of solutions to a problem, and then attempt to produce new generations of solutions which are better than the previous ones. This is a direct analogue of the Darwinian principle of the 'survival of the fittest' – i.e. let good solutions survive and cull bad solutions. GAs operates through a simple cycle consisting of the following stages: population creation, selection, reproduction and evaluation <Figure 3>.

The starting point for a genetic algorithm is the creation of a population of 'members' which represent candidate solutions to the problem being solved.



<Figure 3> Genetic Algorithm Cycle

A top level description of GA is presented below:

1) Initialize a population of chromosomes.
2) Evaluate each chromosome in the population.
3)Create more chromosomes by applying the crossover and mutation.
4) Delete members of the population to make room for the chromosomes.
5) Evaluate the new chromosomes and insert them into the population.
6) Repeat steps (3-5) until some problem criterion is reached

3. Fuzzy based CBR Modeling with GA Feature Weights

3.1 Fuzzy Indexing and Retrieval for Feature Transformation

Since CBR involves finding similar case from the case base and using them to construct new solutions, indexing and retrieving of cases play critical roles in case-based problem solving. Unless the cases are properly indexed and ready for retrieval, they may not be useful. Generally speaking, case indexing and retrieving are implemented on the attribute level.

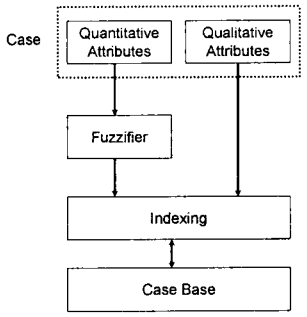
A case is composed of many attributes available for indexing. For example, the risk of a firm may be assessed by a set of financial ratios such as quick ratio and operating income to sales. Case attributes can be divided into two categories: qualitative and quantitative. Qualitative attributes accept nominal values. A firm's quality of management, for instance, is a qualitative attribute whose value may be excellent, good, average, or poor. Quantitative attributes allow values to be measured on a numerical scale.

Fuzzy indexing and retrieval are useful in domains where cases have quantitative attributes. For cases with qualitative attributes only, indexing can be performed on attributes directly. For example, the quality of management can be classified as excellent, good, average, poor, or bad (five classes). We can easily index firms by their risks or quality of management. If we also want to include quick ratio, however, indexing becomes more complicated. The value of quick ratio can be any positive real number. Like numerical attributes, it has an infinite number of possible values and is easier to index with a proper transformation into a few discrete classes.

The major value of fuzzy indexing and retrieval is that they can effectively offset in case retrieval. Fuzzy indexing and retrieval allow multiple class memberships to be defined on a single attribute.

Fuzzy indexing is a two-stage process as shown in Figure 4. Quantitative attributes are first processed by the *fuzzifier* (called *fuzzification)* and then indexed on the resulting classed (*indexing*) before being stored in the case base. The fuzzification process includes the following steps(Jeng & Liang, 1995):

1. When a case is encountered, quantitative attributes are identified;
2. For each quantitative attribute, proper classes are determined based on practical needs;
3. The membership function of each class and its associated a-cut are determined;
4. Numerical values of each case are converted into proper classes for indexing.



<Figure 4> The Fuzzy Indexing Process

Once cases are indexed and stores in the case base, they can be used for problem solving. When a new case is encountered, the CBR engine searches the case base to retrieve similar cases. The retrieval process also needs fuzzy treatment if quantitative attributes are involved. The fuzzy retrieval process includes the following steps:

1. Quantitative attributes are converted into fuzzy terms based on membership functions defined in the fuzzifier;
2. The resulting fuzzy terms combined with known qualitative attributes are used as keys for searching similar cases;
3. The matched cases are retrieved as candidates, and the one that has the highest similarity is used to construct a solution to the new case.

3.2 GA Optimization for Feature Weighting

This study proposed the integration methodology of GA and case based systems for feature weighting of the case indexing and retrieving process to the searching and

learning capabilities of evolutionary algorithms. Based upon the natural evolution concept, the GA is computationally simple and powerful in its search for improvement and is able to rapidly converge by continuously identifying solutions that are globally optimal with in a large search space. To determine a set of optimum weighting values, the search space is usually quite huge. This is because the search process must consider countless combination of variant possible weighting values for each of the feature against all of the cases stored in the case base.

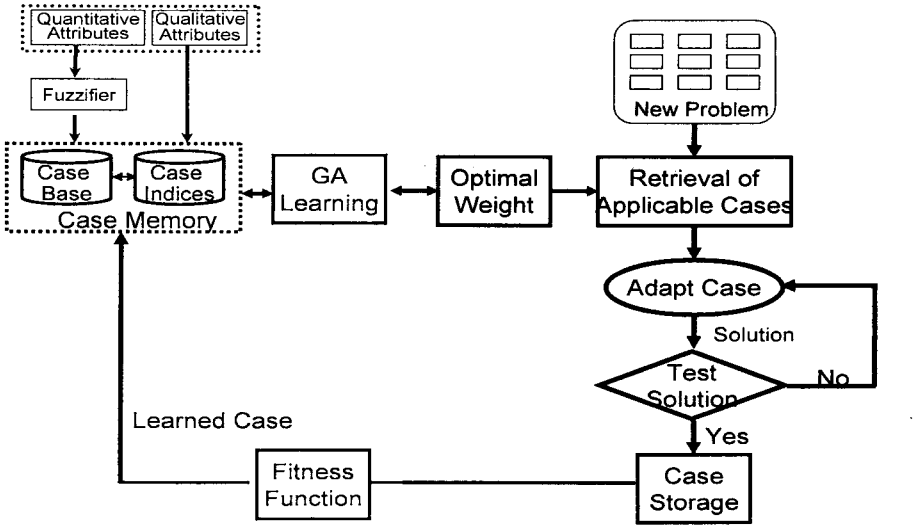
To solve a problem, the GA randomly generates a set of solutions for the first generation. Each solution is called a chromosome that is usually in the form of a binary string. According to a fitness function, a fitness value is assigned to each solution. The fitness values of these initial solutions may be poor. However, the fitness values will rise as better solutions survive in the next generation. For the controlling parameters for experiment, the GA is an iterative cycle composed by the following steps: (1) evaluation of individuals, (2) selection, (3) crossover, and (4) mutation. This iteration process is repeated until all the cases are correctly classified or a certain amount of iteration reached.

3.3 Fuzzy based k-nn Modeling with GA Optimization

The objective of the proposed approach is to determines a set of weighting values that can best formalize the match between the input case and the previously stored cases and integrates fuzzy set concepts into the case indexing and retrieval process.

The GA is used to search for the best set of weighting values that are able to promote the association consistency among the cases. The fitness value in this study is defined as the number of old cases whose solutions match the input cases solution. In order to obtain the fitness value, many procedures have to be executed beforehand.

Also this study tries to transform financial values into category ones using fuzzy logic approach for performance of credit evaluation. Fuzzy set theory allows numerical features to be converted into fuzzy terms to simplify the matching process, and allows greater flexibility in the retrieval of candidate cases. Figure 5 presented the overall structure of hybrid approach.



<Figure 5> Hybrid Framework of GA-CBR System

4. Application: Bankruptcy Prediction

4.1 Experimental Design

Our study uses the fuzzy based k-NN algorithm with GA feature weighting in the problem of bankruptcy prediction. We have seen that the weighted k-nearest neighbor algorithm can outperform the pure nearest neighbor algorithm in a specific data set (Wettschereck et al., 1997). This study, therefore, is devoted to studying methods that can improve the performance of weighted k-NN algorithm, and feature transformation. We will investigate the issue of how to assign the values of importance that will enhance the classification accuracy of weighted k-NN in the test set. And, we will examine the data transformation using the fuzzy set theory for numerical information. This is followed by a comparison of simple k-NN, k-NN with AHP weighting, and k-NN with GA weighting before data transformation and after data transformation..

The application process of the bankruptcy prediction model consists mainly of three parts: (1) sample selection and collection of data (variables and sample characteristics and size), (2) selection of a method and the specific variables (financial and non-financial variables) to develop the evaluation model, and (3) model validation, i.e. statistical significance and accuracy of results.

To study the validation of the proposed approach for bankruptcy prediction, this study presented the two results as follows: (i) predictive performance of case-based retrieval and indexing using weight vectors obtained by various feature weighting methods including statistical evaluations and AHP weights are compared with those that used the financial variables only, and (2) using the financial and non-financial variables, this study compared the empirical results of pure k-NN and various weighted k-NN methodologies.

4.2 Data and Variables

Credit evaluation is an ill-structured decision problem, which involves the analysis of a complex array of a firm's historical data. The database used in this study was obtained from the Industrial Bank of Korea, which is one of the leading banks in Korea, with the public policy role of promoting growth among small and medium-sized enterprises in the country. All of the failure cases are medium-sized firms, which went bankrupt between 1995 and 1998. There are also a small number of bankruptcy cases that can be compared to non-bankruptcy cases. Therefore, it is possible to select a sound case which is included in the same industry and of similar size as that of a bankruptcy case. For the purposes of this study, the experimental sets consisted of the same number of bankruptcy and non-bankruptcy cases. The total sample of 2144 companies includes 1072 bankruptcy and 1072 non-bankruptcy cases placed in random order.

In choosing financial variables, we apply statistical methods. Most studies that have been performed by using statistical methods, such as discriminant analysis, or logistic regression, have selected the independent variables using stepwise selection. Initially, financial variables are selected for the evaluation model by factor analysis and one-way ANOVA. We also reduce the number of financial variables into a manageable set of 13 using two different variable selection methods: stepwise and the t-test.

The selection of non-financial variables are based upon two characteristics: (a) their usefulness in previous studies and (b) experience of the experts for credit evaluation at financial institutions and banks. The 15 non-financial variables are those which were found to be significant in credit evaluation by previous studies or have been used in practice to assess credit evaluation. The 15 qualitative attributes are modeled according to an ordinal scale. All these financial variables, as well as non-

financial variables, are presented in Table 2.

<Table 2> List of Variables (Medium-Sized)

Financial Criteria	
Stability Ratios	
SETA	Stockholder's Equity to Total Assets
FSEL	Fixed assets to Stockholder's Equity and long-term
Liabilities	
QR	Quick Ratio
TBPT	Total Borrowings and bonds Payable to Total assets
Profitability Ratios	
OITA	Ordinary Income to Total Assets
OIS	Operating Income to Sales
FES	Financial Expenses to Sales
Activity Ratios	
TAT	Total Assets Turnover
OAT	Operating Assets Turnover
Productivity Ratios	
GVTF	ratio of Gross Value added to Tangible Fixed assets
GVAS	ratio of Gross Value Added to Sales
Growth Ratios	
GRPE	Growth Rate of Property, plant and Equipment
GRS	Growth Rate of Sales
Non-Financial (Qualitative) Criteria	
Business Profitability	
GP	Growth Potential
PP	Profit Perspective
MN	Market Niche/trend
IP	Industry Position
Competitive Advantage of Firms	
PS	Personnel and Staff hiring policy
TD	Technology Development and quality innovation
PC	Pricing Competitive advantage
IC	International Competitive advantage
Management Capacity	
QM	Quality of Management
RL	Relationship between Labor and capital
WC	Working Conditions and welfare facilities
Reliabilities	
PP	Past Payment record (trade)
IR	Industry Reputation
Others	
HF	History of Firm
SZ	Size

4.3 Feature Transformation based on Domain Knowledge

The discretization of real value attributes is an important task in data mining, particularly for the classification problem. Empirical results are showing that the quality of classification methods depends on the discretization algorithm used in preprocessing step. In general, discretization is a process of searching for partition of attribute domains into interval and unifying the value over each interval. Hence discretization problem can be defined as a problem of searching for a suitable set of cuts on attribute domains.

This study examines the data discretization approach based on domain knowledge and fuzzy logic. This study is to transform numeric values into discrete ones in accordance with the knowledge of experts in credit analysis domain. In the classification problems for bankruptcy prediction, it uses financial information and cross sectional data. The data discretization based on domain knowledge is classified as an endogenous method. The discretization criteria for bankruptcy prediction as follows <Table3>;

<Table 3> The definition of norms for financial variables in medium sized company

Attributes	Category				
	5	4	3	2	1
SETA	≤ 41.4	[41.3-28.2]	[28.1-19.7)	[19.6-14.9)	< 14.9
FSEL	> 56.4	(56.4-80.6)	(80.7-109.1]	(109.2-143.7]	≥ 43.8
QR	≤ 129.9	[129.8-92.8]	[92.7-62.6)	[62.5-45.9)	< 45.9
OITA	> 17.7	[17.0-30.9]	[31.0-43.1)	[43.2-53.1)	≥ 53.2
OIS	≤ 8.9	[8.8-4.9]	[4.8-2.2)	[2.1-0.8)	< 0.8
FES	≤ 11.2	[11.1-8.4]	[8.3-5.4)	[5.3-3.5)	< 3.5
TAT	≤ 1.94	[1.93-1.49]	[1.48-1.11)	[1.10-0.88)	< 0.88
OAT	≤ 6.6	[6.5-4.5]	[4.4-3.2)	[3.1-2.4)	< 2.4
GVTF	≤ 344.3	[344.2-160.9]	[160.8-90.5)	[90.4-59.6)	< 59.6
GVAS	≤ 42.7	[42.6-35.3]	[35.2-27.8)	[27.7-22.9)	< 2.9
GRPE	≤ 74.7	[74.6-24.5]	[24.4-9.1)	[9.0-3.4)	< 3.4
GRS	≤ 47.7	[47.6-16.9)	[16.7-(-0.7))	[(-0.8)-(-3.3))	<-13.3

5. Experimental Results

Table 4 describes the average classification accuracy in the holdout sample data. In this experiment, we use the pure CBR model as benchmarking. Weight based model outperform the pure CBR model. The classification results by pure retrieval with equal weights have an accuracy of 74.08% on average. Our GA weighted k-NN used the domain knowledge of each case as assigned by the expert. The classification accuracies of AHP CBR, GA CBR before feature transformation model, GA CBR after feature transformation using domain knowledge, and GA CBR after feature transformation using fuzzy set are 84.52%, 84.78, 86.32% and 86.34% respectively. Therefore, our proposed model outperformed the pure CBR.

<Table 4> Results of Classification Accuracy (Medium sized Firm)

Model	Classification Accuracy					
	1st Fold	2nd Fold	3rd Fold	4th Fold	5th Fold	Average
Pure k-NN	77.1%	68.4%	76.2%	73.6%	75.1%	74.08%
AHP weighted k-NN	83.0%	84.2%	84.4%	86.3%	84.7%	84.52%
GA weighted k-NN before FT	84.3%	85.7%	82.1%	85.6%	86.2%	84.78%
GA weighted k-NN after FT (Domain)	86.8%	84.9%	84.5%	87.4%	88.0%	86.32%
GA weighted k-NN after FT (Fuzzy)	86.3%	84.2%	85.1%	88.6%	87.5%	86.34%

We used the McNemar test to examine whether or not the classification performance of the hybrid approach is significantly higher than that of other techniques. The McNemar test is a non-parametric test of the hypothesis that two related dichotomous variables have the same mean. As we are interested in the correct classification of cases, the measure for testing is classification accuracy (the number of correct classifications to the total number of holdout samples). Table 5 shows the results of McNemar testing in comparing the classification ability between benchmark models and a proposed model (fuzzy based GA weighted k-NN Model) for holdout samples. The proposed model shown an outstanding prediction accuracy for k=10. The fuzzy based GA weighted k-NN Model performs significantly better than either the pure CBR or the AHP CBR model at the 5% level, and marginally better than the GA CBR using domain knowledge transformation at the 10% level.

<Table 5> McNemar values for the pairwise comparison of performance (Medium)

	AHPCBR	GA k-NN(FT)	GA k-NN(D)	GA-k-NN Fuzzy
Pure CBR	14.714***	15.861***	16.025***	19.211***
AHP CBR		0.235	0.652	4.667**
GA k-NN(FT)			1.258	3.431*
GA k-NN(D)				3.195*

(*** significant at the 1% level, ** significant at the 5% level, * significant at the 10% level)

6. Conclusion and Remarks

In this chapter, we have proposed a fuzzy set based approach that uses fuzzy membership functions to convert numerical attributes into qualitative terms for indexing and retrieval. We have shown that this new approach allows numerical data to be handled easily. We have also shown that the proposed approach to determine appropriate feature weighting using GA for effective case

retrieval. The results show significant promise for credit evaluation insights that are complex, unstructured, and mixed with qualitative and quantitative information. This hybrid model has not only demonstrated its better performance for prediction but also the ability to understand a model.

References

Altman, E.I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 589-609

Altman, E.I. & Saunders, A. (1997). Credit risk measurement: development over the last 20 years. *Journal of Banking & Finance*, 21(11/12), 1721-1742.

Boritz, J.E., & Kennedy, D.B. (1995). Effectiveness of neural network types for prediction of business failure, *Expert System with Applications*, 9(4), 503-512.

Buta, P. (1994). Mining for financial knowledge with CBR, *AI Expert*, 9(2), 34-41.

Bryant, S.M. (1997). A Case-based reasoning approach to bankruptcy prediction modeling. *International Journal of Intelligent Systems in Accounting, Finance and Management*, 6(3), 195-214.

Cattle, J. (1991). On chaining continuous attributes into ordered discrete attributes. In Y. Kodratoff. (ed), Machine Learning EWSL-91, Porto, Portugal, LNAI, 164-178.

Cheng, C.H. and Mon D.L. (1994). Evaluating weapon system by analytic hierarchy process on fuzzy scales. Fuzzy Sets and Systems, 63, 1-10.

Coakley, J.R., & Brown, C.E. (2000). Artificial neural networks in accounting and finance: Modeling issues. *International Journal of Intelligent Systems in Accounting, Finance and Management*, 9(2), 119-144.

Davis, L. (1991). Handbook of Genetic Algorithms, Van Nostrand Reinhold, New York.

Deboeak, G. (1994). Trading on the Edge. John WILEY & Sons, Inc.

Dimitras, A.I., Zanakis, S.H., & Zopounidis, C. (1996). A Survey of business failure with an emphasis on prediction methods and industrial applications. *European Journal of Operational Research*, 90(3), 487-513.

Dogois, D., & Prade, H. (1978). Operations od fuzzy number. International Journal of System Science, 9(6), 613-626.

Dougherty, J., Kohavi, R., and Sahami, M. (1995). Supervised and unsupervised discretisation of continuous features. In Proc. Twelfth International Conference on Machine Learning, Los Altos.

Goldberg, D.E. (1989). Genetic algorithms in search. Optimization and Macjine Learning. Addison-Wesley, Reading, MA.

Han, I., Chandler, J.S., & Liang, T.P. (1996). The impact of measurement scale and correlation structure on classification performance of inductive learning and statistical methods. *Expert System with Applications*, 10(2), 209-221.

Holland, J.H., (1987). Genetic algorithms and classifier systems: foundations and future directions. Proceedings of the 2nd International Conference on Genetic Algorithms.

Jeng, B.C., & Liang, T.P. (1995). Fuzzy indexing and retrieval in case-based system. *Expert System with Applications*, 8(1), 135-142.

Jo, H., & Han, I. (1996). Integration of case-based forecasting, neural network, and discriminant analysis for bankruptcy prediction. *Expert Systems with Applications*, 11(4), 415-422.

Kibler, D. and Aha, D.W. (1987). Learning representative exemplars of concepts: An initial case study. In *Proceedings of the International Workshop on Machine Learning*, (pp. 24-30). Irvine.

Kingdon, J., and Feldman, K. (1995). Genetic Algorithms for bankruptcy prediction. Search Space Research Report No 01-95, Search Space Ltd, London.

Kolodner, J.L. (1993). *Case-based reasoning*. San Mateo, CA: Morgan Kaufmann.

Kaufmann, A. and Gupta, M.M., (1991). Introduction to fuzzy arithmetic theory and applications. Van Nostrand Reinhold, New York.

Martin, D., (1997). Early warning of bank failure: A logit regression approach. *Journal of Banking and Finance*, 1, 249-276.

Matsatsinis, N.F., Doumpos, M., & Zopounidis, C. (1997). Knowledge acquisition and representation for expert systems in the field of financial analysis. *Expert Systems with Applications*, 12(2), 247-262.

Mohri, M. &Tanaka, H. (1994). An optimal weighting criterion of case indexing for both numeric and symbolic attributes. *Tech. Rep. WS-94-01, Case-Based Reasoning: Papers from the 1994 Workshops*. Menlo Park, CA: AAAI Press.

Park, C.S,. & Han, I,G. (2002). A case-based reasoning with the feature weights derived by analytic hierarchy process for bankruptcy prediction. *Expert Systems with Applications*, 20(3), 255-264.

Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1, 81-106.

Shaw, M. & Gentry, J. (1998). Using an expert system with inductive learning to evaluate business loans. *Financial Management*, 17(3), 45-56.

Shin, K.S., & Han, I.G. (1999). Case-based reasoning supported by genetic algorithms for corporate bond rating. *Expert Systems with Applications*, 16(2), 85-95.

Wettschereck. D., & Aha, D.W. (1995). Weighting features In *Proceedings of the First International Conference on Case-Based Reasoning* (pp. 347-358). Sesimbra

Wolf, M.F. (1995). New technologies for customer rating: Integration of knowledge-based systems and Human judgment. *Intelligent Systems in Accounting, Finance and Management*, 4(4), 289-301.

Zadeh, L.A. (1965). Fuzzy sets. Information and Control, 8, 338-353.

Zhang, G., Hu, M.Y., Patuwo, B.E., & Indro, D.C. (1999). Artificial neural networks in bankruptcy prediction: general framework and cross-validation analysis. *European Journal of Operational Research*, 116(1), 16-32.