Association for Information Systems
AIS Electronic Library (AISeL)

PACIS 2000 Proceedings

Pacific Asia Conference on Information Systems (PACIS)

December 2000

A Case-based Approach using Inductive Learning for Corporate Bond Rating

Kyung-Shik Shin
Ewha Womans University

Ingoo Han
Korea Advanced Institute of Science and Technology

A Case-based Approach Using Inductive Learning for Corporate Bond Rating

Kyung-shik Shin
College of Business Administration,
Ewha Womans University, Seoul 120-750, Korea

Ingoo Han
Graduate School of Management,
Korea Advanced Institute of Science and Technology, Seoul, Korea

Abstract
Case-based reasoning is a problem solving technique by re-using past cases and experiences to find a solution to the problems. The central tasks that CBR methods have to deal with are to identify the current problem situation, find a past case similar to the new one, use that case to suggest a solution to the current problem, evaluate the proposed solution, and update the system by learning from this experience. Among these tasks, one of the critical issues in building useful CBR system lies in indexing of cases that supports the retrieval of relevant cases to the problem.
This paper investigates the effectiveness of integrated approach using induction techniques to case indexing process for business classification tasks. We suggest this approach as a unifying framework to combine general domain knowledge and case specific knowledge. The proposed approach is demonstrated by applications to corporate bond rating.

Keywords: Corporate Bond Rating, Case-based Reasoning, Inductive Learning

1. Introduction

Case-based reasoning (CBR) is a problem solving technique that is fundamentally different from other major AI approaches. Instead of relying on making associations along generalized relationships between problem descriptors and conclusions, CBR benefits from utilizing case specific knowledge of previously experienced problem situations. A new problem is solved by finding a similar past case and reusing it in the new problem situation.

Wide range of applications of CBR have been reported (Brown and Gupta, 1994; Chi *et al*., 1993; Hansen *et al.*, 1995; Mechitov *et al.*, 1995; Morris, 1994; O'Roarty *et al.*, 1997; Riesbeck and Schank, 1989), including business classification for decision making such as bond rating (Buta, 1994; Shin and Han, 1998; Shin *et al.*, 1997) and bankruptcy prediction (Bryant,1997).

The central tasks that CBR methods have to deal with are to identify the current problem situation, find a past case similar to the new one, use that case to suggest a solution to the current problem, evaluate the proposed solution and update the system by learning from this experience (Kolodner, 1993; Riesbeck and Schank, 1989; Slade, 1991).
Among these major tasks, one of the major issues lies in the retrieval of appropriate cases (Hansen *et al.*, 1995). An index used to retrieve cases from memory may fail even if there is a relevant case in memory (Kolodner, 1991). This happens when the index does not

correspond to the one used to index the case. For this reason, integration of general domain knowledge into case indexing and retrieving processes is highly recommended in building a successful CBR system. The indexing problem (Kolodner, 1993) refers to the task of storing cases for an effective and efficient retrieval.

In this paper, we discuss the implementation of effective indexing methods to build case-based system. Our particular interest is an integrated approach using induction technique and CBR to retrieve more relevant cases. This approach is aimed at unifying case-specific and general domain knowledge within the system. The proposed approach is demonstrated by applications to corporate bond rating.

This paper is organized as follows. The following section provides a brief description of prior research on corporate bond rating studies. Section 3 explains the integrated approach for an effective CBR system. Section 4 and 5 report the experiments and empirical results of corporate bond rating application. The final section discusses the conclusions and future research issues.

2. Prior Research on Bond Rating

Corporate bond rating informs the public of the likelihood of an investor receiving the promised principal and interest payments associated with the bond issues. Bond ratings by independent rating agencies characterize the risk of the investments and affect the cost of borrowing for the issuer. Since market yields correspond to bond ratings, indicating an association between rating and risk, the study of the rating process is of interest not only to bond issuers but to investors as well.
Numerous bond rating studies have traditionally used statistical techniques such as ordinal least squares (OLS) (Horrigan, 1966), multiple discriminant analysis (Pinches and Mingo, 1973), and logit (Ederington, 1985) models.

Recently, however, a number of studies have demonstrated that artificial intelligence approaches can be used as alternative methodologies for bond rating applications. Shaw and Gentry (1990) applied inductive learning method to risk classification application and found that inductive learning's classification performance was better than probit or logit analysis. They have concluded this result can be attributed to the fact that inductive learning is free from parametric and structural assumptions that underlie statistical methods.

Back-propagation neural networks have been found to be successful predictors for business classification problems. 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 that the developed models are 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 discriminant analysis.

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% of accuracies 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 (MDA) (58%-61%).

Few studies have applied case-based reasoning for bond rating. Buta (1994) developed a CBR that predicts corporate bond rating. Using an inductive indexing scheme, Buta showed an accuracy of 90.4%.

Shin *et al.* (1997) developed a corporate bond rating model using Korean bond rating data. They applied case-based reasoning using an inductive indexing method without controlling the depth of the trees. Despite the optimistic hope that inductive indexing methods can improve the effectiveness of case reasoning resulting higher classification accuracy, the experimental results were rather disappointing. Although the proposed model failed in respect to classification accuracy, the exercise has provided some valuable insights. That is, the success of the case-based reasoning system using an inductive indexing approach largely depends on the appropriateness of induction trees, underlining the necessity of optimizing decision trees.

The recent study of Shin and Han (1998) proposed a new hybrid approach using genetic algorithms to case-based retrieval process in an attempt to increase the overall classification accuracy. 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.

3. Integrating Induction Technique and Case-based Reasoning

Inductive learning and case-based reasoning are classification techniques that can be

applied in financial decision making. Induction and case-based reasoning techniques are compared by considering that the first technique makes direct use of past experiences at the problem solving stage while the second one only uses an abstraction of the cases. Induction compiles past experiences into general knowledge, which are then used to solve problems. Case-based reasoning directly interprets past experiences.

The integration of the induction technique and case-based reasoning reaps the benefits of both systems for the following reasons. First, we can retrieve more relevant cases through generalized domain knowledge derived by the induction technique. CBR has some drawbacks in that the technique ignores general knowledge for the domain. This means useful information may not be utilized. Since the induction method extracts explicit knowledge from the data, CBR can benefit from an integrated approach. Second, the integrated approach can enhance the efficiency of the system because only a small subset of data needs to be considered during retrieval.

Several inductive clustering methods can be used to do this job. Induction algorithms such as ID3 (Quinlan, 1986) and CART (Classification And Regression Trees), determine which features do the best job in discriminating cases and generate a tree type structure to organize the cases in memory.

We integrate the inductive clustering technique and case-based reasoning by using an inductive indexing scheme. As the first step, we build a decision tree for case indexing. A decision tree is built upon a database of training cases. For example, the partitioning procedure of ID3 uses a preference criterion based on the entropy measure of information gain. At each node in the induction tree, the information gain is evaluated for all the attributes that are relevant and the one which yields the highest increase is selected.

The success of inductive indexing approach, however, largely depends on the appropriateness of decision trees for case retrieval (Kolodner, 1993; Shin *et al.*, 1997). To find an optimal or near optimal decision tree, we apply four different stopping conditions for the tree. The stopping criterion denotes the maximal depth of the tree, defining the maximal number of levels an induction tree can have. Figure 1 illustrates the levels of depth of the decision tree.

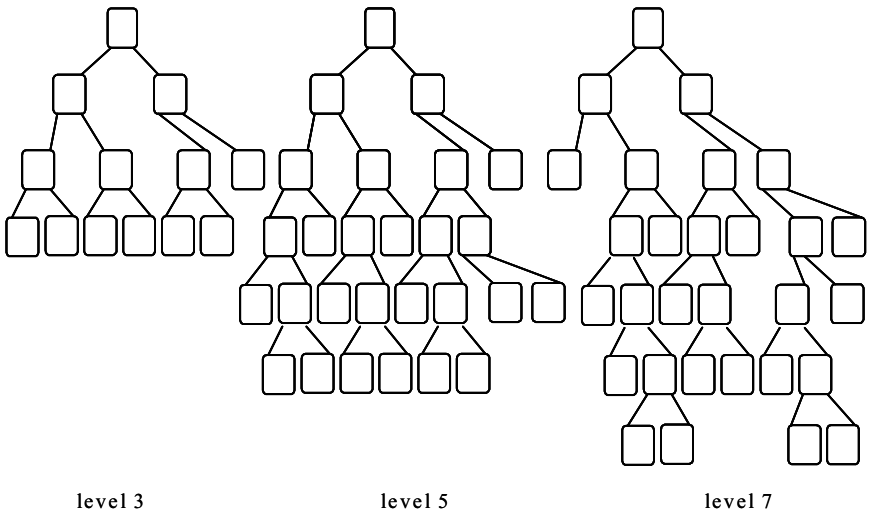

Figure 1. Depths of the decision tree

The first model integrates an induction tree having 3 levels of depth and case-based reasoning by applying the nearest-neighbor algorithm at the end of induction tree. This allows to determine the most similar cases to the current situation, and to choose the most probable value in this subset of cases. The second, third and fourth models follow the same procedure except that the induction trees have 5, 7 and 9 levels of depth, respectively. As a partitioning criterion, the information gain measure based on Shannon's entropy (Shannon and Weaver, 1947) is used. Figure 2 illustrates the hybrid structure of IND-CBR system.

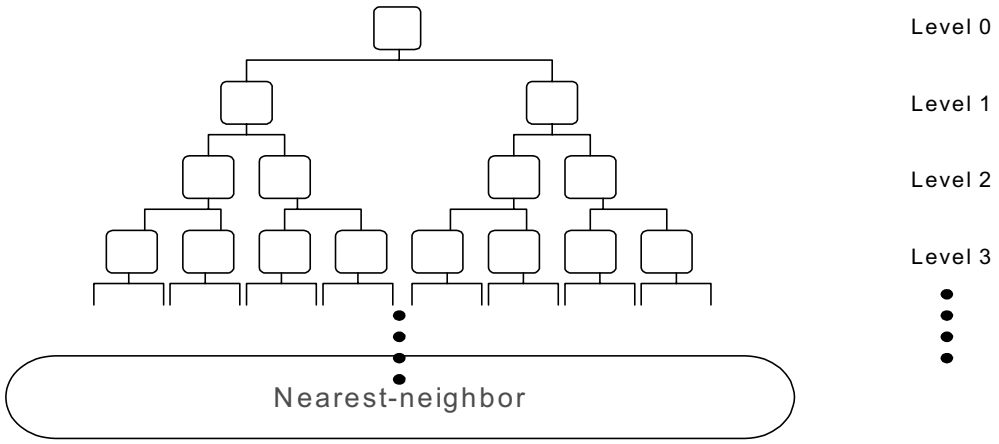

Figure 2. The hybrid structure of IND-CBR system

4. Experimental Designs and Data

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

Table 1. Number of companies in each rating

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

We apply two stages of input variable selection process. At the first stage, we select 27 variables (23 quantitative / 4 qualitative) by factor analysis, 1-way ANOVA (between input variable and credit grade as output variable) and Kruskal-Wallis test (for qualitative variables). In the second stage, we select 12 financial variables (10 quantitative / 2

qualitative) using stepwise method of MDA to reduce the dimensionality. We select input variables satisfying the univariate test first, and then select significant variables by stepwise method for refinement. In choosing qualitative variables, the four variables have been initially selected. However, audit opinion and audit firm are excluded by the expert's opinion. Two of the four qualitative variables selected are firm classification by group (conglomerate) types and firm types. We classify conglomerates into five categories, namely, top-ten conglomerates, top-twenty conglomerates, top-thirty conglomerates, top-forty conglomerates and non-conglomerates. The four types of firms are: listed, registered, externally audited and others. Table 2 illustrates the selected variables for this study.

Table 2. Definition of variables

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

Each data set is split into two subsets, a reference set and a validation (holdout) set. The reference data are used to form a decision tree to index cases and also as a case base for retrieval. The validation data are used to test the model's results with the data which have not been used to develop the system. The number of the reference cases and the validation cases are 3,486 and 400, respectively.

5. Results and Analysis

To study the effectiveness of the integrated approach for case indexing in the context of the corporate bond rating problem, the results obtained are compared with those of other indexing techniques such as CBR-pure model and CBR-expert model. The CBR-pure model uses a nearest-neighbor algorithm that has equal weights among attributes. The CBR-expert model applies importance weights assigned by experts.

For this experiment, we have had experts designate the importance of an attribute by assigning the 5 qualitative values by interview. We have selected 5 experts, three from the bond rating department of a credit rating agency, and two from the credit analysis department of a commercial bank. The selected experts' work experience related to credit analysis ranged from two to eight and half years while the average of experience is 4 years and 2 months. The five qualitative values are: "most important," "very important," "important," "less important," and "ignored" which are associated with numbers for computation as 1.0, 0.8, 0.6, 0.4, and 0.2, respectively. Table 3 shows the assigned

importance to each attribute by expert opinion. Importance values are ranged from 0.4 to 0.88.

Table 3. Importance weights assigned by experts

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

As mentioned above, we apply four predetermined stopping conditions. Integrated model (1) applies the decision tree which has 3 levels of depth. Integrated model (2), (3) and (4) follow the same procedure except the decision trees have 5, 7 and 9 levels of depth, respectively. Table 5 represents different stopping conditions and the corresponding figures of the decision tree. The decision trees are built with the software package KATE™.

Table 4. Figures depend on different stopping conditions

Model	Stopping condition (depth)	Number of nodes	Number of leaves	Number of cases per leaf	Average number of questions	Average depth
Integrated (1)	3	15	8	435.8	2	3
Integrated (2)	5	63	32	108.9	3.4	5
Integrated (3)	7	229	115	30.3	4.7	6.9
Integrated (4)	9	511	256	13.6	5.5	8.4
Full (no condition)		1,491	746	4.67	6.9	11.8

As shown in Table 4, the important figures of the decision tree are dramatically affected by different stopping conditions. For example, the number of leaves increases from 8 to 746 depending on the depth of tree. Since the number of leaves corresponds to the number of inductive clusters for case organization, 8 and 746 leaves denote 8 and 746 clusters for cases, respectively. Since the main role of induction in this context is to extract general domain knowledge from database, we can easily expect that a higher number of clusters

does not ensure an effective case-based model.

Table 5. Classification accuracies (%)

Methods		A1	A2	A3	B	C	Avg.
MDA		57.7	69.8	58.3	55.0	77.8	60.0
ID3		65.4	55.8	47.5	72.9	33.3	59.0
CBR	Pure	65.4	66.3	58.3	66.4	0.0	62.0
	Expert	69.2	65.1	58.3	63.6	0.0	61.0
	Integrated (1)	84.6	74.4	64.7	70.7	22.2	69.3
	Integrated (2)	80.8	72.1	66.9	72.9	22.2	70.0
	Integrated (3)	76.9	62.8	63.3	73.6	11.1	66.5
	Integrated (4)	61.5	61.6	58.3	71.4	11.1	62.8

Table 5 represents 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 popular classification techniques such as multiple discriminant analysis (MDA) and ID3 are also presented as benchmarks to verify the applicability of the proposed model to the domain.

Among the techniques, the integrated models have the highest level of accuracies (Integrated (2): 70.0%, Integrated (1): 69.3%, Integrated (3): 66.5%) in the given data sets, followed by CBR-pure model (62.0%). MDA and ID3 have similar levels of classification accuracy. As we expected, the classification accuracies are affected by the depth of the decision tree. Figure 3 shows that corresponding accuracies depend on the depth of the trees.

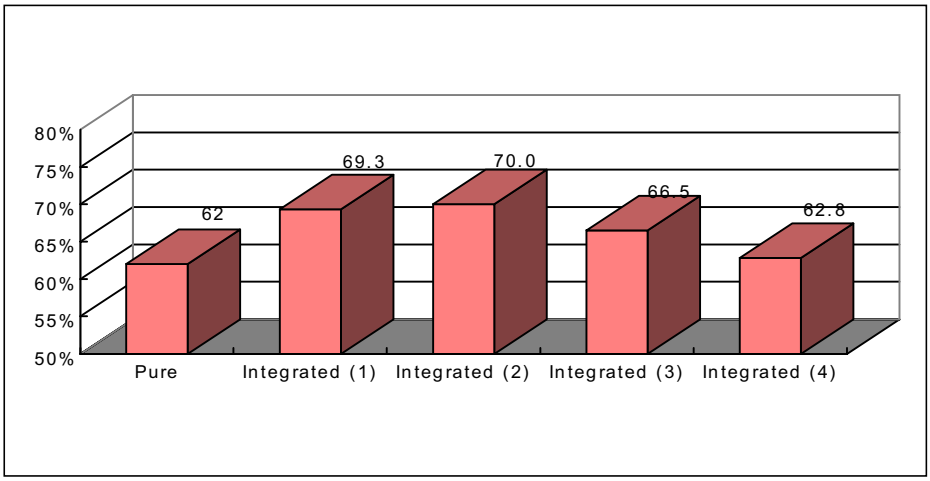

Figure 3. Classification accuracies by integrated models (%)

A comparison of integrated models indicates that higher level of depth in induction does not guarantee higher performance for integration. The accuracies of integrated model (3) and (4) decrease as the depth of the decision tree increases. This result underlines the necessity of optimizing decision trees to apply in a case-based retrieval and not simply leaving it to the induction technique itself.

We use the McNemar tests to examine whether the predictive performance of integrated approach is significantly higher than that of other techniques. McNemar test is a nonparametric test of the hypothesis that two related dichotomous variables have the same means. This test is useful for detecting changes in responses due to experimental intervention in 'before and after' designs using the chi-square distribution. Since we are interested in the correct prediction of cases, the measure for testing is the classification accuracy rate (the number of correct classification from the number of whole holdout samples).
Table 6 shows the results of McNemar tests to compare the classification ability between benchmark models and an integrated model (2) using a decision tree which has 5 levels of depth for holdout samples.

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

	MDA	ID3	CBR-Pure	CBR-Expert	CBR-Integrated (2)
MDA	-	0.0500[a]	0.2988	0.0529	9.7500 ***
ID3	-	-	0.7857	0.3182	12.6644 ***
CBR-Pure	-	-	-	0.1000	8.3904 ***
CBR-Expert	-	-	-	-	8.1441 ***

[a.] Chi-square values / * significant at 10% / ** significant at 5% / *** significant at 1 %

The result shows the integrated model (2) performs significantly better than every benchmark model proposed for this study a 1% significance level. Based on the results, we conclude that the integrated approach using induction is effective, enhancing the overall classification accuracy of the case-based system, for the application domain.

7. Concluding Remarks

This paper examined the potential effectiveness of using an induction technique to support case-based reasoning for classification tasks. In this approach, the induction technique is used to cluster and organize cases for an efficient and effective retrieval. Our experimental results have shown that this approach support an effective retrieval of cases and increases overall classification accuracy significantly.

Findings of this study are as follows. First, this integration reaps the benefits of both systems, ensuring practical applicability to the domain. That is, the induction technique provides general knowledge for the application domain, and CBR extracts case-specific knowledge through case retrieving. The experimental results of the bond rating problem support this finding. Second, the inductive clustering technique is an effective method of knowledge extraction for case-based systems, although we have burdensome tasks to optimize the induction tree.

This study has a few limitations that need further research. First, the determination of

decision trees using the different stopping conditions has a critical impact on the performance of the system. However, we did not suggest theoretically sound procedures to determine optimal stopping conditions including the depth and the criterion itself. We plan to find general method to determine stopping conditions for future research. This includes the issue of more appropriate stopping criterion than the depth of trees such as information gain measure.

The second issue for future research related to the use of more accountable target variables. The target variables used for the study are ratings by credit analysts of bond rating agency. This means that, if those historical ratings that human raters evaluated do not correspond to the correct credit level of the company, the system also can not provide correct credit information to the users. To overcome this limitation, use of more accountable target variables that represent the credit level of the company correctly may be considered for the future research.

The aim of integrating different techniques is to make more powerful and efficient systems by taking advantage of the strength of each technique. Therefore, developing more effective methods using synergistic integration is continuing research issue for the future.

References

Brown, C. E., and Gupta, U. G. "Applying case-based reasoning to the accounting domain," *Intelligent Systems in Accounting, Finance and Management,* (3), 1994, pp. 205-221.

Bryant, S. M. "A case-based reasoning approach to bankruptcy prediction modeling," *Intelligent Systems in Accounting, Finance and Management,* (6), 1997, pp. 195-214.

Buta, P. "Mining for Financial Knowledge with CBR," *AI EXPERT*, (9:2), 1994, pp. 34-41.

Chi, R. T., Chen, M., and Kiang, M. Y. "Generalized case-based reasoning system for portfolio management," *Expert Systems with Applications,* (6:1), 1993, pp. 67-76.

Dutta, S., and Shekhar, S. "Bond rating: A non-conservative application of neural networks," *Proceedings of IEEE International Conference on Neural Networks,* (2), San Diego, CA, 1988, pp. 443-450.

Ederington, H. L. "Classification models and bond ratings," *Financial Review*, (20:4), 1985, pp. 237-262.

Hansen, J., Meservy, R. D., and Wood, L. E. "Case-based reasoning: application techniques for decision support," *Intelligent Systems in Accounting, Finance and Management,* (4), 1995, pp. 137-146.

Horrigan, J. O. "The determination of long term credit standing with financial ratios," *Journal of Accounting Research,* supplement, 1966, pp. 44-62.

Kim, J., Weistroffer, H. R., and Redmond, R. T. "Expert systems for bond rating: A comparative analysis of statistical, rule-based and neural network systems," *Expert Systems*, (10:3), 1993, pp. 167-172.

Kolodner, J. “Improving human decision making through case-based decision aiding,” *AI Magazine,* (12:2), 1991, pp. 52-68.

Kolodner, J. *Case-Based Reasoning,* Morgan Kaufmann, San Mateo, CA, 1993.

Kwon, Y. S., Han, I. G., and Lee, K. C. “Ordinal pairwise partitioning (OPP) approach to neural networks training in bond rating,” *Intelligent Systems in Accounting, Finance and Management*, (6), 1997, pp. 23-40.

Maher J. J., and Sen, T. K. “Predicting bond ratings using neural networks: A comparison with logistic regression,” *Intelligent Systems in Accounting, Finance and Management,* (6), 1997, pp. 59-72.

Mechitov, A. I., Moshkovich, H. M., Olson, D. L., and Killingsworth, B. “Knowledge acquisition tool for case-based reasoning system,” *Expert Systems with Applications,* (9:2), 1995, pp. 201-212.

Moody, J., and Utans, J. “Architecture selection strategies for neural networks application to corporate bond rating,” in Refenes, A. (ed.), *Neural Networks in the Capital Markets*, John Wiley, Chichester, 1995, pp. 277-300.

Morris, B. W., “SCAN: A case-based reasoning model for generating information system control recommendations,” *Intelligent Systems in Accounting, Finance and Management,* (3), 1994, pp. 47-63.

O'Roarty, B., Patterson, D., McGreal, S., and Adair, A. “A case-based reasoning approach to the selection of comparable evidence of retail rent determination,” *Expert Systems with Applications,* (12:4), 1997, pp. 417-428.

Pinches, G. E., and Mingo, K. A. “A multivariate analysis of industrial bond ratings,” *Journal of Finance*, (28:1), 1973, pp. 1-18.

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

Riesbeck, C. K., and Schank, R. C. *Inside Case-Based Reasoning*, Lawrence Erlbaum Associates, Hillsdale, NJ, 1989.

Shannon, C. E., and Weaver, W. *The Mathematical Theory of Computation,* University of Illinois Press, 1947.

Shaw, M., and Gentry, J. “Inductive learning for risk classification,” *IEEE Expert*, February 1990, pp. 47-53.

Shin, K. S., and Han, I. “Using genetic algorithms to support case-based reasoning: Application to corporate bond rating,” *Proceedings of 2nd Asia-Pacific Decision Sciences Institute Conference*, Taipei, 1998, pp. 341 - 344.

Shin, K. S., Shin, T. S., and Han, I. “Using induction technique to support case-based reasoning: a case of corporate bond rating,” *Proceedings of MS/OR Society Conference*,

Seoul, Korea, 1997, pp. 199-202.

Singleton, J. C., and Surkan, A. J. “Neural networks for bond rating improved by multiple hidden layers,” *Proceedings of the IEEE International Conference on Neural Networks,* (2), 1990, pp. 163-168.

Singleton, J. C., and Surkan, A. J. “Bond rating with neural networks,” in Refenes, A.N. (ed.), *Neural Networks in the Capital Markets,* London Business School, England 1995.

Slade, S. “Case-based reasoning: A research paradigm,” *AI Magazine,* (12:1), 1991, pp. 42-55.