한국경영과학회/대한산업공학회 '97 춘계공동학술대회, 1997년 4월 25일-26일, 포항공과대학교

Using Induction Techniques to Support Case-Based Reasoning : A Case of Corporate Bond Rating

Kyung-shik Shin, Taeksoo Shin and Ingoo Han

Graduate School of Management, Korea Advanced Institute of Science and Technology,
207-43 Cheongryangri-Dong, Dongdaemoon-Gu, Seoul, 130-012, Korea.
Telephone : 02-958-3673, Fax: 02-958-3604, e-mail: ingoohan@msd.kaist.ac.kr

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 is able to benefit from utilizing the specific knowledge of previously experienced, concrete problem situations. A new problem is solved by finding a similar past case, and reusing it in the new problem situation [9].

Central tasks that case-based reasoning methods have to deal with are to identify the current problem situation, find a past case similar to the new one, use that case to suggest a solution to the current problem, evaluate the proposed solution, and update the system by learning from this experience[7, 12, 13].

Among these major tasks, one of the biggest issue lies in the retrieval of appropriate cases [6]. An index used to retrieve cases from memory may fail even if there is a relevant case in memory [7]. This happens when the index does not correspond to the one used to index the case. The indexing problem [13] refers to the task of storing cases for effective and efficient retrieval.

In this paper, we discuss implementation of effective indexing methods to solve classification tasks. Our particular interest is an integrated approach using induction technique to retrieve more relevant cases. We demonstrate our method and results applying to the corporate bond rating problem. Despite the optimistic hope that integration methods can improve effectiveness of case reasoning resulting higher classification accuracy, the experimental results are rather disappointing in the corporate bond rating application. However, the present exercise suggests some valuable insights.

INDEXING AND RETRIEVING

Case indexing involves assigning indices to cases to facilitate their retrieval[9]. Indices organize and label cases so that appropriate cases can be found when needed. Analysis of remindings coupled with experience in building case-based reasoning systems has led the CBR community to propose several guidelines for choosing indexes for particular cases: predictiveness, usefulness, abstractness and concreteness. Both manual and automated methods have been used to select indices. Choosing indices manually involves deciding a case's purpose with respect to the aims of the reasoner and deciding under what circumstances the case will be useful.

The second issue of indexing cases is how to structure the indices so that search through case library can be done efficiently and accurately. Given a description of a problem, a retrieval algorithm, using the indices in the case-memory, should retrieve the most similar cases to the current problem or situation. The retrieval algorithm relies on the organization of the memory to direct the search to potentially useful cases.

Indices can either index case features independently for strictly associative retrieval or arrange cases from most general to most specific for hierarchical retrieval [2]. Weighted links are used for associative retrieval. This approach involves the assessment of similarity between stored cases and the new input case, based on matching a weighted sum of features. The biggest problem here is to determine the weights of the features. The limitation of this approach include problems in converging on the correct solution and retrieval times. In general the use of this method leads to the retrieval time increasing linearly with the number of cases. Therefore this approach is more effective when the case base is relatively small.

When a case library is large, there is a need to organize cases hierarchically so that only some

small subset needs to be considered during retrieval. This subject, however, must be likely to have the best-matching or most useful cases in it [9]. In hierarchical retrieval, cases are stored in a decision tree where the top node contains common properties of all cases while nodes further down the tree are indexed based on their differences with other cases. An intermediate node represents a general description of cases under that particular node [2].

Several inductive clustering methods [4, 5] can be used to do this job. Inductive clustering methods generally look for similarities over a series of instance and form categories based on those similarities [9].

INTEGRATING INDUCTION TECHNIQUE AND CASE-BASED REASONING

Induction is a technology that automatically extracts knowledge from training samples. Induction algorithms determine which features do the best job in discriminating cases, and generate a tree type structure to organize the cases in memory [1].

An induction tree is built upon a database of training cases. The partitioning procedure uses a preference criterion based on the information gain. At each node in the induction tree, the information gain is evaluated for all the attributes which are relevant and the one is picked which yields the highest increase if the information gain measure. This approach is useful when a single case feature is required as a solution, and where that case feature is dependent upon others. An induction tree that is induced for sample cases is illustrated in Figure 1.

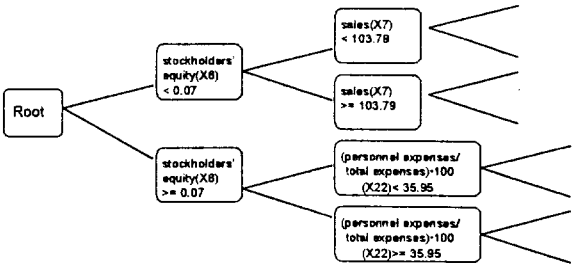

Figure 1. An induction tree (partial)

We can compare induction and case-based reasoning techniques by considering that the first technique makes direct use of past experience at the problem solving stage while the second one only uses an abstraction of the cases. In other words, induction compiles past experiences into general heuristics which are then used to solve problems. Case-based reasoning directly interprets past experience. [11]

We believe that the integration of induction and case-based reasoning is one very interesting issue for improving the experience-based expert systems. We can retrieve more relevant case through generalized concept descriptions by building a case-based reasoner that uses inductive techniques.

In this study, we apply the case-based reasoning at the end of the induction tree. This allows to determine the most similar cases to the current situation, to choose the most probable value in this subset of cases.

CASE STUDY : CORPORATE BOND RATING APPLICATION

Data and Variables

The sample consists of financial ratios and corresponding bond ratings of Korean companies. The ratings have been performed by National Information and Credit Evaluation, Inc., one of the most prominent bond rating agency in Korea. The total sample available includes 2,651 companies whose commercial papers have been rated in 1991, 1992, 1993, 1994, and 1995, respectively. Credit grades are defined as outputs and classified as 5 grade groups according to credit levels as follows. Table 1 shows the organization of data set.

Table 1. Number of companies in each rating

year	A1	A2	A3	B	C	total
91	-	-	-	-	83	83
92	55	184	276	275	68	858
93	51	153	279	332	8	823
94	53	183	257	294	3	790
95	7	20	32	34	4	97
sum	166	540	844	935	166	2,651

We devide data into two sets: the reference data (2,551samples) and the test data (100 samples). The reference data are used for constructing induction tree and case base for case-based reasoning. The test data are used for evaluating the established model.

In this study, 27 input variables (4 qualitative, 23 quantitative) are selected by stepwise regression method, factor analysis, 1-way ANOVA (between input variable and credit grade as output variable), Kruskal-Wallis test and expert's opinion. Especially, 4 qualitative input variables are selected by Kruskal-Wallis test and expert's opinion. 23 quantitative input variables are selected by 1-way ANOVA, stepwise regression model, factor analysis, and expert's opinion. To reduce the dimensionality,

we select 27 financial variables from the total variables using both the statistical method and the expert's opinions method. This combinational input variable selection approach is to select input variables satisfying both the significant levels of two methods, and then select significant variables by stepwise regression method. Table 2 illustrates selected variables for this study.

Table 2. A list of selected input variables

Var.	Name
X1	firm classification by group types
X2	firm types
X3	auditing opinion
X4	auditing company
X5	total assets
X6	stockholders' equity
X7	sales
X8	year after founded
X9	(operating income/business capital)*100
X10	gross profit to sales
X11	cash flow to total stockholders' equity
X12	this year trends when gross loss happens
X13	(stockholder's equity/total assets)*100
X14	net cash flow to total assets
X15	(investment + other assets + fixed assets)/(fixed liability + stockholders' equity)*100
X16	net cash flow trends for 3 years
X17	equipment investment efficiency
X18	(financial expenses/sales)*100
X19	dependence for liability
X20	(sales cost/sales)*100
X21	liability repayment parameter
X22	(personnel expenses/total expenses)*100
X23	incremental personnel expense rate per employee
X24	(short-term liability/total liability)*100
X25	(depreciation/ total expenses)*100
X26	quick ratio
X27	working capital turnover

RESULTS

To study the effectiveness of integrated approach for the corporate bond rating problem, the results obtained by applying induction and case-based reasoning method are compared with results from individual classification techniques. Results for multiple discriminant analysis are presented as benchmarks for the other techniques.

We apply two stopping conditions to construct induction tree. The stopping criterion is the maximal depth of the tree. This criterion defines the maximal number of levels an induction tree shall have. We use 5 for Induction (1) model and 10 for Induction (2) model. The induction trees we use are built by KATE tool [10], which uses the information gain measure based on Shannon's entropy.

Retrieving method applied in pure CBR model is nearest-neighbor retrieval. Varying numbers in CBR denote the number of nearest neighbors retrieved. In general, classification accuracy is affected by the number of cases retrieved. Table 3 presents the comparison of the results of the single classification techniques. Each cell contains the accuracy of the various classification techniques.

Table 3. Classification accuracy of the individual techniques (%)

		A1	A2	A3	B	C	Avg.
MDA		70.5	66.1	53.6	60.3	40.0	58.6
Induction (1)		28.6	30.0	65.6	70.6	14.3	54.0
Induction (2)		42.9	55.0	50.0	64.7	14.3	53.0
Pure CBR	1	71.4	70.0	62.5	67.7	14.3	63.0
	2	71.4	50.0	65.6	73.5	28.6	63.0
	3	71.4	65.0	62.5	64.7	14.3	61.0
	4	71.4	65.0	62.5	70.6	14.3	63.0
	5	28.6	60.0	59.4	76.5	14.3	60.0
	6	28.6	60.0	59.4	73.5	14.3	59.0
	7	28.6	60.0	68.8	76.5	14.3	63.0
	8	28.6	60.0	62.5	79.4	14.3	62.0
	9	28.6	55.0	62.5	73.5	14.3	59.0
	10	28.6	45.0	59.4	76.5	14.3	57.0

Among the techniques, the pure case-based reasoning has the highest level of accuracy(63%) in the given data sets, followed by multiple discriminant analysis(58.6%) and induction(54%). The results also indicate that classification accuracy is affected by the number of cases retrieved and the depth of the induction tree.

The results of integrated approach using induction and case-based reasoning are summarized in Table 4. First model integrates induction tree which has 5 level of depth and case-based reasoning by applying CBR at the end of induction tree. Second model follows the same procedure except induction tree has 10 level of depth.

Despite the optimistic hope that integration methods can improve effectiveness of case reasoning resulting higher classification accuracy, the experimental results are rather disappointing.

Table 4. Classification accuracy of the integrated approach (%)

		A1	A2	A3	B	C	Avg.
	1	71.4	50.0	59.4	58.8	14.3	55.0

IND (1) + CBR	2	71.4	60.0	62.5	67.7	14.3	61.0
	3	57.1	55.0	59.4	70.6	14.3	59.0
	4	57.1	55.0	68.8	73.5	14.3	63.0
	5	42.9	60.0	65.6	67.7	14.3	60.0
	6	42.9	55.0	65.6	67.7	14.3	59.0
	7	42.9	50.0	68.8	61.8	14.3	57.0
	8	42.9	45.0	71.9	61.8	14.3	57.0
	9	42.9	50.0	75.0	61.8	14.3	59.0
	10	42.9	50.0	71.9	64.7	14.3	59.0
IND (2) + CBR	1	71.4	50.0	50.0	58.8	14.3	52.0
	2	71.4	50.0	62.5	58.8	14.3	56.0
	3	57.1	50.0	68.8	64.7	14.3	59.0
	4	57.1	55.0	62.5	64.7	14.3	58.0
	5	42.9	60.0	68.8	67.7	14.3	61.0
	6	42.9	60.0	68.8	67.7	14.3	61.0
	7	42.9	50.0	62.5	58.8	14.3	54.0
	8	42.9	55.0	65.6	58.8	14.3	56.0
	9	42.9	50.0	62.5	58.8	14.3	54.0
	10	42.9	55.0	65.5	58.8	14.3	56.0

Although Induction(1)+CBR model shows the highest accuracy with three nearest neighbors (63%), this model fails to beat the pure CBR model which has the 63% of accuracy for a given data set. In addition, the performance of the pure CBR model is less sensitive to the number of cases retrieved than integrated model. As can be seen from an inspection of Table 4, there is a fair degree of variation in the results of Induction(1)+CBR depending in the number of cases retrieved.

A comparison of Induction(1)+CBR and induction(2)+CBR indicates that higher level of depth in induction does not guarantee higher performance for integration. This underlines the necessity of optimizing induction tree to apply in a case-based retrieval and not simply leaving it to the induction technique itself to do the job.

CONCLDING REMARKS

This paper examines the potential effectiveness of using induction technique to support case-based reasoning for classification tasks. Although the integration methods to improve effectiveness of case reasoning failed in the experiment, the present exercise suggests some valuable insights to us :

(1) Improving effectiveness in case retrieval using induction technique is not going to be easy, i.g. you are not sure to retrieve the most similar cases because of the static indexing scheme

(2) We should keep in mind the issue of keeping the induction tree optimal. We find keeping such an induction tree optimal can be very expensive. Kolodner [9] defines the optimal network (tree) is one in which retrieval time is kept as small as possible and the cases that are retrieved are those that are best for the job. However, she says she can give no guidelines for choosing a method that will set up a network with these conditions satisfied.

(3) In the positive side, using induction tree and CBR can speed up the retrieval process on big database.

Further research is necessary to determine whether it is possible to improve the effectiveness of indexing in case-based reasoning using induction technique. One extension would involve identifying an optimal structure of induction tree for case-based reasoning.

REFERENCES

[1] Althoff, K., et al. (1993). Integrating Inductive and Case-Based Technologies for Classification and Diagnostic Reasoning. *Proc. ECML-93 Workshop on Integrated Learning Architectures (edited by E. Plaza)*

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

[3] Buta, P. (1994). Mining for Financial Knowledge with CBR. *AI EXPERT February 1994*, pp. 34-41.

[4] Cheeseman, P., et al. (1988). AutoClass: A Bayesian classification system. In *Proceedings of the Fifth International Machine Learning Workshop*. San Mateo, CA: Morgan Kaufmann.

[5] Fisher, D. (1987). Knowledge acquisition via incremental conceptual clustering. In *Machine learning*, vol. 2.

[6] Hansen, J., et al. (1995). Case-Based Reasoning: Application Techniques for Decision Support. *Intelligent Systems in Accounting, Finance and Management Vol. 4*, pp.137-146.

[7] Kolodner, J. (1991). Improving human decision making through case-based decision aiding, *AI Magazine, 12, No 2*, pp. 52-68.

[8] Kolodner, J. (1992). An introduction to case-based reasoning. *Artificial Intelligence Review 6(1)*, pp. 3-34.

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

[10] Manago, M., et al. (1991). Induction of Decision Trees from Complex Structured Data. In G. Piatetsky-Shapiro and W.J. Frawledy, eds., *Knowledge Discovery in Databases*, Combridge, MA:AAAI/MIT Press.

[11] Manago, M., et al. (1993). Induction and Reasoning from Cases. In: Richter, Wess et al, pp.313-318.

[12] Reisbeck, C.K, & Schank, R.C. (1989). Inside Case-Based Reasoning. Lawrence Erlbaum Associates, Hillsdale, NJ, US.

[13] Slade, S. (1991). Case-based reasoning: A research paradigm. *AI Magazine Spring 1991*, pp. 42-55.