International Journal of Fuzzy Logic and Intelligent Systems, vol. 4, no. 2, September 2004 pp. 205-210

Efficient Extraction of Hierarchically Structured Rules Using Rough Sets

Chul-Heui Lee, Seon-Hak Seo

Dept. of Electrical and Computer Engineering, Kangwon National University
192-1 Hyojadong, Chunchon, Kangwondo, Korea
Tel : 82-31-250-6296 Fax : 82-31-241-3775
chlee@kangwon.ac.kr seonhak@hitel.net

Abstract

This paper deals with rule extraction from data using rough set theory. We construct the rule base in a hierarchical granulation structure by applying core as a classification criteria at each level. When more than one core exist, the coverage is used for the selection of an appropriate one among them to increase the classification rate and accuracy. In Addition, a probabilistic approach is suggested so that the partially useful information included in inconsistent data can be contributed to knowledge reduction in order to decrease the effect of the uncertainty or vagueness of data. As a result, the proposed method yields more proper and efficient rule base in compatability and size. The simulation result shows that it gives a good performance in spite of very simple rules and short conditionals.

Key Words : rule extraction, inconsistent data, hierarchical granulation structure, coverage, probabilistic approach, rough set

I. Introduction

Today knowledge discovery which extracts the useful and meaningful knowledge from a large volume of data becomes one of the important issues[1]. Its application covers a wide range of areas such as system modeling and control, time series analysis and forecasting, decision making, and data mining, etc.. Knowledge discovery includes the extraction of rules as a kernel of the problem, and the resultant rules should be explicit and uncomplicated.

The rule generation procedure usually consists of two major operations; first the given data is categorized into several classes by measuring the similarity based on the different attributes of the data, and then their features are extracted and formulated into the rules. However, it is not easy to discover the efficient and effective rules since there exists the uncertainty in the available data due to the repetition of some specific data, the corruption of data, and the inclusion of inconsistent data, etc..

A variety of rule extraction methods, such as induction of decision trees, artificial neural networks, clustering, and rough set theory, have been suggested[1-5]. In general, the more the rules are induced, the more detailed and exact the description of the objects can be. Nevertheless, a method which results in fewer and simpler rules is preferred by virtue of easiness in understanding and explaining of the discovering process, even though its performance is possibly more or less inferior[5]. The induction of decision tree methods expose a defect that they are very sensitive to the data size and may yield too many rules. Neural networks based methods have a critical drawback that they can't explain why the resultant rules are induced. Also their learning speed becomes slow down, and the classification ability is lowered. Although clustering based methods are effective in case of quantitative data, the calculation burden is heavy, so they are not appropriate for vast data. In addition, the interpretation of the results obtained by clustering is not easy.

Rough set theory, introduced by Pawlak, is emerging as a powerful tool for knowledge discovery from the incomplete and inconsistent data. As it provides an efficient and systematic frame through the set theoretic treatment of knowledge that can extract the common essence from the individual fragmentary knowledge about objects, it is widely and successfully used in rule extraction and reduction[6,7].

When rough set theory is applied to rule extraction, the choice of classification criteria and handling of inconsistent data are the very delicate but important problems. The choice of classification criteria has a great influence on the features of induced rules such as size, length, and hierarchy, etc.. If the cores and reducts as the candidates for classification criteria are searched for without consideration of the frequency of inconsistent data belonging to the same equivalent class, necessary attributes may be discarded or unnecessary ones may not be eliminated unfortunately.

Therefore, an efficient rule extraction method that can handle these problems effectively is presented in this paper. The rule base is constructed in a hierarchical granulation structure by applying cores as the classification criteria at each level. When more than one core exist, the coverage degree is used to select an appropriate one among them to increase the classification rate. Furthermore, a probabilistic approach[8] is provided so that the loss of the partially useful information included in inconsistent data by excluding them entirely can be avoided. As a result, the effect of the uncertainty on the knowledge reduction procedure is decreased to some extent.

Manuscript received May. 7, 2004; revised Sep. 2, 2004.

The proposed method produces more proper and effective rule base in compatibility and size.

2. Summary of Rough Set Theory[2]

Rough set theory is an extension of the classical set theory in which the notion of classification is incorporated. Its mathematical basis is the indiscernibility relation which says the objects characterized by the same information are indiscernible in view of the available information about them. Any set X is expressed in terms of approximation regions defined by lower and upper approximation[2] [6-11].

(1) Indiscernibility

Let $S=(U, A)$ be a database, where U is a finite set of objects called the universe, ana A is a set of attributes. For each $R \subseteq A$, the equivalence relation, called indiscernibility relation, I_R is defined as:

$$I_R=\{(x, y) \in U \times V : \forall a \in R,\ a(x)=a(y)\} \quad (1)$$

where $a(x)$ denotes the value of attribute a for element x.

The family of all equivalence classes of I_R will be denoted by U/I_R, or simply U/R, and an equivalence class of I_R, i.e., a block of the partition U/R containing x will be denotes by $R(x)$.

(2) Approximation

Let $X \subseteq U$ a rough concept. The B-lower and B-upper approximation of X are defined as follows:

$$B_*(X)=\bigcup_{x \in U}\{B(x) : B(x) \subseteq X\} \quad (2)$$

$$B^*(X)=\bigcup_{x \in U}\{B(x) : B(x) \cap X \neq \emptyset\} \quad (3)$$

The set $BN_B(X) = B^*(X) - B_*(X)$ will be referred to as the B-boundary region of X. If the B-boundary region of X is empty set, then the set X is crisp with respect to B; otherwise, if $BN_B(X)=\emptyset$, the set X is referred to as rough with respect to B.

(3) Reduct and Core

Let P, Q be families of equivalence relations over U. The P-positive region of Q, denoted by $POS_P(Q)$ is defined as:

$$POS_P(Q)=\bigcup_{x \in U/Q} P_*(X) \quad (4)$$

$POS_P(Q)$ is the set of all elements of U that can be uniquely classified to blocks of the partition U/Q by means of P. $R \in P$ is Q- indispensable in P if $POS_{I_P}(I_Q) \neq POS_{I_{P-R}}(I_Q)$. If every $R \in P$ is Q-indispensable, then P is Q-independent. If R is Q-independent and $POS_{I_R}(I_Q) = POS_{I_P}(I_Q)$, we say that R is a Q-reduct of P, denoted by $RED_Q(P)$. Reduct is not unique in general. P may have many reducts. Also we define Q-core of P be a set of all indispensable relations. Core can be expressed with respect to reduct as follows:

$$CORE_Q(P)=\bigcap RED_Q(P) \quad (5)$$

Reduct is a subset of P that is sufficient and necessary to describe the same knowledge represented by P, and core is the essential part of that knowledge.

3. Extraction of Hierarchical Rules Using Cores

An important step in rule discovery process is reducing dimensionality of data[1]. In rough set theory, this knowledge reduction is carried out by using cores and reducts which can be calculated from the decision table or the discernibility matrix[2, 12]. Decision table is a kind of prescription that describes the relation between the conditions and the decisions. It can be curtailed by eliminating unnecessary attributes and their value by means of cores and reducts. For details, confer to [2].

3.1 Rules with Hierarchical Granulation

Granulation of a universe involves grouping of the similar elements into the clusters of indistinguishable objects, the granules[13, 14]. The granularity of knowledge is due to the indiscernibility of the objects caused by lack of sufficient information about them. In terms of equivalent granules, a finer relation produces smaller granules than a coarser relation. This property may be linked with the hierarchical granulation structure. In a higher level with coarser granulation, one obtains less accurate rough set approximations. On the other hand, in a lower level with finer granulation, one obtains more accurate rough set approximations, but also has more rules. At the lowest level, the number of the rules is equal to that of objects, and each rules includes the conditions corresponding to the number of attributes[15]. Therefore, one may search the layered granulations to find the suitable granulation for the approximation.

In this paper, the rules with hierarchy are induced by taking cores as classification criteria because they are the essential attributes of objects. Hence the initial classification accuracy and rate are increased in spite that the number of the rules is small and the conditionals in the rules are short.

Consider an example of rotary clinker kiln [2] whose decision table is given by Table 1. It consists of 13 objects, 4 conditional attributes $\{a, b, c, d\}$, and 2 decision attributes $\{e, f\}$.

First, applying the same knowledge reduction procedure in [2] without eliminating unnecessary reduct attribute b, cores for each rows in decision table is obtained. Since there is no inconsistent data in this example, a probabilistic processing of them, presented in next section, is not required. The shaded and encircled attributes in Table 1 are cores.

Next a hierarchical granulation structure is obtained by

sequentially applying cores as classification criteria from top to bottom level of granulation. Of course, if there remains no more core, resucts are used for classification. The resultant structure is shown in Fig.1, where core c is used at the highest level.

Table 1. Decision table of rotary clinker kiln

U	a	b	c	d	e	f	D
1	3	3	2	②	2	4	
2	③	2	2	②	2	4	I
3	③	2	2	①	2	4	
4	②	2	2	1	1	4	II
5	②	2	2	2	1	4	
6	3	2	2	③	2	3	
7	3	3	2	③	2	3	III
8	4	3	②	3	2	3	
9	4	3	③	3	2	2	
10	4	4	3	3	2	2	
11	4	4	3	2	2	2	IV
12	4	3	3	2	2	2	
13	4	2	3	2	2	2	

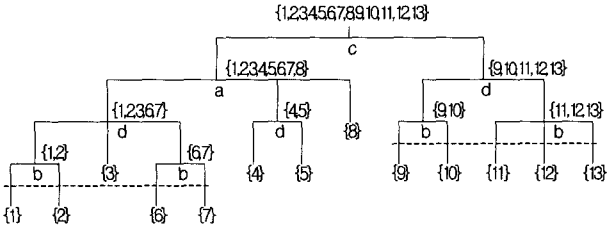

Figure 1. Hierarchical granulation structure with core c

The accuracy measure[16, 17], a measure for the uncertainty of knowledge, is defined by cardinality ratio of lower approximation to upper approximation as follows:

$$\alpha_B(X) = \frac{card(B_*(X))}{card(B^*(X))} \tag{6}$$

The above definition implies that the exactness of knowledge is dependent on the size of the boundary region of a set. That is, the wider the boundary region of a set is, the lower the exactness of knowledge represented by it is.

The accuracy of approximation for decision class $D=$ I is given in Table 2.

Table 2. Accuracy of approximation for D=I

level	lower approximation	upper approximation	accuracy
1	Φ	U	0
2	Φ	{1,2,3,4,5,6,7,8}	0
3	Φ	{1,2,3,6,7}	0
4	{1,2,3}	{1,2,3}	1
5	{1,2,3}	{1,2,3}	1

Granulation level 4 and 5 have the same value of accuracy, 1. However, classification is performed with cores only at level 4 while both cores and reducts are used in classification at level 5. It indicates the fact that, in most cases, classification with hierarchical granulation structure can be carried out with only cores. The necessary condition of classification using cores only is that product of the number of the attribute values of each core is not less than the number of decision classes.

3.2 Selection of cores by coverage

When more than one cores are found, one must select one of them for classification criteria at each level so that the better and more efficient hierarchical granulation structure can be induced. Here we propose an effective selection method using coverage in which core with higher total weighted coverage(TWC) used as classification criteria for upper level.

Coverage[16] of R under D, denoted by $\kappa_R(D)$, is defined as

$$\kappa_R(D) = \frac{card([x]_R \cap D)}{card(D)} \tag{7}$$

Coverage is a probability of R under the condition of D, and it measures the degree of necessity of a proposition $R \rightarrow D$. Therefore, the higher coverage the attributes has, the more useful it is for rule extraction.

Selection procedure of cores is as follows.

First, calculate coverage of cores under the equivalent classes of decision, and then select ones with $\kappa_R(D_i) > \delta$. Next add all coverage of each selected core multiplied by weights given as the cardinality ratio of decision class to universe. That is, calculate TWC defined as

$$\kappa_R'(D) = \sum_i \frac{card(D_i)}{card(U)} \kappa_R(D_i) \tag{8}$$

Core with higher TWC is applied to upper level for classification. As a result, the initial classification accuracy is increased, and the more efficient classification structure is obtained.

Let's apply the proposed method to decision table in Table 1. TWC of attribute a is 12/13, that of c is 1, and that of d is 8/13. If core with lowest value of TWC, d, is used in classification at top level, the resultant classification structure is different from the case using core c, as shown in Fig. 2.

With the structure in Fig. 1, the rule for decision class $D=$ IV is "If $c=3$, Then $D=$ IV". On the other hand, the rule for same decision class induced from the structure in Fig. 2 is "If ($d=3$ & $c=3$) or ($d=2$ & $c=3$), Then $D=$ IV".

From the above, we know that the former produce minimal

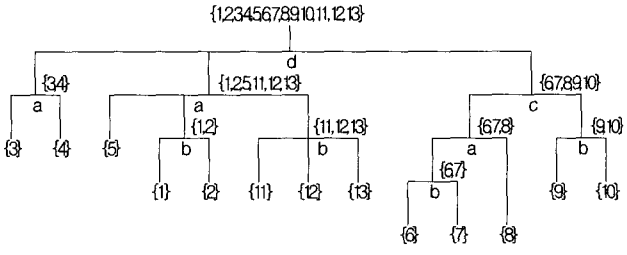

Figure 2. Hierarchical granulation structure with core d

rules, while the latter requires additional calculation as well as yields more complicate rules.

4. Probabilistic Processing of Inconsistent Data

Conventional methods for processing of inconsistent data in the simplification procedure of decision table is to discard them all in order to exclude the effect of uncertainty involved in them. However, the partially useful information contained in inconsistent data is also lost, so the rule classification accuracy is degraded in consequence[8, 16].

Thus we suggest an alternative technique base on probabilistic approach for handling of inconsistent data. Inconsistent data occurred frequently than a certain value is treated as consistent in the proposed method.

The probability in sense of frequency is defined as follows[8].

$$p(r_i) = \frac{supp(r_i)}{supp(r_j)} \tag{9}$$

where $supp(r_i)$ is the number of objects supporting rule(object) r_i, and r_j is a rule having the same conditional attributes as r_i.

If this probability $p(r_i)$ is greater than $\beta(0.5 \le \beta \le 1)$, data corresponding to rule r_i are regarded to be consistent: otherwise, they are discarded. β is the critical value for inclusion of inconsistent data in rule extraction, and functions as a control factor.

As $p(r_i) = 0.5$ means the probability of the existence of uncertainty is equal to that of the existence of useful information , so $\beta = 0.5$ implies that inconsistent data is regarded as consistent one if the existence of useful information is more probable. If $\beta = 1$, inconsistent data is excluded entirely, and only consistent data are used in rule extraction.

The following decision table(Table 3) is an example to show the probabilistic processing of inconsistent data.

Table 3. Decision table with inconsistent data

Object	a_1	a_2	a_3	a_4	a_5	a_6	D
1	1	1	4	3	2	3	1
2	1	1	4	2	2	2	1
3	2	1	4	3	2	3	1
4	2	②	2	2	2	2	1
5	3	3	2	③	④	3	1
6	2	2	3	3	3	3	2
7	2	④	2	2	2	2	2
8, 9	3	3	2	②	4	2	2
10	4	4	1	4	3	4	2
11	4	4	1	④	4	4	2
12	3	3	2	3	③	3	3
13	3	3	2	2	4	2	3
14	4	4	1	③	4	4	3

It consists of 14 objects, 6 conditional attributes $\{a_1, a_2, \cdots, a_6\}$, and 1 decision attribute $\{D\}$. The objects 8, 9, and 13 in Table 3 are inconsistent because their decision attribute is different with each other while their conditional attributes are the same. If we choose $\beta = 0.6$, then the objects 8 and 9 are treated as consistent data because $p(r_{8,9}) = 2/(2+1) = 2/3$ is greater than β. On the other hand, since $p(r_{13}) = 1/3 < \beta$, the object 13 is discarded before rule extraction.

If no probabilistic processing is performed, cores of this decision table are $a_2 =$ ② of the object 4, $a_5 =$ ④ of the object 5, $a_2 =$ ④ of the object 7, $a_4 =$ ④ of the object 11, $a_5 =$ ③ of the object 12, and $a_4 =$ ③ of the object 14. But after probabilistic processing with $\beta = 0.6$, $a_4 =$ ③ of the object 5 and $a_4 =$ ② of the object 7 are added to cores.

This result reveals the fact that cores needed for rule extraction may be eliminated if an appropriate and careful treatment of inconsistent data. As shown above, the probabilistic processing of inconsistent data makes it possible that the partially useful information included in inconsistent data is contributed to knowledge reduction procedure in order to decrease the effect of the uncertainty or vagueness of data.

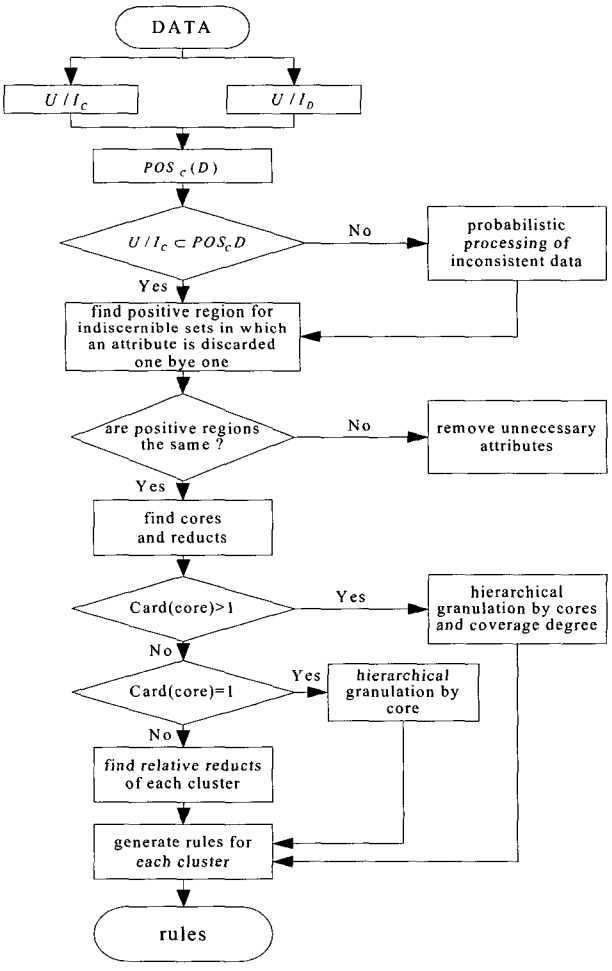

Figure 3. Algorithm of the proposed method

5. Algorithm

The algorithm of overall rule extraction method proposed in this paper is shown in Fig. 3.

First, the probabilistic processing of inconsistent data is performed if necessary. Next find cores. Finally the rules are induced in hierarchical granulation structure by applying cores in the order of their value of total weighted coverage(TWC).

6. Simulations

Here Wisconsin Breast Cancer Database[18] is used for the simulations. This data has 9 conditional attributes, 1 decision attribute that consist of 2 classes(benign, malignant), and 699 cases. We select 369 cases(group I) for learning data to find attribute classification rules, but among them only 342 cases except 27 incomplete data are practically used in the learning stage. The remaining 330 data(group II) are used in the test stage.

The rules created by the proposed method are as follows.

1. if ($a_1 \leq 6$ and $a_2 = 1$) *then* $d =$ *class* 1
2. if ($a_1 \leq 6$ and $a_3 \leq 2$) *then* $d =$ *class* 1
3. if ($a_1 \leq 6$ and $a_3 \geq 3$) *then* $d =$ *class* 2
4. if ($a_1 \geq 7$) *then* $d =$ *class* 2

With respect to these rules, the classification rate of learning data is 97.95%, and that of test data is 95.45%, as given in Table 4.

Table 4. Classification result of WBC data

group	class	original data	classified data	accurate classification	inaccurate classification	accuracy
I	1	174	167	167	0	97.95%
	2	168	175	168	7	
II	1	257	246	244	2	95.5%
	2	73	84	71	13	

Table 5. Classification Accuracy Comparison

classification method	data	the number of rules/accuracy
Neuro-fuzzy	learning data	hidden layer = 10
	test data	100.0%
		95.0%
C4.5	learning data	rules = 16
	test data	96.8%
		95.6%
CN2	learning data	rules = 30
	test data	100.0%
		94.4%
proposed method	learning data	rules = 4
	test data	93.5%
		95.45%

As shown in the simulation results, the classification accuracy of the proposed method is satisfactory even though the number of the classification rules is only 4. The performance comparison with another existing methods[19] is given in Table 5. Compared with another classification methods, the proposed algorithm has the same or higher accuracy. Moreover it takes an advantage that it generates simpler rules and shorter conditionals. By virtue of this, the speed of the classification process becomes faster, and it is easier to understand and explain the results of classification. Also it is more powerful in constructing a classification model since it adopts hierarchical rule structure.

7. Conclusions

We suggest an efficient rule extraction method using rough set. Rough set is used to classify the objects of interest into the similarity classes and to investigate the granularity of knowledge. Also the hierarchical granulation structure is adopted to find classification rules effectively.

The rule base is constructed in a hierarchical granulation structure by applying core as a classification criteria at each level, and coverage is used to select an appropriate core to increase the classification rate and efficiency. In addition, a probabilistic approach is provided so that the loss of information included in inconsistent data is minimized.

As shown in simulation results, the proposed method produces more proper and effective rule base in compatibility and size, and gives a good performance in spite of simple rules and short conditionals. Therefore it may be used in the wide area of applications such as system modeling and control, time series analysis and forecasting, decision making, and data mining, etc..

References

[1] I. Witten and E. Frank, *Data Mining*, Morgan Kaufmann Publisher, 2000.

[2] Z. Pawlak, *Rough Sets : Theoretical Aspects of Reasoning about Data*, Kluwer Academic Publishers, Dordrecht, 1991.

[3] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, Academic Press, 1999.

[4] M. S. Chen, Jong Soo Park and Philip S. Yu, "Efficient Data Mining for Path Traversal Patterns", IEEE Trans. on Knowledge and Data Engineering, vol. 10, no. 2, pp. 209-221, 1998.

[5] A. Berson, S. Smith and K. Thearling, *Building Data Mining Applications for CRM*, McGraw-Hill, 1999.

[6] Z. Pawlak, "Why Rough Sets?", Proc. of the 5th IEEE International Conference on Fuzzy Systems, vol. 2, pp. 738-743, 1996.

[7] C. C. Chan, "A rough set approach to attribute generalization in data mining", Information Sciences, vol. 107, pp. 169-176, 1998.

[8] E. A. Kwon and H. G. Kim, "Reduction of Approximate Rule Based on Probabilistic Rough Sets", Trans. KIPS,

vol. 8-D, no. 3, pp. 203-210, 2001.
[9] Y. Cho, H. Noh, Y. Lee and M. Park, "Fuzzy Modeling by Occupancy Degree and Optimal Partition of Projection Using Rough Set Theory", Trans. KIEE, vol 46, no. 9, pp. 1388-1394, 1997.
[10] Y. Yang and T. C. Chiam, "Rule Discovery Based On Rough Set Theory", Proc. of the Third International Conference on Information Fusion, Vol.1, TuC4-11-16, 2000.
[11] M. B. Gorzalczany and Z. Piasta, "Neuro-fuzzy approach versus rough-set inspired methodology for intelligent decision support", Information Sciences, vol. 120, no. 1, pp. 45-68, 1999.
[12] W. C. Bang and Z. N. Bien, "Determination of Input/Output Relations and Rule Generation for Fuzzy Combustion Control System of Refuse Incinerator Using Rough Set Theory." Proc. of KFIS Fall Conf. '97, pp. 81-86, Seoul, 1997.
[13] Y. Y. Yao, "Rough Sets, Neighborhood Systems, and Granular Computing", Proc. of the IEEE Canadian Conference on Electrical and Computer Engineering, pp. 1553-1558, 1999.
[14] Y.Y. Yao, "Stratified Rough Sets and Granular Computing", Proc. of the 18th International Conference of the North American Fuzzy Information Processing Society, pp. 800-804, 1999.
[15] C. H. Lee, S. H. Seo and S.C. Choi, "Rule Discovery using Hierarchical Classification Structure with Rough Sets", Proc. of Joint 9th IFSA World Congress and 20th NAFIPS International Conference (IFSA/NAFIPS 2001), pp.447-452, 2001
[16] S. Tsumoto, ""Extraction of Experts' Decision Rules from Clinical Databases Using Rough Set Model", Intelligent Data Analysis, vol. 2, pp. 215-227, 1998.
[17] Z. B. Xu and J. Y. Liang, "Inclusion degree: a perspective on measure for rough Set data analysis", Information Sciences, vol. 141, pp. 227-236, 2002.
[18] W. H. Wolberg and O. L. Mangasarian, "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", Proc. of the National Academy of Sciences, vol. 87, pp. 9193-9196, 1990.
[19] P. Clark and T. Niblett, "The CN2 Induction Algorithm", Machine Learning Journal, vol. 3. pp.261-283, 1989.

Chul-Heui Lee

He received the B.S., M.S. and Ph. D. degree in electrical engineering from Seoul National University, Seoul, Korea, in 1983, 1985 and 1989, respectively. Since 1990, he has joined the faculties of the Department of Electrical Engineering(now Department of Electrical and Computer Engineering) in Kangwon National University, Chunchon, Korea, where he is currently a professor. His research interests include intelligent systems and soft computing(fuzzy, neural and genetic algorithms), rough sets and knowledge discovery, and adaptive control and signal processing, etc..

Seon-Hak Seo

He received the B.S. and M.S. degree in electrical engineering from Kangwon National University, Chunchon, Korea, in 1995 and 1997, respectively. He is currently a Ph. D. candidate of Department of Electrical and Computer Engineering in Kangwon National University. He is interested in applying rough sets and soft computing(fuzzy, neural and genetic algorithms) techniques to intelligent control, signal processing, and knowledge discovery.