Finding More Accurate Decision Models Based on Rough Set Theory

Hyontai Sug

Division of Computer and Information Engineering, Dongseo University, Busan, 617-716 South Korea

Summary
For the time when we deal with the difficulty of determining a decision attribute among several candidates to find more accurate decision models, we propose a new method to help users find a good decision attribute based on the overall correctness of the target database. The overall correctness is measured by the degree of non-conflicting decisions based on rough set theory. By assessing the measure among possible candidates, we can have more accurate decision models for the target databases.
Key words:
Decision trees, Rough set theory, clustering

1. Introduction

There are two principal data mining tasks [1]; Classification maps a datum into one of the predefined classes. Clustering is needed if the user does not define classes. But, yet little research on the situation between classification and clustering has been done.

Most KDD (knowledge discovery in databases) systems assume a good training data set has been selected during the data selection process, so we usually assume that a decision attribute is predefined. But, in real world databases we may have some difficulty in determining a decision attribute, if we have limited domain knowledge about the database. Moreover, when the target database is a relational database, usually several relations are joined together for data warehousing, and if several relations are joined together, we may have a similar problem in the case that we want to select a decision attribute from the new database table.

But direct generation of decision models like decision trees for each candidate decision attribute to determine the best one may take very long time especially when the size of target data set is very large and it has many continuous attributes [2]. So, direct testing for each candidate may not be practical. One may think that sampling is an alternative for the problem, but sampling may not be satisfactory, because there is the possibility of sampling error and sample size problem.

Moreover, when we generate a decision model, we may need to find very accurate decision models in some domains. For example, in the thyroid application experimented by Quinlan et. al. [3] 91 ~ 95% of high level accuracy can be obtained by ignoring all attribute values and classifying all cases as normal, thus he could get very simple decision model. But knowledge discovery process may need to find some hidden causes that cover only 5 ~ 9% of the cases. In this respect a decision attribute that will generate not only simpler decision model but also higher accuracy may be more significant than other candidate decision attributes.

We suggest a method to solve the problems of choosing a good decision attribute based on an approach developed from rough set theory to generate more accurate decision model.

We will first discuss related works in section 2, in section 3 we present our method in detail and in section 4 we illustrate our method through experimentation. Finally in section 5, we present conclusions.

2. Related Work

Many classification systems have been implemented including decision tree systems [1], neural networks [4], rough set-based systems [5], etc. The basic assumption in all of these approaches is that we know the class of each example beforehand. For example, decision tree systems have one decision attribute that depends on other values that are under condition attributes.

Although most classification systems assume that a decision attribute is given, this assumption can be a limitation for the applicability of these systems, since it may not be always clear that we know exactly which attribute in a database table is the decision attribute. Due to this fact an elaborate selection process is needed unless the database has been arranged in a very simple structure.

When we don't know exactly the class of each object or row of a database table, we should rely on clustering. Earlier work in this area has been mostly done outside of artificial intelligence under the name of cluster analysis

Manuscript revised August 22, 2005.

using numerical taxonomy and distance measures to define similarity between objects. Much work in pattern recognition is based on this method [6]. Conceptual clustering methods have been developed to take advantage of the object's semantics. CLUSTER/S [7] uses background knowledge to determine the class of an object. But there is no guarantee that the built--in background knowledge can be very helpful for near-optimal clustering, and evaluation using LEF(Lexical Evaluation Function) in CLUSTER/S is prone to be arbitrary and difficult to justify. COBWEB [8] defines classes as a probability distribution over the values of attributes of objects and generates a hierarchy of classes. The system uses a category utility measure which is similar to Gini index [9]. The tree COBWEB generates has the property that all the nodes except the root node define a class. COBWEB generates a taxonomy based on the probability distribution. But this taxonomy may differ from a human's since no external knowledge is used for the classification, so that it must rely on human's verification of the generated tree.

3. Suggesting Method

Rough set theory is a mathematical tool to deal with vagueness and uncertainty of imprecise data. After being introduced by Z. Pawlak in the early 1980's [10], the theory has been developed and expanded to include applications in the fields of decision analysis, data analysis, pattern recognition, machine learning, and knowledge discovery in databases.

If we are given a finite set $\mathbf{U} \neq \phi$ of objects, called a universe, and a family of equivalence relations over **U**, called **R**, then a relational system **K**=(**U**, **R**) is a knowledge base. An equivalence relation represents the set of values that each object can have as one of its properties.

Definition: The degree of dependency between **P** and **Q** where **P**, **Q** $\subseteq$ **R** is defined as follows.

$\mathbf{P} \Rightarrow_k \mathbf{Q}$ where k= (card POS**p**(**Q**))/(card **U**)

1) k = 1: **Q** totally dependent on **P**
2) 0<k<1: **Q** roughly dependent on **P**
3) k = 0: **Q** totally independent on **P**

In the above definition, 'card' and 'POS' stand for cardinality and positive region respectively. Our basic idea is to select a decision attribute that has the largest positive region and the smallest boundary region based on rough set theory.

In order to compute the value based on the size of the positive region repeat the following steps for the each user-selected potential decision attribute.

Repeat

- Select a candidate decision attribute.
- Sort the database table using all attributes of the table other than this decision attribute.
- Find the size of the positive region: count the number of rows having the same value on the condition attributes but having different values on the decision attribute.

Until no more candidate decision attribute;

Note that in the third procedure of the above counting the number of rows is equal to counting the rows belonging to the boundary region. By subtracting the value from the total number of rows, we obtain the number of rows belonging to the positive region.

So, we have the following equation:

The score of candidate decision attribute = |X| / |T|

where X is the number of rows belonging to positive region, and T is the total number of rows of the database table. So, the larger a positive region is, the more overall dependency between condition and decision attributes. Note that attributes in the database tables correspond to equivalence relations in rough set theory.

4. Experimental Consideration

We used random data sets generated from dgp2 data generation program in UCI machine learning repository [11], since the program generates random data in normal distribution with no conflicting classification. Thus, all data are in positive region. We used C4.5 decision tree generation system [12] as a decision model for our experiment.

In order to make conflicting decision values after generating 50,000 objects, duplicate objects were made. Among them randomly 20% and 10% were made to have conflicting classification values. The size of positive region is reduced from 100,000 to 79,644 and 89,876 respectively. Table 1, 2, 3 show decision trees generated from C4.5 for original data set, 10% conflicting data set, and 20% conflicting data set with various pruning confidences. Default value of pruning confidence in C4.5 is 25%.

Table 1: Decisions trees of original data

Tree size	Estimated error rate (%)	Pruning confidence (%)
14458	7.7	25

9752	9.2	15
5069	11.3	5
1461	13.5	0.5
471	15.6	0.01
377	15.8	0.001

Table 2: Decision trees of 10% conflicting data

Tree size	Estimated error rate (%)	Pruning confidence (%)
9875	12.3	25
5586	13.7	15
2451	15.4	5
620	17.2	0.5
392	19.2	0.01
380	19.3	0.001

Table 3: Decision trees of 20% conflicting data

Tree size	Estimated error rate (%)	Pruning confidence (%)
5857	16.7	25
3015	17.9	15
1244	19.4	5
373	20.9	0.5
195	22.9	0.01
195	23.0	0.001

If we compute the scores based on positive regions;

for original data: 100000/100000 = 1,
for 10% conflict: 89876/100000 = 0.89876, and
for 20% conflict: 79644/100000 = 0.79644.

So, we will choose decision attribute of original data as the best decision attribute among three candidate decision attributes. For comparison, table 4 shows the size of decision trees generated from C4.5 when estimated error rate is 15.8%.

Table 4: Decision trees with estimated error rate of 15.8% for each data set

Data sets	The number of nodes
Original	377
10% conflict	2219
20% conflict	7517

5. Conclusions

Although it is difficult to say that there are only two situations of classification and clustering that we surely know or absolutely don't know the class of each object, yet little research on the situation between two cases has been done.

As a method to deal with the situation between clustering and classification, we propose a method based on rough set theory. The relative size of potentially accurate knowledge model is measured by the size of positive region that reflects the overall dependency between condition attributes and a decision attribute. So, by considering the factor in determining a good decision attribute among possible candidates, one may get a better decision model of accuracy.

References

[1] I.H. Witten, E. Frank, Data mining: Practical machine learning tools and techniques with Java implementations, 2nd ed., Morgan Kaufman, 2005.
[2] C.L. Tan, H. Liu, F. Hussain, M. Dash, "Discretization: an enabling techniques," Data Mining and Knowledge Discovery, 6(6): 393-423, 2002.
[3] J.R. Quinlan, P.J. Compton, K.A. Horn, L. Lazarus, "Inductive knowledge acquisitions: a case study," In J.R. Quinlan, editor, Applications of Expert Systems. Addison Wesley, 1987.
[4] M.T. Hagan, H.B. Demuth, M.H. Beale, Neural network design, Martin Hagan, 2001.
[5] S. Hirano, M. Inuiguchi, S. Tsunmoto, Rough set theory and granular computing. Springer Verlag, 2003.
[6] R.O. Duda, P.E. Hart, Pattern classification and scene analysis, John Wiley and Sons, 1993.
[7] R.E. Stepages III, R. Michalski, "Conceptual clustering: Inventing goal-oriented classification of structured objects," In J.G. Carbonell, R.S. Michalski, T.M. Mitchell, editors, Machine learning

volume 2: an artificial intelligence approach, 471-498. Morgan Kaufmann publishers, 1986.
[8] D.H. Fisher, "Interactive optimization and simplification of hierarchical clustering," Journal of artificial intelligence research, 4:147-179, 1996.
[9] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Wadsworth International Group, Inc., 1984.
[10] Z. Pawlak, Rough sets: theoretical aspects of reasoning about data, Kluwer, Netherland, 1991.
[11] S. Hettich, S.D. Bay, The UCI KDD archive, Technical report, University of California, Irvine, Department of Information and Computer Science, 1999.
[12] J.R. Quinlan, C4.5: program for machine learning, Morgan Kaufmann publishers, Inc., 1993.

Hyontai Sug received the B.S. degree in Computer Science and Statistics from Busan National University, South Korea, M.S. degree in Computer Science from Hankuk University of Foreign Studies, South Korea, and Ph.D. degree in Computer and Information Science and Engineering from University of Florida, U.S.A., in 1983, 1986, and 1998 respectively. During 1986-1992, he stayed in Agency for Defense Development (ADD) as a researcher. He is now with Dongseo University, South Korea, as an assistant professor. His research interest includes data mining, machine learning, and database applications.