A RULE DISCOVERY COMPARISON OF STASTICS To INDUCTION: KENDALL'S VS. ID3 IN THE DERMATOLOGY AND LIVER DISORDER DIAGNOSIS DOMAIN

Shu-Chen Kao

Graduate Student
Department of Business Administration
National Cheng Kung University
E-mail: r4888108@ccmail.ncku.edu.tw

Hae-Ching Chang

Department of Business Administration
National Cheng Kung University

Chin-Ho Lin

Department of Industrial Management Science
National Cheng Kung University
E-mail: linn@mail.ncku.edu.tw

Chien-Hsing Wu

Department of Information Management
Kun Shan University of Technology
E-mail: wuch@mail.ksut.edu.tw

ABSTRACT

Recently, Data Mining (DM) gradually becomes an active domain as many mining techniques were developed. ID3, an induction-based method of data mining, applies information theory to get the entropy of the attribute by which the decision tree can be developed. On the other hand, Kendall's correlation is a statistics-based method to help us to find the correlation coefficient of the attributes and the result. According to the coefficient, the decision tree can be generated. The comparison of the above methods is made to find out the way can extracts concise rules from huge datasets, which is seldom discussed in the past. In this research, two different datasets are used in comparing and the criterions are the number of rules and the depth of the generated decision tree. The result shows the ID3 performs better than Kendall's correlation.

Key words: data mining; ID3; Kendall's correlation; entropy; correlation coefficient

1.Introduction

The methodology of knowledge discovery from sets of data requires the understanding and regulation of several complex tasks. Techniques that have been widely utilized in generating rules from the selected dataset mainly are based on induction and statistics[1]. Induction based approaches center on the determination of a root (regarded as an attribute) and the generation of decision tree by obtaining the gained information while statistics decides the relationship via their correlation coefficient, and consequently the order of attributes is determined.

Based on an inductive approach, Chang et al. [2] address the process that combines fuzzy measure theory and ID3 algorithm to enhance the accuracy of the generated decision trees. They also propose a model that employs fuzzy measure theory and ID3 algorithm to help generate rules. Whereas, Tsatsaraskis et al. [3] mention that inductive learning algorithms have been suggested as alternatives to knowledge acquisition for expert systems. However, the application of machine learning algorithms often involves a number of subsidiary tasks to be performed as well as algorithm execution itself. These activities are often called preprocessing and postprocessing. In induction process, data conversion is important when the datasets contains numeric attribute. One of the existing unsupervised techniques used to convert continuous data to discrete data is fuzzy measure. The fact that different numeric value will reflect a different strength of participation becomes a critical drawback for these techniques. This paper discusses issues related to the application of the ID3 algorithm and an integrated approach to the application of ID3.

In additions, statistics-based approach also has been proposed to solve the problems of knowledge acquisition. Arndt et al. [4] reveal a question about which correlation

coefficient to use in a study but are unaware of the strengths and weakness of the alternative correlation measures when using statistical analysis on the selection of attributes. They compare Pearson, Spearman and Kendall's correlation coefficient using a large sample of subjects with schizophrenia spectrum disorders who were evaluated with 7 different psychiatric rating scales. The result suggests that Kendall's tau has many advantages over Pearson's and Spearman's. Moreover, Rudolfer et al. [5] compare and contrast two types of model-logistic regression and decision tree induction for the diagnosis of carpal tunnel syndrome using four ordered classification categories. Result shows that there is no significant difference between the two methods. Further to this investigation, it presents a detailed comparison of the structure of bivariate versions of the models. The result indicates that the classification accuracy of the bivariate model is slightly higher than that of the multivariate ones.

Although these approaches based on induction and statistics were proposed, most of them focused on providing a way to solve the problems in data mining instead of discussing effectiveness. In fact, not only the capability but effectiveness also plays an important role in the process of rule generation. In this research, we compare ID3, a methodology in induction domain [6] with Kendall's tau correlation, a methodology in statistics domain [4] as to their effectiveness in the process of rule generation. For consistency, the continuous data is converted into discrete data with fuzzy membership function while applying ID3 to generate rules. The datasets include a set of real diagnosis records in the domain of dermatology and liver disorder. And the criterion used to evaluate these two methods is the number of rules and the number of levels generated.

This paper is organized as follows. In section 2, the research framework and methodology are described in detail. The results of comparison are expressed in section 3. The conclusion is made and future work is proposed in final section.

2. Research Framework

Although several methods that based on induction, statistics, etc. were proposed in the past, the comparison of effectiveness was always ignored. In this research we take two methods, one based on induction and another based on statistics, to compare their effectiveness of rule generation. First, we choose the methods to be the targets of comparison. One method we choose is ID3, an induction-based technique and another one is Kendall's tau correlation, a statistics-based technique. Before the decision tree is generated, the root has to be decided. The ID3 chooses the attribute which gains maximum entropy as the root. On the other hand, the Kendall's tau correlation chooses the attribute which has maximum correlation coefficient as the root. Because of the restriction of ID3, the trained dataset must be discrete. Consequently, the attributes are supposed to be converted into discrete if they are continuous. After generating the decision tree, we can compare the results get from the above two different ways according to their evaluation criterion: the number of rules and the number of attributes used. The whole process about this research architecture is described as Figure 1.

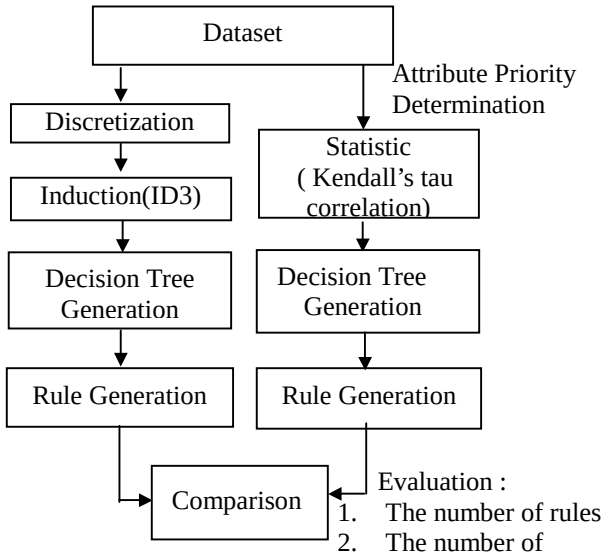

Figure 1: The research architecture

2.1 ID3

Quinlan [6] presents an ID3 (Interactive Dichotomizer 3) algorithm to help generating decision tree as a mechanism of knowledge discovery. The ID3 algorithm is a data-driven approach that uses top-down induction technique to generate its decision tree. ID3 algorithm basically adopts the information theory to determine the nodes of the decision tree. That is, the position of each node is determined by the gained entropy. It resolves the problems of selecting the attributes as the nodes of the decision tree.

2.2 Kendall's tau correlation coefficient

In Kendall's tau correlation, a nonparametric measure of association for ordinal or ranked variables that take ties into account [4]. The sign of the coefficient indicates the direction of the relationship, and its absolute value indicates the strength, with larger absolute values indicating stronger relationships. Possible values range from -1 to 1, but a value of -1 or +1 can only be obtained from square tables.

2.3 Membership function

The data used in ID3 must be discrete so that data conversion is needed when the data is numeric. In this research, the technique of triangle membership functions with three levels is used in converting data. First, the maximum value and the minimum value of the attribute have to been found out. Then, the triangle membership functions are determined according to the level number. Once the membership functions are decided, every value on the function will reflect the probability from zero to one. For example, a membership functions defined for an

attribute “Age” illustrated in Figure 2. The maximum and minimum values from the observed values are 75 and 0. Therefore, each linear equation can be determined via two points. The observed numeric data can be then converted to the corresponding level with the probability of the event happens. Table 1 shows the discrete level and its probability that get from the equation line. If the patient is 30 years old, he will be viewed as “medium” with probability 0.7. After the conversion, all the data will be processed with ID3 algorithm, the root of decision tree will first be selected according to the maximum entropy. All the decending node of the tree will then be determined.

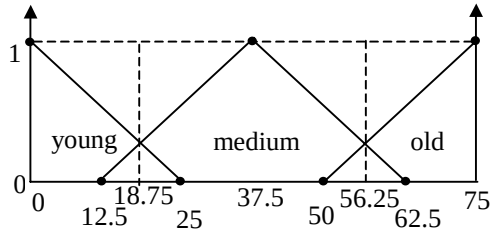

Figure 2: Triangle membership Function for Attribute “Age”

Table 1. Different level and its corresponding probability

Age (x)	Probability
$0 \le x \le 12.5$	y= - 0.04 x +1
$12.5 \le x \le 18.75$	y = max (- 0.04 x + 1 , 0.04 x – 0.5)
$18.75 \le x \le 25$	y = max (- 0.04 x + 1 , 0.04 x – 0.5)
$25 \le x \le 37.5$	y = 0.04 x - 0.5 ,
$37.5 \le x \le 50$	y = - 0.04 x + 2.5
$50 \le x \le 56.25$	y = max (- 0.04 x + 2.5 , 0.04 x – 2)
$56.25 \le x \le 62.5$	y = max (- 0.04 x + 2.5 , 0.04 x – 2)
$62.5 \le x \le 75$	y = 0.04 x - 2

3. Data description and research results

For comparing the performance of ID3 and Kendall’s tau correlation, this research choose two databases which contains continuous data. After data conversion with triangle membership function, ID3 chooses the attribute pertaining the most gain entropy as a root. On the other hand, Kendall’s correlation chooses the attribute having the most significant correlation coefficient as the root. After building the decision tree, the comparison will be made according to the number of rules and the number of levels. The databases description and the result will be illustrated in detailed as follows.

3.1 Data description

The datasets used to test the performance of mining power include Dermatology [7] and Bupa [7]. The dermatology dataset exists 34 attributes : 33 of which are linear valued and the rest is nominal. The differential diagnosis of erythemato-squamous diseases is a real problem in dermatology. It contains 366 instances and 6 classes that entail 6 different kinds of dermatology diseases. Another dataset, Bupa, collected from BUPA Medical Research Ltd. It contains seven attributes and five of them are the data of blood tests. In additions, the rest attributes are the number of half-pint equivalents of alcoholic beverages drunk per day and the classification. This dataset contains 345 instances and 2 classes.

3.2 Attribute priority determination

After calculating information gain and pertaining correlation coefficient, we determine the attribute priority. Table 2 shows the attribute priority from ID3 and Kendall’s tau correlation for the dermatology databases. From the left part of the result, we choose the attribute with the highest priority in ID3 as the root of the decision tree. The attribute with second priority is chosen as the determination of level two in the decision tree, etc. Figure 3 shows part of generated decision tree of dermatology with ID3. On the other hand, the highest priority in Kendall’s tau correlation, shown in the upper right in Table 2, is chosen as the root of the decision tree too. The generated decision tree of dermatology with Kendall’s correlation is illustrated in Figure 4.

We repeat the same way to generate the decision trees for the dataset of Bupa according to the results listed in Table 3. After generating the decision trees, we begin to compare the above different technique according to the criteria – the number of rules and the number of levels.

Table 4 is the result of the comparison. This table indicates that ID3 perform better than Kendall’s tau correlation in discovering rules for the dermatology and bupa database. Especially, in the dataset of dermatology, the number of rules that ID3 generates is much less than that Kendall’s tau correlation dose.

Table 2. Attribute priority of dermatology database

Attribute priority	ID3		Kendall’s tau correlation	
	Attr. name	Gained info.	Attr. name	Correlation coefficient
1	Y	1.2540	V	-0.6670
2	O	1.1808	T	-0.6590
3	M	1.1367	X	-0.5370
4	F	0.9028	J	-0.5340
5	U	0.7829	N	-0.5040
6	P	0.7806	W	-0.4630
7	AB	0.6717	O	0.4560
8	AG	0.6494	B	-0.4050
9	T	0.6396	I	-0.3990
10	V	0.6021	S	-0.3880

11	L	0.5767	U	-0.3830
12	AC	0.5751	Z	-0.3700
13	AA	0.5686	G	0.3600
14	H	0.4713	AE	0.3490
15	I	0.4503	AD	0.3420
16	C	0.3940	C	-0.3260
17	J	0.3429	P	0.2650
18	D	0.3134	AB	0.2270
19	B	0.3021	A	-0.2084
20	E	0.3011	K	-0.1700
21	S	0.2949	AH	-0.1630
22	G	0.2644	Y	0.1230
23	AE	0.2582	AC	0.1190
24	X	0.2546	L	0.1160
25	N	0.2448	AG	0.1160
26	AD	0.2144	F	0.1150
27	AH	0.2043	H	0.1130
28	Z	0.1995	AA	0.1100
29	K	0.1684	Q	0.0990
30	W	0.1486	D	0.0700
31	Q	0.1447	R	-0.0680
32	A	0.1420	E	-0.0640
33	AF	0.0957	AF	-0.0370
34	R	0.0876	M	-0.0130

Table 3. Attribute priority of Bupa database

Attribute priority	ID3		Kendall's tau correlation	
	Attr. name	Gained info.	Attr. name	Correlation coefficient
1	C	0.0230	E	0.181
2	D	0.0220	D	0.121
3	A	0.0171	C	-0.111
4	F	0.0051	B	-0.101
5	B	0.0003	A	-0.087
6	E	0.0002	F	0.034

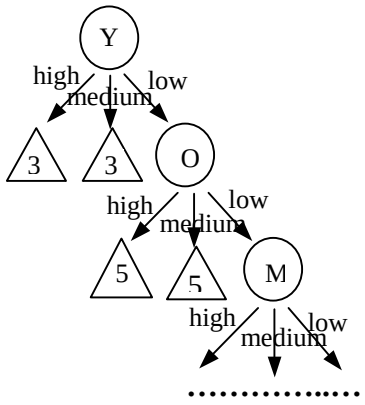

Figure 3: Decision tree generated in ID3 for dermatology

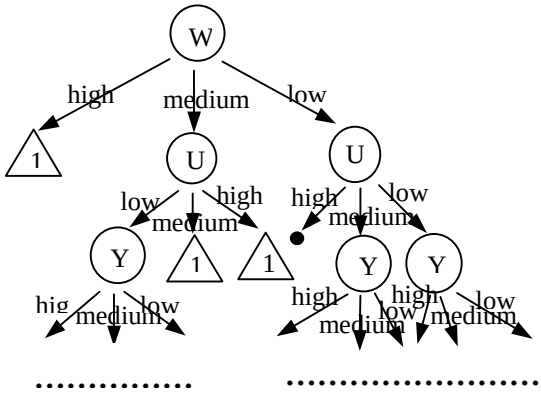

Figure 4: Decision tree generated in Kendall's correlation for dermatology

Table 4: The evaluation results

Database name\ Evaluation criteria	Dermatology		Bupa	
	ID3	Kendall's	ID3	Kendall's
The number of rules	54	96	16	20
The number of levels	25	28	5	6

4. Conclusion and future work

Recently, many researches have proposed different methodologies of knowledge discovering which based on either induction or statistics. But the comparison of the method is seldom discussed. In this research, we choose ID3, induction-based method, and Kendall's tau correlation, statistics-based method, to compare their performance of rule generation. To prove the confidence, we choose the real databases, dermatology and Bupa, as the comparison targets. The evaluated criterions in this research are the number of rules and the level of decision tree. It is obvious that ID3 performs better than Kendall's tau correlation in respect to the selected criterions. In addition to the above criterions, more criteria such as run time, accuracy and the number of nodes seems to be taken into consideration. In this research, run time is not measured due to the lack of computerization and systemization. Accuracy is not considered for ID3 because of the entire checking for the dataset while building the decision tree. The research, therefore, concentrates on the concision that is used to reflect the number of nodes in the tree.

In this paper we use two different databases in this research and conduct the comparison. However, it is necessary to test several different domain datasets separately for the reliability in the future. On the other hand, the annual rule discovery from a set of data is a time consuming task because of the high computation process. It will be helpful to develop a system that can carry out the entire processes from reading dataset, discretization, determination of the attributes, and decision tree generation.

Reference

[1] Ming-Syan Chen, Jiawei Han, and Philip S. Yu, “Data mining : an overview from a database perspective”, IEEE Transactions on knowledge and data engineering, Vol. 8, No.6, December 1996

[2] Kang Chang, Chien-Hsing Wu, “Decision tree generation using fuzzy measure theory”, Journal of the Chinese Institute of Industrial Engineers, Vol.14, No.2, pp115-122, 1997

[3] Charalambos Tsatsaraskis, D. Sleeman, “Supporting preprocessing and postprocessing for machine learning algorithms: a workbench for ID3”, Knowledge acquisition, Vol. 5, No. 4, December 1993

[4] Stephan Arndt, Carolyn Turvey, and Nancy C. Andreasen, “Correlating and predicting psychiatric symptom ratings : Spearman’s r versus Kendall’s tau correlation “, Journal of Psychiatric Research 33, pp97-104, 1999

[5] Stephan M. Rudolfer, Georgios Paliouras, and lan S. Peers, “A comparison of logistic regression to decision tree induction in the diagnosis of carpal tunnel syndrome”, Computers and biomedical research, Vol.32, No. 5, October 1999

[6] Quinlan, J. R., “Induction of Decision Tree”, Readings in Machine Learning, Shavlik, J. W. and Dietterich, T. G., eds., Morgan Kaufmann Publishers, Inc., California, pp57-69, 1986

[7] http://www.ics.uci.edu/~mlearn