Rule Induction using Screening Strategy

Shusaku Tsumoto
Department of Medical Informatics, Shimane Medical University
89-1 Enya-cho, Izumo, Shimane 693-8501 Japan
Phone: +81-853-20-2172, Fax: +81-853-20-2170
email:tsumoto@computer.org

ABSTRACT: Conventional rule induction methods use confidential ratio(classfication accuracy) and support ratio (coverage) to extract strong rules. However, these ratios lost information about prevalence in original databases. In this paper, a new rule induction method with screening strategy is introduced. This method induces a special rule, the confidence ratio of which is equal to its support ratio. This constraint is equivalent to minmax strategy, and also useful to restrict the search space for rule generation. These characteristics are discussed in the paper. Furthermore, the method was evaluated by several medical databases, which suggests that induced rules perform better than rules obtained by conventional methods.

KEYWORDS: rough sets, knowledge discovery, data mining, screening strategy

INTRODUCTION

Conventional rule induction methods(Quinlan, 1996; Langley, 1996) use confidential ratio(classification accuracy) and support ratio(coverage) to extract strong rules. These indices can be viewed as measures of relations between a decision attribute *D* and a conjunctive formula *R*. Let us consider the following two-way table (Table 1):

Table 1: Two-way Table

	D	**~D**	
R	a	b	a+b
~R	c	d	c+d
	a+c	b+d	N=a+b+c+d

From this table, classification accuracy is defined as a/(a+b) and coverage is defined as a/(a+c). It is easy to see from the table that both of the indices focus on the intersection of *R* and *D (a)* with respect to *R(a+b)* and *D(a+c)*. If this table is regarded as a contigency table, another measure, such as χ^2-statistics, is easy to define as a function of *a,b,c,d*: $(a+b+c+d)(ad-bc)^2/ (a+b)(b+c)(c+d)(d+a)$.

One of the most important problems of these measures is that they, as the ratios, lost information about prevalence in original databases. In this paper, a new rule induction method with screening strategy is introduced. This method induces a special rule, the confidence ratio of which is equal to its support ratio. This constraint is equivalent to min-max strategy, and also useful to restrict the search space for rule generation. These characteristics are discussed in the paper. Furthermore, the method was evaluated by several medical databases, which suggests that induced rules perform better than rules obtained by conventional methods.

The paper is organized as follows: Section 2 discusses the definition of probabilistic rules based on rough set model. Section 3 shows rules based on screening strategy. Section 4 presents an algorithm for induction of screening rules. Section 5 gives experimental results. Finally, Section 6 concludes this paper.

DEFINITION OF PROBABILISTIC RULES

ACCURACY AND COVERAGE

In the subsequent sections, we adopt the following notations, which is introduced in (Skowron and Busse, 1994). Let *U* denote a nonempty, finite set called the universe and *A* denote a nonempty, finite set of attributes, i.e., $a: U \to V_a$ for $a \in A$, where V_a is called the domain of *a*, respectively. Then, a decision table is defined as an information system, $\mathbf{A}=(U, A \cup \{d\})$. The atomic formulas over $B \subset A \cup \{d\}$ and *V* are expressions of the form $[a=v]$, called descriptors over B, where $a \in B$ and $v \in V_a$. The set $F(B,V)$ of formulas over *B* is the least set containing all atomic formulas over B and closed with respect to disjunction, conjunction and negation.

For each $f \in F(B,V)$, f_A denote the meaning of $f \in A$, i.e., the set of all objects in *U* with property f, defined inductively as follows.

1. If *f* is of the form $[a=v]$, then, $f_A=\{s \in U \mid a(s)=v\}$.
2. $(f \wedge g)_A = f_A \cap g_A$; $(f \vee g)_A = f_A \cup g_A$; $\neg f_A = U - f_A$

By the use of this framework, classification accuracy and coverage, or true positive rate is defined as follows(Tsumoto, 1998).

(Definition 1) [Accuracy and Coverage]
Let *R* and *D* denote a formula in $F(B,V)$ and a set of objects which belong to a decision *d*. Classification accuracy and coverage(true positive rate) for $R \to d$ is defined as:

$$a_R(D) = \frac{|[x]_R \cap D|}{|[x]_R|} (= P(D \mid R)), \quad k_R(D) = \frac{|[x]_R \cap D|}{|D|} (= P(R \mid D)),$$

where R_A denotes the meaning of a formula *R*. ÿ

In the case of Table 1, since $|R_A|$ is equal to $a+b$ and $|R_A \cap D|$ is equal to *a*, classification accuracy $a_R(D)$ is equal to a/(a+b). In the same way, coverage $k_R(D)$ is equal to a/(a+c).

It is notable that $a_R(D)$ measures the degree of the sufficiency of a proposition, $R \to D$, and that $k_R(D)$ measures the degree of its necessity. For example, if $a_R(D)$ is equal to 1.0, then $R \to D$ is true. On the other hand, if $k_R(D)$ is equal to 1.0, then $R \to D$ is true. Thus, if both measures are 1.0, then $R \Leftrightarrow D$.

Also, Pawlak(1998) recently reports a Bayesian relation between accuracy and coverage:

$$a_R(D)P(D) = P(R|D)P(D) = P(R,D) = P(R)P(D|R) = k_R(D)P(R)$$

This relation also suggests that *a priori* and *a posteriori* probabilities should be easily and automatically calculated from database.

By the use of these two measures, a probabilistic rule is defined as:

$$R \xrightarrow{a,k} d, \; s.t. \; R = \vee_{i,j}[a_i = v_j], \quad a_{[a_i=v_j]} \geq d_a, \; k_{[a_i=v_j]} \geq d_k.$$

This rule is a kind of probabilistic proposition with two statistical measures, which is an extension of Ziarko's variable precision model(VPRS)(Ziarko, 1993). This probabilistic rule is also a kind of rough modus ponens introduced by Pawlak(1998).

PROBLEMS OF PROBABILISTIC RULES BASED ON ACCURACY AND COVERAGE

One of the most important problems of accuracy and coverage is that they lost information about prevalence of *R* and *D* in the universe *U* because both measures are ratios of several counting measure in two-way table. For the above

example in Table 1, accuracy and coverage will lost information about $a+c$ and $a+b$. Even if both of the values of $a+c$ and $a+b$ are very small , those values may be high. Although those information may contribute to discovery of unexpected patterns in a database, most of the rules of high accuracy and coverage have only typical rules, which are trivial for domain experts. Moreover, this ratio representation will have a problem when attributes have many missing values.

SCREENING STRATEGY

As discussed in Section 2, the main characteristic of problems with accuracy and coverage is that they do not include any information about prevalence of R and D because the calculus of a ratio loses constraints between the nominator and denominator. In order to solve the problem with two indices as ratios, we introduce one constraints between accuracy and coverage. Construction of discriminant rules is called *screening strategy*, which is defined as follows.

(Definition 2) [Screening Strategy]
Let R and D denote a formula in $F(B,V)$ and a set of objects which belong to a decision d. Screening strategy is defined as search for rules, the cardinality of the meaning of conditional part $|R_A|$ is equal to that of decision part $|D|$. In other words, searching strategy is to find rules such that:

$$R \xrightarrow{a,k} d,\ s.t.\ R = \vee_{i,j}[a_i = v_j], \quad a_R = k_R \geq d_k\ (|R_A| = |d_A| = |D|).$$

ÿ

In the case of Table 1, since $|R_A|$ is equal to $a+b$ and $|D|$ is equal to $a+c$, $|R_A| = |D|$ means $b=c$. This also means that classification accuracy is equal to coverage. Thus, this constraint is to restrict the freedom of b and c. In other words, $|R_A \cap \neg D|$ is equal to $|\neg R_A \cap D|$.

This constraint is also related with misclassification cost between R and d. If the discrimination between these two classes is very important, then b and c are terms with misclassification. Even a misclassification cost for each case is different, such as l or r, total misclassification cost is equal to $lb+rc=(l+r)b$. Since high accuracy and coverage are equivalent to low misclassification rate, that is low b, to find a rule with the above condition is to find a rule with low misclassification cost. It is easy to see that this strategy is equivalent to min-max strategy(Hand, 1997).

A RULE INDUCTION ALGORITHM

A rule induction algorithm for screening rules is given in Fig.1.

```
procedure Induction of Screening Rules;
var i: integer, M,Li:List;
begin
   L1:= Lir; /* Candidates for Inclusive Rules *
   i:=1; M:={};
   while (i=1 or M ¹{} ) do
     for i:=1 to n do /* n: Total number of attributes */
      begin
       while (Li ¹{}) do
         begin
           Select one formula R from Li ;
           Li := Li - {R};
           if (a[ai=vjj](D)=k[ai=vjj](D)>dk)}  then do Sir(D):= Sir +{R} ;     /* Include R as Classification Rule */
           else  M := M + {R} ;
         end
       Li+1:=(A list of all possible conjunctions of two element subsets from M);
      end
end {Induction of Screening Rules};
```

Figure 1: A Rule Induction Algorithm for Screening Rules

EXPERIMENTAL RESULTS

This algorithm was evaluated on a medical dataset on differiential diagnosis of headache, RHINOS domain, whose training samples consist of 2301 samples, 10 classes, and 20 attributes. In these experiments, d_a, is set to 0.75.

The experiments were performed by the following four procedures. First, these samples were randomly split into half (new training samples) and half (new test samples). For example, 2301 samples were split into 1150 training samples and 1151 training samples. Secondly, this algorith, PRIMEROSE-REX(Tsumoto, 1998) and C4.5(Quinlan, 1993) were applied to the new training samples. Finally, the induced results were tested by the new test samples. These procedures were repeated for 100 times and average all the estimators over 100 trials. Three methods were compared with respect to accuracy, length of induced rules, similarity and number of rules. Experimental results on the performance of this system are summarized in Table 2. Expert's rules are acquired manually from medical experts and evaluated by the test samples generated by the above sampling scheme. For a similarity measure, Jaccard's coefficient was adopted because of its simplicity(Everitt, 1996).

Table 2: Experimental Results(Headache, Averaged)

Method	Accuracy	Length	Similarity	Number of Rules
This Algorithm	86.2%	2.3	0.57	162
PRIMEROSE-REX	89.6%	3.9	0.71	2217
C4.5	85.8%	2.7	0.51	218
Experts	93.0%	6.4	1.00	53

The above table show the following results: (1) The performance of this algorithm is a slightly worse than PRIMEROSE-REX, but comparable to C4.5. (2) The length of rules induced by the introduced algorithm is shorten than other methods. (3) Similarity of the rules is not better than PRIMEROSE-REX and C4.5. (4) However, the number of induced rules is suppressed, compared with other methods. These results suggest that rule induction based on screen strategy generates compact rules faster than other methods.

CONCLUSIONS

In this paper, a new rule induction method with screening strategy is introduced. Screening strategy searches for a rule whose accuracy is equal to coverage, which is equivalent to min-max strategy. Experimental results show that induced rules perform slightly better than those generated by conventional methods and that this strategy is useful to restrict the search space for rule generation.

REFERENCES

Everitt, B. S., 1996, "Cluster Analysis", 3rd Edition, John Wiley & Son, London.
Hand, D. 1997, "Assessment of Classification Rules", John Wiley & Son, London.
Langley, P., 1996, "Elements of Machine Learning", Morgan Kaufmann, CA, 1996.
Lin, T.Y. , 1998, "Fuzzy Partitions: Rough Set Theory", Proceedings of Seventh International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems(IPMU'98), Paris, France, pp. 1167-1174, 1998.
Pawlak, Z., 1991, "Rough Sets", Kluwer Academic Publishers, Dordrecht, 1991.
Pawlak, Z., 1997, "Conflict analysis.", Proceedings of the Fifth European Congress on Intelligent Techniques and Soft Computing (EUFIT'97)}, pp.1589--1591, Aachen, Germany, 1997.
Pawlak, Z. 1998, "Rough Modus Ponens", Proceedings of Seventh International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems(IPMU'98), Paris, France, 1998.
Pawlak, Z. 1998, Rough Sets and Decision Analysis, Fifth IIASA workshop on Decision Analysis and Support, Laxenburg, Austria, 1998.
Polkowski, L. and Skowron, A., 1996, "Rough mereology: a new paradigm for approximate reasoning", Intern. J. Approx. Reasoning 15, 333-365, 1996.
Quinlan, J.R., "C4.5 - Programs for Machine Learning", Morgan Kaufmann, Palo Alto, 1993.

Skowron, A. and Grzymala-Busse, J. 1994, From rough set theory to evidence theory. In: Yager, R., Fedrizzi, M. and Kacprzyk, J.(eds.) “Advances in the Dempster-Shafer Theory of Evidence”, pp.193-236, John Wiley & Sons, New York.
Tsumoto, S. , 1998, Automated Induction of Medical Expert System Rules from Clinical Databases based on Rough Set Theory. Information Sciences 112, 67-84, 1998.
Tsumoto, S., 1998. Extraction of Experts' Decision Rules from Clinical Databases using Rough Set Model, Journal of Intelligent Data Analysis, 2(3), 1998.
Zadeh, L.A., 1997, Toward a theory of fuzzy information granulation and its certainty in human reasoning and fuzzy logic. Fuzzy Sets and Systems, 90, 111-127.
Ziarko, W., 1993, Variable Precision Rough Set Model. Journal of Computer and System Sciences, 46, 39-59.