Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, 18-21 August 2005

RULE INDUCTION FOR INCOMPLETE INFORMATION SYSTEMS

HONG-ZHEN ZHENG[1], DIAN-HUI CHU[2], DE-CHEN ZHAN[3]

[1,2]Department of Computer Science & Technology, Harbin Institute of Technology at Weihai, China, 264209
[1]Department of Computer Science & Technology, Harbin Institute of Technology, Harbin, China, 10060
E-MAIL:hithongzhen@163.com, dechen@hit.edu.cn

Abstract:

we proposed a midified rule generation algorithm(MRG) to generate a minimal set of rule reducts and proposed a generalized rule generation algorithm(MRGI) to generate a minimal set of rule directly from the original incomplete information system Based on MRGI, with each rule reduct represents a unique decision rule. We developed a rule generation and rule induction prototype(RGRIPI) to extract certain rules directly from the incomplete information system. RGRIPI can automatically generate a minimal set of decision rules directly from an incomplete data set. We build a probability function combining the plausibility and probability of missing values to compute the possible rules for incomplete information systems.

Keywords:

Rough sets; Knowledge discovery; Reduce

1. Introduction

It is noted that for the typically huge size of today's information systems, real world data tend to be incomplete due to missing values. Therefore, discovering knowledge from incomplete information systems has received more and more attention in recent years.

Certain rules induced directly from incomplete information system may not provide enough knowledge for enterprises or decision makers to predict uncertain situations or provide strategies. Hence, generating possible rules by probability analysis can improve the expediency of rule extraction approach for incomplete information systems.

In real world applications, certain rules induced directly from incomplete information systems may not provied enough knowledge for enterprises or decision makers to predict uncertain situations or provide strategies. Hence, genetating possible rules by probablity analysis can improve the expediency of rule extraction approach for incomplete information systems.

Several techniques have been developed to extract decision rules from an incomplete information system. A key factor among them is using different methods to manage the missing data. The simplest is removing the objects with unknown values[1]. Other simple methods include replacing missing values with possible values calculated by statistical analysis[2] or observed from the feature values indicated in the corresponding input feature[3]. These methods all try to transform the incomplete system into a complete system by smoothing or extending the data.

However, due to the probability of missing values, rules induced by these methods are all incorporated with "uncertainty", i.e. we cannot tell what and how many rules are "certain rules" even though we know they may exist.

The rough set theory provides a natural method to cope with incomplete or inconsistent information, which has been the main impediment to the classification and rule induction of objects. In view of rough set theory, other groups of techniques deal with incompletes systems without changing the size of the data sets or making assumptions about the missing valuses[4,5].

These methods intend to induce every certain rules directly from the original data sets. No matter what the missing values might be, they won't affect the induction rules.

Rough-set-based rule induction approaches which induce decision rules using approximation of a set become inefficient when dealing with large scale databases. Hence we proposed a midified rule generation algorithm(MRG) to generate a minimal set of rule reducts.Developing a similar rule induction algorithm that can induce certain rules effectively and directly from incomplete systems. We discussed rule induction from complete information system including inconsistent data and describes similar rule extraction from incomplete information systems.

2. Rough sets

Most rough sets-based rule induction techniques employ two basic concepts:lower and upper approximations of a set. A lower approximation means the elements that

certainly belong to the set, and the upper approximaton denotes the elements possibly belong to the set.

Let A be a set of condition attributes used to desvribe objects in U and R be a sub sets of A that we wish to investigate as a possible identifier or classifier of X. R can be used to define "indifference"or "indiscernible"classes of U,denoted *Ind(R)*, in the sense that elements of U x and $y \in$ *Ind(R)* if and only if x and y are indiscernible by the valuse of condition attributes R.Finally we donote elementary set with respect to R containing object x in U as $[x]_{IND(R)}$ or *R(x)*.

We define a lower approximation LX_R and upper approximation UX_R of the set X with respect to the set of condition attributes R as follows:

$$LX_R=\{x_i \in U \mid [x_i]_{ind(R)} \subset X\}$$

$$UX_R=\{x_i \in U \mid [x_i]_{ind(R)} \cap X \neq \phi\}$$

The difference between LX_R and UX_R is called the boundary BX_R of X in U with respect to R.That is $BX_R=UX_R-LX_R$

3. Rule Induction for Complete Information System

3.1. An Reduct generation algorithm(RG)

A reduct generation algorithm based on the developments in Pawlak[6] will be present next.The RG algorithm is as follows:

Step1 Initialize object number i=1, feature number j=1;

Step2 Seletc feature $j=1\sim n$, for all $k \neq i$, if $a_{ij} \neq a_{kj}$ or $a_{ij}=a_{kj} \wedge d_i=d_k$ then a_{ij} can generate r-reducts. If all found, go to *Step3;*

Step3 Set $i=i+1$. If all objects have been considered, go to *Step 4*; Otherwise, go to *Step2*;

Step4 Select two feature and go to *Step2,* until all *n-1* features r-reducts have been considered.

RG algorithm, which produces a very efficient scheme for generating all possible reducts with minimal set of attributes. However, the proposed RG algorithm although can generate all possible rule-reducts, in general will not generate a minimal set of rule reducts. Some generated rule-reducts are rule- overlapped and cannot uniquely represent decision rules..

3.2. Modified rule generation algorithm

The main idea of the RG algorithm is based on what Pawlak [6] calls "simplitication fo decision tables". It utilizes relation between objects in indiscernibility sets to produce reducts.

We define a redundant r-reduct as one which produces exactly the same decision rule as same other r-reducts which have a smaller set attributes.

Lemma 1 Each one-feature r-reduct is in a minimal form. That is,there is no redundant one-feature r-reduct.

It is clear from *Lemma 1* that redundant reducts will only involve reducts with two or more features. Clearly "redundant" higher- order reducts can always be constructed from lower-order reducts.

Proposition 1 Non-redundant higher-order reducts can be constructed if those and only those combinations,each one forming a lower-order reduct, are removed from further consideration in forming higher-order reducts[7].

Now we give modified MRG algorithm as follows:

*Step1*Arrange input data according to the value of decision attribute.

Step2 Initialize object number i=1, the number of attribute in reduct r=1.

Step3 Scan row i from column j=1. If $a_{ij} \neq$ "*", go to *step 4*, otherwise go to *step 5*.

Step4 For all $k \neq i$, if $a_{ij} \neq a_{kj}$ or $a_{ij}=a_{kj} \wedge d_i=d_k$, then aij can produce r-reduct. It all columns $j=1,...,n$ have been scanned, go to *step5*. Otherwise, go back to *step 3* to scan the next column $j=j+1$.

Step5 Set $j=i+1$, go to *step 3*, until all objects have been considered. Go to *step 6* after all objects have been considered.

Step6 Based on the objects which have the same corresponding feature value, revise the decision table T by replacing the value of $a_{ij} \neq$ "x" used to form the corresponding 1-attribute reduct by "*". Go to *step 7*.

Step7 Based on the revised decision tabel T,begian indentifying higher-order reducts by setting $r=r+1$. If $r=m$, stop.otherwise set i=1 and go to *step 8*.

Stept8 Scan row i to identify r eligible attribute $Fj1,...Fjr$ along with $aij1,...,aijr$ to form an r-attribute reduct. If such an eligible set{ $aij1,...,aijr$ } exists, go to *step9*. otherwise goto *step 10.*

Step9 For all $k \neq i$, if $a_{ij} \neq a_{kj}$ for at least one $j=j_1,...j_r$ or $a_{ij}=a_{kj}$ for j= $j_1,..j_r \wedge d_i=d_k$, then{ $aij1,...,aijr$ } forms an r-attribute reduct. Based on the objects which have the same corresponding eligible set, each $aij1,...,aijr$ is then marked by the same symble such as "*r" to identify that the combination { $aij1,...,aijr$ } has been used to form an r-reduct and the combination shoud not be used to form any part of any further reduct. Return to *step8*.

Step 10 Set i=i+1. if i>* of objects in U, go to *step 7*.otherwise go to *step 7*.

The key different between the RG and MRG algorith is that MRG contains two-level mechanisms. The MRG algorithm can tell us when to stop due to no higher order

reducts are existed. It can help the decision-maker predict the outcomes of new cases effectively.

4. Modified rule generation algorithm for incomplete systems-MRGI

The MRG algorithm proposed can generate all possible ruleducts in complete information systems including inconsistent data.However, it can not deal with incomplete information systems.

We develop a algorithm to induce rules directly from the original information systems, which means: extract certain rules without changing the size of initial information systems.

Based on the MRG algorithm generating the minimal set of r-reducts, a generalized ruel-generation algorithm for incomplete information systems(MRGI)is developed as follows:

Step 1 :Arrange input data according to the value of decision feature.

Step2: Initialize object number i=1,the number of features in reduct r=1.

Step3: Scan row i from column j=1. If $a_{ij} \neq$ "*" and $a_{ij} \neq$ M, go to *step 4*, otherwise go to *step5*.

Step4: For all $k \neq i$, if ($a_{ij} \neq a_{kj} \neq$ M) or ($a_{ij} \neq a_{kj} \wedge d_j=d_k$), and if a_{kj}=M $\wedge$ $d_j=d_k$, then a_{ij} can produce R-reduct. If all columns $j=1,\ldots,n$ have been scanned, go to *step5*. otherwise,go back to s*tep 2* to scan the next colunm$j=j+1$.

Step5: Set i=i+1, go to *step3*,until all objects have been considered. Go to *step 6* after all objects have been consicered.

Step6: Revise the decision table T by replacing the value of $a_{ij} \neq$ 'x' used to form the corresponding 1-feature reduct by '*'. Go to *step 7*.

Step7: Based on the trvised decision table *T*,begin identifying higher-order reducts by setting $r=r+1$. if $r=m$,stop. Otherwise set $i=1$ and go to *step8*.

Step8: Scan row i to identify r eligible features $F_{ji},\ldots,F_{jr}$ along with $a_{ij1},\ldots,a_{ijr}$ to form an r-feature reduct.(The set $\{ a_{ij1},\ldots,a_{ijr} \}$ is eligible to form an r-feature reduct if none of its proper subjects form a lower-order reduct or the entire combination is not marked as having been used to form an r-reduct already.see *step 9*). If such an eligibel set $\{ a_{ij1},\ldots,a_{ijr} \}$ exists, go to step 9. otherwise go to *step 10.*

Step9: For all $k \neq i$, if $a_{ij} \neq a_{kj} \neq M$ or $a_{ij}=a_{kj} \neq M \wedge d_j=d_k$, for at least one $j=j_1,\ldots j_r$ or $a_{kj} \supseteq M$ & $a_{ijp=} a_{kjp}$ for $j_p=(j_1,\ldots j_r)$-$j_m \wedge d_j=d_k$, then$\{a_{ij1},\ldots,a_{ijr}\}$ forms an r-feature reduct. Each $a_{ij1},\ldots,a_{ijr}$ is then marked by the same symbol such as "*r*" to identify that the combination $\{a_{ij1},\ldots,a_{ijr}\}$ has been used to form an r-reduct and it should not be used to form any part of any furtuer reduct. Return to *step 8.*

Step10: Set $i=i+1$. If $i>$# of objects in U, go to *step 7*. otherwise go to *step 8*.

The above MRGI algorithm generates a minimal set of rule reduct directly from Information systems.

As we can see that the MRGI is modified formMRG on step4 and step9 by considering the missing values to check if the possible value of the missing entry will conflict the generation of the r-reduct. In another words, when considering a_{ij}, if the missing entry has the possibility of conflicting the identification of decision feature, then a_{ij} can not generate a "certain rule".

Like MRG algorithm, the MRGI algorithm not only generates the minimal set of rule reducts but also revises the input decision table*T'*. We build a probability function combining the plausibility and probability of missing values to compute the possible rules for incomplete information systems.

5 . Ruel induction prototype for incomplete system

Based on MRGI, it is easy to realize that the minimal set of rule reducts represent "certain" rules. We developed a rule generation and rule induction prototype(RGRIPI) which can automatically generate a minimal set of decision rules directly from an incomplete data set.

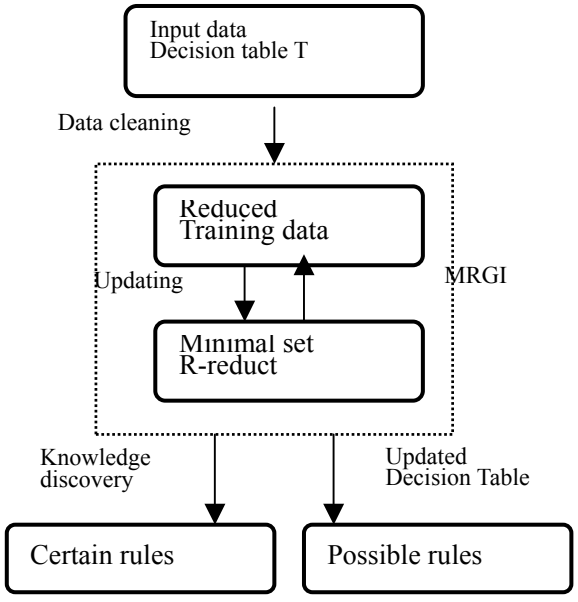

Figure 1. The RGRIPI

The first step of mechanism of RGRIPI is to transform the data set into a decision table T'. This generally involves data cleaning to eliminate obvious redundant or superfluous data that may exist in a typical real world data set. The main step of RGRIPI is use MRGI to generate a minimal set of r-reducts and revise the training data set iteratively. Certain rules of the original data will be directly induced from such a minimal set of rules reducts. Missing data are involved at this stage only in so far as they eliminate rules that would have been classified as "certain", if there were no missing data.

6. Conclusions

In this paper, we proposed a midified rule generation algorithm(MRG) to generate a minimal set of rule reducts and proposed a generalized rule generation algorithm(MRGI) to generate a minimal set of rule directly from the original incomplete information system Based on MRGI, with each rule reduct represents a unique decision rule. We developed a rule generation and rule induction prototype(RGRIPI) to extract certain rules directly from the incomplete information system. RGRIPI can automatically generate a minimal set of decision rules directly from an incomplete data set. We build a probability function combining the plausibility and probability of missing values to compute the possible rules for incomplete information systems.

Acknowledgements

This paper is supported by National Doctor Specialized scientific research Foundation

References

[1] Chmielewski, M.R.;GRzymala-Busse.J.W (1996) "Global discretization fo continuous attributes as preprocessing for machine learning"International journal of approximate resoning,15,319-331.

[2] Lingra,P.J.;Yao,Y.Y.(1998),"Data Mining Using Extensions of the Rough Set Model". Journal of the American society for information scienct.49(5), 415-422.

[3] Chmielewski,M.R.;Grzymala-Busse,J.W.; Peterson, N. W.; Than,S.(1993)"The rule induction system LEAS-A version for personal computers": Found, Compute. Decision Science 18.181-212.

[4] Kryszkiewicz,M..(1998),"Rough set approach to incomplete information systems", Information Sciences,112,39-49.

[5] Kryszkiewicz, M(1999). "Rules incomplete information systems" Information Sciences,113,39-49

[6] Pawlak,Z., "Rough Sets": Theoretical Aspects of Reasoning About Data,Kluwer,Boston.1991.

[7] Jia-Yuarn Guo, "Rough Set-Based Approach to Data Mining",2003.ph.D.