Venkata Naga Sudheer T, Dr. A. Rama Mohan Reddy/ International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 1, Issue 3, pp.677-682

Decision Tree Construction Based on Degree of Rough Sets

Venkata Naga Sudheer T
Department of Computer Science and Engineering
Sri Venkateswara University
Tirupathi, Andhra Pradesh, INDIA.

Dr. A. Rama Mohan Reddy
Department of Computer Science and Engineering
Sri Venkateswara University
Tirupathi, Andhra Pradesh, INDIA.

***Abstract*— Many Data Sets are incomplete, i.e, are affected by missing attribute values. This paper adopts the rough set approach to construct decision tree for incomplete data sets. Here we provide the concept of classification Set Degree (CSD) as the selection criteria for splitting attribute. This approach gives more accurate results when compared with the existing methods based on the Entropy.**

Keywords*—Classification Accuracy, Correlation Factor, Decision Tree, Classification Set Degree (CSD), Rough Set*

I. INTRODUCTION

Decision tree classification is one of the best approach for classification, when compared with the Bayesian Network and Neural Networks. Decision Tree learning is an inductive learning process which is based on example learning, focusing on pushing out rules of decision tree's representation form from cases of no order, no rules. Decision tree approach uses top-down approach, dividing and rule to divide the search space into several disjoint subsets, forming a structure similar to flow chart. This method is faster and very easy to convert into classification rules which are simple and easy to understand. The early existing algorithms like IDE3, C4.5, etc., uses the Entropy and Information Gain as the selection criteria for splitting attribute and those algorithms deal with the complete data sets. Moreover the entropy and information gain consider only the mutual information between attributes. That is the impact from attribute fo of decision results.

Rough Set theory proposed Polish mathematician as a mathematical theory of data analysis by Z. Pawlak in 1982. Rough set approach is used to deal with uncertain and imprecise information. Its characteristic Its characteristic is that it does not need to in advance assign certain characteristics and attributes which are described in quantity, but discovers this question's inherent laws from the given question's description set directly. Its basic philosophy is closer the realistic situation. Now rough set theory has applied to decision tree on part of the study, such as literature [1], first to the sample set do attribute reduction, and then build decision tree according to the core, the decision tree built by this method removes noise and redundant attributes by using the attribute reduction. Literature [6] gives the definition of resolution, and uses resolution as the standard of splitting attributes to build the decision tree. Literature [7] uses the attribute classification roughness of rough set as the standard split attribute, build decision tree on the basis of classification roughness; in addition this paper proposes using the variable precision rough set method to remove noise. Literature [8][9] both use Boundary Region as the standard of splitting attributes, Where [9] in order to avoid the decision tree over-refined introduces inhibitory factor, when the inhibitory factor is less than a certain value, the tree will not expand. Literature [12] proposes to use core attributes and identify matrix to select the attribute which does the greatest contribution to the classification. Literature [13] proposes to use decision attribute on condition attribute dependency as heuristic information to select Properties.

In a rough set approach, it is possible to interpret missing attribute values. There are two types of missing values: lost values and don't care values. Lost values are interpreted as originally given, but currently unavailable due to actions such as incidental deletion, lack of care in recording, etc. A rough set approach to handle the incomplete tables is presented in Literature[14].

Literature[14] uses rough set approach to incomplete data sets in which all attribute values were "do not care" conditions was presented for the first time, where a method for rule induction was introduced in which each missing values was replaced by all values from the domain of the attribute.

Here we present the input data sets as Decision Table. Rows in the decision table represents cases and columns are represented as variables. The set of all cases in the decision table are called instant space. Independent variables are called conditional attributes and dependent variables are called decision attributes. The lost values in the decision table are represented using "?" and don't care values are represented using "*". Additionally we assume that for each case of the decision table must be specified.

For an attribute if there exists a case of lost values, then the case should not be included in blocks of all values of the attribute.

For an attribute if there exists a case of don't care values values, then the case should include in blocks of all values of the attribute.

This paper proposes to use the degree of rough classification to build decision tree. Decision tree based on the degree of rough classification takes the attribute classification accuracy and decision attribute on condition attribute dependency as the standard of splitting attribute. The greater the degree of rough classification of the attribute, the more determining factors are included in the attribute and the greater dependency between the attribute and the decision attribute. After the analysis of many examples, in the process of splitting attribute, the attribute classification accuracy selected by decision tree algorithm based on the degree of rough classification precedes the attribute with maximum information gain selected by ID3 algorithm. Compared with the ID3 algorithm through experiments, under the situation of a slight increase in the number of generated rules, the accuracy has been significantly improved and stable.

II. ROUGH SET THEORY

Information system $S= \{U, A, V, f\}$,
Where
U: Finite set of objects
A: Finite set of attributes $A=C \cup D$
C: Subset of condition attributes
D: Subset of decision attributes
V: Range of the attribute
F: $U \times A \rightarrow V$ is generic function, such that for every $Xi \in U$, $q \in A$,there is $f(Xi.q) \in Vq$

In information $S = \{U, A, V, f\}$, let $X \subset U$ is a subset of individual whole domain, a subset of attributes $P \subseteq A$, then:

Lower approximation of X : $\underline{P}X = \{Y \in U/P : Y \subseteq X\}$

Upper approximation of X : $\overline{P}X = \{Y \in U/P : Y \cap X \neq \phi\}$

Boundary region of X : $Bnd_P(X) = \overline{P}X - \underline{P}X$

$\underline{P}X$ is the set of items which are inevitably classified on $X \subset U$. According to the subset of attribute P, all the items in U which are inevitably classified to the set X, i.e. the greatest definable set included in X.

$\overline{P}X$ is the set of items which are probably classified on U. According to the subset of attribute P, all the items in U which are inevitably and probably classified to the set X, i.e. the least definable set included in X.

$Bnd_P(X)$ is the set of items which are neither classified on $X \subset U$, nor on U-X . The bigger the set X 's $Bnd_P(X)$,the smaller the degree of certainty is.

The positive region about the subset of attributes P to decision attribute D is:
$POS_P(D) = \cup \{ \underline{P}X : X \in U / D \}$
$POS_P(D)$ represents the set constituted by the items which are inevitably classified to the set X on U according to the subset of attributes P .

Approximation accuracy of X on S is:

$$\mu_P(X) = \frac{\mu_P(X)}{\mu_P(X)} = \frac{card(\underline{P}X)}{card(\overline{P}X)} = \frac{card(\underline{P}X)}{card(Bnd_P(X)) + card(\underline{P}X)}$$

Where *Card* is the cardinal of a set

Approximation accuracy reflects the accuracy about the attribute P to the classification of the set X . For an attribute, it is the rate about determined number of samples in positive region to the sum of samples in positive region and boundary region.

Degree of correlation (also degree of dependence) between the subset of condition attributes $P \subseteq C$ and decision attribute D is:

$$K(P, D) = \frac{card(POS_P(D))}{card(U)}$$

Where, $0 \leq K(P,D) \leq 1$, $K(P,D)$, $K(P,D)$ is the degree of correlation between the subset of attributes P and decision attribute D.

If $K(P,D) = 1$, then D totally dependent on P ;
If $0 < K(P,D) < 1$, then D partially dependent on P ;
If $K(P,D) = 0$, then D completely independent on P .

III. AN ALGORITHM FOR DECISION TREE CONSTRUCTION BASED ON DEGREE OF ROUGH SETS

A. Algorithm theory

Information system $S= \{U, A, V, f\}$,

Where U is the finite set of objects and is divided into a finite set of samples X1,X2 ……..Xm, such that $Xi \subseteq U$, $Xi \neq \varnothing$, $Xi \cap Xj = \varnothing$ $(i \neq j)$, i,j=1,2,…m,

$$\bigcup_{i=1}^{m} X_i = U$$

A: Finite set of attributes $A=C \cup D$
C: Subset of condition attributes
D: Subset of decision attributes

An attribute $p \in P \subseteq C$, then the degree of rough classification is defined as:

$$CSD(p, C, D) = K(P, D) * \sum_{i=1}^{m} \mu_P(X_i)$$

Where $\sum_{i=1}^{m} \mu_p(X_i)$ represents the sum of classification accuracy of the attribute P to every set of decision attributes.

$\mu_p(X_i)$ represents the classification accuracy of the set of sampl Xi.

The bigger $\mu_p(X_i)$ is, i.e, $\frac{card(\underline{P}X_i)}{card(Bnd_p(X_i))+card(\underline{P}X_i)}$ is bigger, the less uncertain factors *BndP(Xi)* brings and the result of correlation is better on contrary. The result of classification is not obvious.

K(P,D) represents the importance of the attribute P, the bigger K(P,D) is, the degree of correlation between the attributes set P and decision attribute D is grater. When the attributes set P has only the one attribute p, there is the following formula,

$$k(P,D)=\frac{card(POS_P(D))}{card(U)}=\frac{card(\sum_{i=1}^{m}\underline{p}X_i)}{card(U)}$$ where i=1,2,…m.

Literature [6] proofs that there is compatibility relation between mutual information and *card*(*POSP*(*D*)), for an information system *S* , *card*(*U*) is unchanged. Therefore, there exists compatibility relation between *K*(*P*,*D*) and mutual information, i.e. *K*(*P*,*D*) could well reflect the dependence of attribute *p* on the decision attribute. When *CSD*(*p*,*C*,*D*) =0 , i.e. *K*(*P*,*D*) =0 or $\sum_{i=1}^{m}\mu_p(Xi)=0$,i.e, $\mu_p(Xi)=0$. It represents that the lower approximation of *Xi* on the attribute *p* is null. The classification accuracy and lower approximation of the attribute are 0,then the classification contribution of the attribute to the samples is 0.

Synthesizing the above analysis, we use the Classification Set Degree *CSD*(*p*,*C*,*D*) as the standard of splitting attributes. This is to ensure efficient construction of decision trees, and he dependency between condition attributes and decision attributes.

B. Algorithm flow

Calculate the *CSD*(*p*,*C*,*D*) of every attribute as the standard of splitting attributes, and select the attribute whose *CSD*(*p*,*C*,*D*) is maximum as the splitting attribute, i.e. take the attribute which has the greatest influence on the decision results and the higher classification accuracy as a splitting node.

The procedure of the algorithm for decision tree construction based on degree of rough classification as following:

Input: *training samples, set of attributes*

Output: *a decision tree*

1. Create node N
2. If Samples are in the same class C, return N as leaf node, marked by class C
3. If attribute_list is null, return N as leaf node, marked by the most common class in Samples
4. Calculate lower and upper approximation of every attribute in attribute_list. If lower approximation is null, then *CSD*(*p*,*C*,*D*) =0
5. Calculate degree of rough classification *CSD*(*p*,*C*,*D*) of every attribute attribute_list. Select the attribute test_attribute which has the biggest *CSD*(*p*,*C*,*D*) in attribute_list
6. Mark node N as the splitting attribute test_attribute
7. For every known value *a* in test_attribute;
   - Grow a branch whose condition is test_attribute= ai from the node N
   - Let si as the set which satisfies test_attribute=ai in Samples
   - If si is null,
     add a leaf, use the most common class to mark.
   - Otherwise, add a node returned from Generate_Decision_Tree(Si,attribute_list-test_attribute)
8. The procedure of constructing tree is a recursive
9. the terminal condition is:
   - All the samples of the given node belong to the same class(step 2)
   - No remaining attributes can further divided samples (step 3)
   - No samples satisfy test_attribute = a(step 10)

The algorithm for decision tree construction is repeated until the terminal condition is encountered. The algorithm can be applied for both the precise and imprecise data sets. The Generate Decision Tree function is used to generate at each level of the splitting attribute.

C. Example

Table 1 is some part of mud logging explanation result data selected from oil exploration database, and choose one of the important conclusions of the 8 which affect interpretation conclusion to make up the set of condition attributes, a decision attribute, a set of training samples made up with 25 objects. Using the algorithm for decision tree construction based on degree of rough sets to construct decision tree, we have discrete values which are continuous values of the samples before the construction. The procedure of constructing decision tree as following:

Using the algorithm for decision tree construction based on degree of rough sets to construct decision tree. The set of samples classified by the decision attribute "Interpretation result":

X1 ={1, 4,11,13,15,16, 21, 22, 24, 25}

X 2 = {2, 3, 5, 6, 7, 8, 9,10,12,14,17,18,19, 20, 23, 26}

First, calculate lower approximation and upper approximation of every attribute *pi*∈*C* on the set of samples, then calculate Approximation accuracy $\sum_{i=1}^{m}\mu_p(X_i)$ and the correlation factor *K(P,D)* at the last get the *CSD(p,C,D)*.

Table 1
INCOMPLETE DECISION TABLE

	Conditional Attributes									*Decision Attribute*
	Noof wells	***Depth of top Boundary*** ***P1***	***Depth of Bottom boundary*** ***P2***	***Base value of methane*** ***P3***	***Base value of ethane*** ***P4***	***Base value of Butane*** ***P5***	***Base value of iso butane*** ***P6***	***methane outlier*** ***P7***	***Ethane Outlier*** ***P8***	
1	Fengshen10	3622	?	10.26	1.736	0.226	0.8	87.18	4.34	Reservoir
2	Fengshen10	3864	3265	5.09	0.815	0.082	0.512	5.95	0.885	Dry Layer
3	Fengshen10	3911	3865	3.03	0.458	0.031	0.113	11.42	1.536	Dry Layer
4	Fengshen10	3920	3468	0.15	0.034	0.005	0.03	0.28	0.05	Reservoir
5	Fengshen10	*	4138	10.63	1.049	0.248	0.182	16.57	1.657	Dry Layer
6	Fengshen10	3957	4028	14.79	1.173	0.081	0.3	19.82	2.111	Dry Layer
7	Fengshen10	4137	3098	17.64	1.509	0.096	0.304	23.85	2.095	Dry Layer
8	Fengshen10	4169	?	7.3	0.975	0.131	0.489	11.27	1.473	Dry Layer
9	Fengshen10	4234	*	8.21	1.102	0.158	0.564	10.48	1.165	Dry Layer
10	Fengshen10	4265	4268	7.52	0.641	0.158	0.711	8.31	0.883	Dry Layer
11	Fengshen6	4269	4275	0.75	0.142	0.036	0.099	1.76	0.389	Reservoir
12	Fengshen6	*	4209	0.24	0.048	0.025	0.098	1.25	0.176	Dry Layer
13	Fengshen6	3465	4023	1.74	0.185	0.018	0.075	2.46	0.229	Reservoir
14	Fengshen6	3501	*	1.04	0.137	0.014	0.075	4.27	0.533	Dry Layer
15	Fengshen6	*	3208	2.62	0.372	0.021	0.087	12.89	1.787	Reservoir
16	Fengshen6	3559	4097	1.14	0.163	0.019	0.076	7.5	0.92	Reservoir
17	Fengshen6	3572	3843	0.48	0.052	0.008	0.024	1.39	0.09	Dry Layer
18	Fengshen6	3656	3574	0.53	0.059	0.008	0.02	0.53	0.059	Dry Layer
19	Fengshen6	3679	*	0.77	0.069	0.01	0.018	0.77	0.069	Dry Layer
20	Fengshen6	*	3743	0.94	0.047	0.005	0.019	0.94	0.047	Dry Layer
21	Fengshen6	3703	3840	0.36	0.026	0.007	0.011	0.93	0.074	Reservoir
22	Fengshen6	3740	?	0.75	0.059	0.005	0.018	1.06	0.175	Reservoir
23	Fengshen6	3837	3939	0.4	0.029	0.007	0.013	2.69	0.175	Dry Layer
24	Fengshen6	3856	3956	0.59	0.055	0.008	0.019	1.69	0.134	Reservoir
25	Fengshen6	3876	*	0.71	0.06	0.007	0.021	0.89	0.083	Reservoir
26	Fengshen6	3981	4023	0.14	0.014	0.004	0.011	0.42	0.099	Dry Layer

*- Don't care value ?- lost value

For the attribute p_8="Ethane outlier", the upper approximation and lower approximation on the set of samples $X1X2$ are:

$\underline{p_8}X1=\{1\}$, $\underline{p_8}X2 = \varnothing$, $\overline{p_8}X1=\{1,2,3,...,26\}$, $\overline{p_8}X2=\{2,3,...,25,26\}$.

The argument, the $CSD(pi,C,D)$ of other attributes are:

$CSD(p_1,C,D)=0$, $CSD(p_2,C,D)=0$, $CSD(p_3,C,D)=0$, $CSD(p_4,C,D)=0$, $CSD(p_5,C,D)=0$, $CSD(p_6,C,D)=0$, $CSD(p_7,C,D)=0$, $CSD(p_8,C,D)=0$.

The maximum of them is p_8 , so the attribute of the first node is "Ethane outlier". Then determine the set of it's branches, where the samples which are ">3.2255" are in the same class, so it reaches the leaf node. By parity of reasoning, the child node of "<=3.2255" is "Base value of methane". At last get the decision tree, as Fig. 1 expresses:

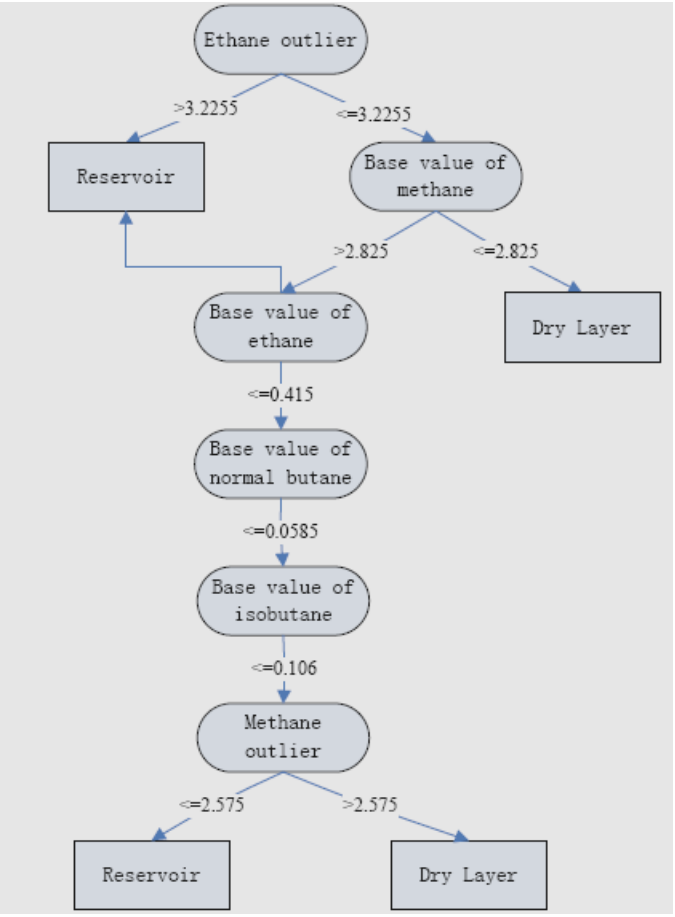

Fig 1: Decision Tree based on Degree of rough sets algorithm.

From the resulted decision tree, we can have the classification accuracy of the decision tree process is increased. The dependence of the decision tree based of degree of rough sets get he advantage.

IV. CONCLUSION

This paper designs a new algorithm for decision tree based on degree of rough sets. On he splitting attribute, this algorithm considers both the classification accuracy of the attribute and the dependence condition attribute on decision attribute. Through a large number of example analysis, it proves that on the condition of considering to the dependence, the algorithm for decision tree based on degree of rough sets is better than ID3 algorithm.

V. REFERENCES

[1] WU Cheng-dong, XU Ke, HAN Zhong-hua, PEI Tao. Approach to Data Mining Based on Rough Sets and Decision Tree [J]. Journal of Northeastern University, 2006, 27(5): P481-484.
[2] QUINLAN J R. Induction of Decision Tree [J]. Machine Learning,1986, P81～106.
[3] ZDZISLAW P. Rough set theory [J]. Kunstliche Intelligenz, 2001, 5(3): P38-39.
[4] HU Keyun, LU Yunchang, SHI Chunyi. Feature ranking in rough sets [J]. AI Communications, 2003 16: P41-50.
[5] Qiao Mei, Han WenXiu. Decision Tree Algorithm Based on Rough Set [J]. Journal of Tianjin University, 2005, 38(9): P842-846.
[6] GAO Jing, XU Zhang-yan, SONG Wei, YANG Bing-ru. New Decision Tree Algorithm Based on Rough Set Model [J]. Computer Engineering, 2008, 34(3): P9-11.
[7] WEI Jinmao. Rough Set based Approach to Selection of Node [J]. International Journal of Computational Cognition, 2003, 1(2):
P25-40.
[8] WEI Jinmao, HUANG Dao, WANG Shuqin. Rough Set based Decision Tree [C]. Proceedings of the 4th World Congress on Intelligent Control and Automation (VCICA), Shanghai, China: IEEE, Jan, 2002: P426-431.
[9] BLEYBERG M Z, ELUMALAI A. Using Rough Sets to Construct Sense Type Decision Trees for Text Categorization [C]. IFSA World Congress and 20th VAFIPS International Conference, Vancouver, BC, Canada: July, 2001: Vol1, P19-24.
[10] SANG W H, JAE Y K. Rough Set-based Decision Tree using the Core Attributes Concept [C]. Second International Conference on Innovative Computing, Information and Control (ICICIC 2007), Japan: IEEE, 2007: P298-301.
[11] WANG Cuiru, OU Fangfang. An Algorithm for Decision Tree Construction Based on Rough Set Theory [C]. 2008 International Conference on Computer Science and Information Technology (ICCSIT 2008), Singapore: IACSIT, 2008: P295-298.
[12] SONAJHARIA M, RAJNI J. Rough Set Based Decision Tree Model for Classification [C]. 5th International conference on data warehousing and knowledge discovery (DaWaK 2003), Prague, Czech Republic: DEXA Society, 2003: Vol2737, P172-181.

[13] SANG W H, JAE Y K. A New Decision Tree Algorithm Based on Rough Set Theory [J]. International Journal of Innovative Computing Information and Control, 2008, 4(10): P2749-2757.
[14] ZHANG et al, An Algorithm for Decision Tree Construction Based on Degree of Rough Classification[C]. International Conference on Artificial Intelligence and Computational Intelligence, 2010, P23-29.
[15] MATTHEW N ANYANWU, SAJJAN G SHIVA, Comparative Analysis of Serial Decision Tree Classification Algorithms[J]. International Journal of Computer Science and Security, 2008, Volume(3): P 230-240
[16] JERRY W GRZYMALA BUSSE, Rough Set and CART approaches to Mining Incomplete Data, 2010, IEEE, P214-219.