Global Journal of researches in engineering
Numerical Methods
Volume 11 Issue 7 Version 1.0 December 2011
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 2249-4596 & Print ISSN: 0975-5861

Decision Tree Construction: A Continues Label Support Degree Based Approach

By N.Madhuri, T.Nagalakshmi, D.Sujatha
ATRI, Parvathapur, Uppal, Hyderabad, India

Abstract - Data mining and classification systems utilize decision tree algorithms since they proffer rapid speediness, advanced exactness and also simple organization of those algorithms. An ideal decision can be built only when the appropriate attributes are chosen. This paper focuses on throwing light on choosing characteristics based on the theory of attribute support degree on account of which a unique decision tree construction algorithm is proposed on the basis of rough set and granular computing theory. It is henceforth proved that the decision tree proposed by the new approach yields far more better results in terms of precision and consistency as compared to the decision trees yielded by ID3, C4.5 and DTBAS.

Keywords : *rough set: decision tree: granular computing: attribute support degree: attribute selection.*

GJRE Classification : *FOR Code: 050299*

Strictly as per the compliance and regulations of :

Decision Tree Construction: A Continues Label Support Degree Based Approach

N.Madhuri[α], T.Nagalakshmi[Ω], D.Sujatha[β]

Abstract - Data mining and classification systems utilize decision tree algorithms since they proffer rapid speediness, advanced exactness and also simple organization of those algorithms. An ideal decision can be built only when the appropriate attributes are chosen. This paper focuses on throwing light on choosing characteristics based on the theory of attribute support degree on account of which a unique decision tree construction algorithm is proposed on the basis of rough set and granular computing theory. It is henceforth proved that the decision tree proposed by the new approach yields far more better results in terms of precision and consistency as compared to the decision trees yielded by ID3, C4.5 and DTBAS.

Keywords : rough set: decision tree: granular computing: attribute support degree: attribute selection

I. INTRODUCTION

Decision sets can be denoted using tree structures with the help of decision tree which is a unique, spontaneous, data illustration scheme and also a competent classifier. Quinlan et al[1] proposed ID3, decision tree algorithm and hence has been persistently augmented which have been advanced to C4.5 [2]. The preeminent attribute is chosen as the existing attribute which is then recursively inflates the decision tree branches unless and until the conditional statement is achieved, which ultimately makes use of top-down greedy algorithms. There are different classification schemes that can be achieved concerning different solutions which poses two issues [3] in decision tree construction. Choosing characteristics for crafting new branches in a tree is one issue while the other one is pruning which is all about omitting and decreasing the tree. DTBAS[10] considered the Assortment of attribute as main concern, which was refined and improved by considering assortment of continuous labels also that is discussed in this paper.

Z. Pawlak et al[4] recommended rough set theory which is an expansion of set theory for studying intelligent systems which is followed up by inadequate and partial data information. There is a thriving submission of the rough set theory in the disciplines of data mining, pattern recognition, machine learning, decision analysis etc in recent times. Models are categorized into various resembling classes that houses imperceptible objects in terms of few attributes. Issues pertaining to feature selection, data reduction and pattern extraction can be amicably taken care of such that it can liberate the system of redundant data in systems containing null values or missing data.

Lin et al[5] proposed the expression of Granular Computing which spans itself covering all aspects of concerning theories, tactics, practices and means essential in solving a problem that makes use of granules. Granular Computing has witnessed vast inputs from different practices such as fuzzy sets, rough sets, shadowed sets, probabilistic sets etc.

A crucial step that needs to be taken care of while building a decision tree is choosing characteristics of nodes of a tree that houses minimum number of branches. Decision tree based on continuous label support degree (DTBLSD) algorithm is introduced which is considered as a splitting criterion on account of rough set theory and granular computing. Trial results have approved the usage of DTBLSD algorithm that assures and provides uncomplicated structures and superior categorization accuracy.

The rest of the paper is organized as follows: Section 2 discusses concepts relevant to rough set theory and granular computing. Section 3 gives a basic introduction to our new method and presents a simple example. Experimental comparison of the proposed method with 103 and C4.S is given in section 4. The final section concludes the research work of this paper.

II. BASIC CONCEPT

Few fundamental concepts of rough set theory [6, 7] and granular computing[8] are first initiated for ease of demonstration.

Definition 1 (Information System) : An information system can be labeled as $S = (U, A, V, f)$ wherein U is a finite set of object known as the universe; $A = C \cup D$, which is a non-vacant finite group of attributes; C and D depict set of condition and decision attributes respectively as also $v = Uv_a, \forall a \in A$, which says that v_a is a value set of the attribute a and

Author α : M.Tech, Department of CSE, ATRI, Parvathapur, Uppal, Hyderabad, India. E-mail : madhuri.neerubavi@gmail.
Author Ω : Sr,Asst.prof, Department of CSE, ATRI, Parvathapur, Uppal, Hyderabad, India. E-mai : nlakshmi.t@gmail.com
Author β : Assoc.Prof and HOD, Department of CSE, ATRI, Parvathapur, Uppal, Hyderabad, India.
E-mail : sujatha.dandu@gmail.com

$f = U \times A \rightarrow V$ is known as the information function also known as total decision function, such that $f(u,a) \in V_a$, for every $a \in A$, $u \in U$.

Definition 2 (Lower and Upper Approximation of Sets, Boundary of $X(BND(X))$ Let $S = (U, A, V, f)$ be an information system, and let $R \subseteq A$ and $X \subseteq U$. X can be estimated using information present in R by building lower and upper estimate $\underline{R}(X)$ and $\overline{R}(X)$ and $BND(X)$ known as boundary of X. We can now designate $\underline{R}(X)$ Ji(x), $\overline{R}(X)$, and $BND(X)$ as follows.

$$\underline{R}(X) = \{x \in U \mid [x]_R \subseteq X\} \ldots (1)$$

$$\overline{R}(X) = \{x \in U \mid [x]_R \cap X \neq \phi\} \ldots (2)$$

$$BND(X) = \overline{R}(X) - \underline{R}(X) \ldots (3)$$

Definition 3 (Indiscernibility Relation) Let $S = (U, A, V, f)$ be an information system and subset $P \subseteq A$ be known as the indiscernibility relation, indicated by IND(P), that can be termed as

$$IND(P) = \{(x, y) \in U \times U \mid \forall_a \in P, f(x,a) = f(y.a) \ldots (4)$$

Where *IND(P)* is an equivalence relationship that separates U into equivalence classes labeled as $\frac{U}{IND(P)}$, which contains group of objects that is unobvious concerning A.

Definition 4 (Granular Degree) suppose $K = (U, R)$ is a repository. $R \subseteq U \times U$ is equivalence relations in the universe U known as knowledge. $GD(R)$ depicts Granular degree of knowledge R. Its definition is as follows:

$$GD(R) = \frac{|R|}{|U \times U|} = \frac{|R|}{|U|^2} \ldots (5)$$

Where $|R|$ is the cardinal number of $R \subset U \times U$.

When in an equivalence relation R, the granular degree of R reaches the minimum size $|U|/|U|^2 = 1/|U|$; When R is a domain relation, the granular degree of R attains the maximum size $|U|^2/|U|^2 = 1$

Definition 5 : Assume R is knowledge of repository $K = (U, R)$, $U/R = \{X_1, X_2, \ldots, X_n\}$, the granular degree of basic knowledge is defined as

$$GD(X_i) = \frac{\sum_{i=1}^{n} |x_i|^2}{|U|^2} \ldots (6)$$

III. Proposed Algorithm

This section aims at familiarizing the algorithm of building a decision tree on the basis of attribute support degree.

a) The Principle of Label and Attribute Selection

The label that represents least average uncertainty is supposed to be chosen as the test label and then choose attribute with less uncertainty as test attribute from the class represented by the selected label, which because it makes apt decisions when compared to existing test attribute selection in different decision tree algorithms.

Definition 6 : (Label support Degree) let Label l representing the group of attributes here is total number of attributes grouped under label .Then average uncertainty represented by label can measure as $\{a_1, a_2, \ldots . a_{ls}\}$

$$avg_{uc}(l) = \frac{\sum_{i=1}^{||} uc(a_i)}{|l|} \ldots (7)$$

Here $avg_{uc}(l)$ represents average uncertainty of label l $uc(a_i)$ Represents uncertainty of attribute a_i of label l

Definition 7 (Attribute Support Degree) Let $S = (U, A, V, f)$, $A = C \cup D$ be an information system l is a label contains subset of attributes represented as $Q \subseteq C$. Attribute support degree can be denoted as follows based on the definitions mentioned above.

$$S(Q, D) = \frac{GD(Q \cup D)}{GD(Q)} = \frac{|IND(QUD)|}{IND(Q)} \ldots (8)$$

Where $|IND(Q)|$ denotes the cardinal number of $IND(Q) \subseteq U \times U$.

D using Q can be estimated with the help of a measure S (Q,D) . Definition 5 states that whenever we get the relations among them, namely, when GD(R) is smaller, the distinguishable degree is stronger and S(Q,D) is greater, thereby Q is better sets of test attribute of D. On the contrary, the smaller S(Q,D) is, the worse we get Q as sets of test attribute of D.

b) The Description of DTBLSD

The basic notion of DTBLSD expresses the point that whenever label support degree with association of label level attribute support degree is made use of as a customary for choosing a test attribute concerning every node in the decision tree. The attribute reduction set assists in selecting a condition attribute that possesses the highest degree of label level attribute which can be put to use at the root of the decision tree. There will be a testing of the remaining condition attributes on each and every branch of the root node

and so, the algorithm persists in a recursive manner by addition of new sub-trees to every division until the leaf is reached.

According to the above idea, using the S(Q,D) as the splitting criterion, we propose our algorithm DTBLSD. Current sample set is depicted by T, set of labels depicted by L, condition attribute set of a label is depicted by the l_{al}. $|l_{al}|$ depicts the number of attributes in the condition attribute set of label l. All attributes of the condition attribute set are discrete and continuous values are discretized by continuous labeling. Following are the specific steps of the algorithm.

Algorithm : A decision tree is created by DTBLSD (T, attribute list) that using the given training data.

Input : The training set samples, represented by discrete valued attributes; the set of condition attribute, attribute list.

Output : A decision tree.

Step1 : create a node N;

Step2 : if samples are all of the same class C, then return N as a leaf node labeled with the class C;

Step3 : if attribute list is empty, then return N as a leaf node labeled with the most common class in the samples;

Step 4: Select a label l that represents average uncertainty is low.

Step5 : select test attribute in l_{al} with the highest degree of attribute support;

Step6 : label node N with test attribute;

Step7 : for each known value aj of test attribute, grow a branch from node N according to the condition test attribute= aj;

Step8 : let Sj be the set of samples in samples for which test attribute= aj;

Step9 : if Sj is empty, then attach a leaf labeled with the most common class in samples;

Step10 : else attach the node returned by . $DTBLSD(s_i, l_{al}, test\_attribute)$.

The top-down recursive divide and conquer approach for construction of a decision tree wherein the recursion related division takes place only when any one criterion mentioned below is gratified. A common class contains :

1. All specimens for a specific branch which restores a leaf that is termed with the concerned class. Here, a large case of voting is provisioned to change the present working node into a leaf that is termed with the concerned class that in in demand from amongst various specimens.
2. In addition, there are no more specimen test attributes and the class division specimens can be placed wherein a leaf is generated and termed with the most featured class in specimens.

IV. Experiments

a) Example Analysis

Table 1 showcases a data tuple training group originated from All Electronics customer records that are implemented using polic mentioned in reference (6). The first step is to estimate the degree of attribute support for each situational attribute or characteristic.

$\frac{U}{IND(al)} = \{ \{l,2,8,9, 11 \},\{3,7,12,13\},\{ 4, 5, 6, 10, 14\}\};$

$\frac{U}{IND(az)} = \{ \{1, 2, 3, 13\},\{4, 8, 10, 11, 12, 14\}, \{ 5, 6, 7, 9\}\};$

$\frac{U}{IND(a3)} = \{\{1, 2, 3, 4, 8, 12, 14\}, \{5, 6, 7, 9, 10, 11, 13\}\};$

$\frac{U}{IND(a4)} = \{ \{l,3,4,5,8,9,10,13\},\{2,6,7,11,12,14\}\};$

$\frac{U}{IND(d)} = \{ \{1 ,2,6,8,14 \},\{3,4,5,7,9, 10, 11,12, 13\}\};$

$\frac{U}{IND(al, D)} = \{\{I, 2, 8\},\{3, 7, 12, 13\},\{4,S, 1 0\},\{6, 14\},\{9, II\}\};$

$\frac{U}{IND(az, d)} = \{\{1, 2\},\{3, 13\},\{4, 10, 11, 12\},\{5, 7, 9\},\{6\},\{8, 14\}\};$

$\frac{U}{IND(az, d)} = \{\{l, 2, 8, 14\}, \{3, 4, 12\}, \{5, 7, 9, 10, 11, 13\}, \{6\}\};$

$\frac{U}{IND(a4, d)} = \{\{1, 8\}, \{2, 6, 14\}, \{3, 4, 5, 9, 10, 13\}, \{7, 11, 12\}\}.$

$$S(a_1, d) = \frac{|IND(a_1 \cup d)|}{IND(a_1)} = 0.636$$

$$S(a_2, d) = \frac{|IND(a_2 \cup d)|}{IND(a_2)} = 0.559$$

$$S(a_3, d) = \frac{|IND(a_3 \cup d)|}{IND(a_3)} = 0.633$$

$$S(a_4, d) = \frac{|IND(a_4 \cup d)|}{IND(a_4)} = 0.580$$

Table I : Training data tuple from the AllElectronics customer database

U	age	income	student	credit	buy
1	<=30	high	no	fair	No
2	<=30	high	no	excellent	No
3	31<=age<=40	high	no	fair	yes
4	>40	medium	no	fair	yes
5	>40	low	yes	fair	yes
6	>40	low	yes	excellent	No
7	31<=age<=40	low	yes	excellent	Yes
8	<=30	medium	no	fair	No
9	<=30	low	yes	fair	Yes
10	>40	medium	yes	fair	Yes
11	<=30	medium	yes	excellent	Yes
12	31<=age<=40	medium	no	excellent	Yes
13	31<=age<=40	high	yes	fair	Yes
14	>40	medium	no	excellent	No

Notations used in example descriptions:

Age➔ , Income➔ , Student➔ , Credit➔ , Buy➔

a_1 is selected as the min root of the decision tree and is tagged with age since $S(a_1,d)$ is the maximum extent of a degree from amongst all the condition attributes as also various number of divisions which are branched in reference to a range of different attributes. In case where age=1, all the specimens that are grouped into this should belong to the same class and hence a leaf should be generated at the end of every division and should be tagged with d=yes. The figure above depicts the final decision tree that is built by DTBLSD.

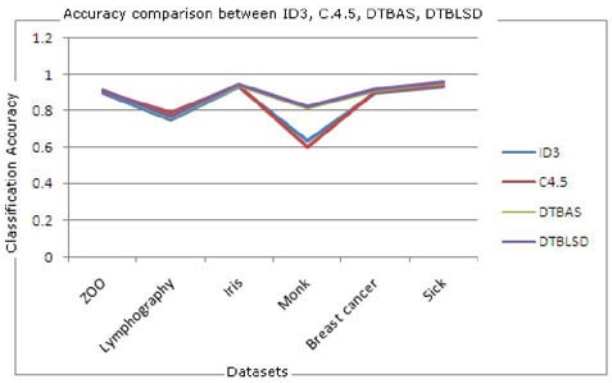
b) Experimentl Comparison

Experimental comparison of DTBLSD with respect to ID3[1], C4.5[2] and DTBAS[10] is discussed in this section. The real datasets that are used in this are approved from University of California, Irvine (UCI), and is known as the machine learning database repository where C++ design language is implemented to form the requisite algorithm. WEKA 3.7 is used for successful accomplishment of ID3[1] and C4.5[2] which is a compilation of machine learning algorithms used for data mining generated and procured by Frank that involved the 10 fold cross estimation to calculate classification authenticity. All experiments were performed on a PC, Intel(R) Pentium(R) 4 CPU, 2.93GHz, 512MB RAM.

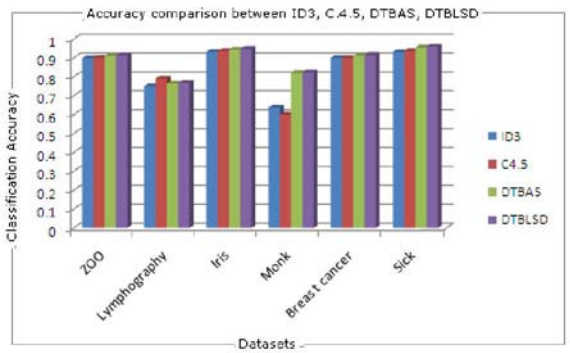
It has been observed from the analytical results that as compared to ID3, C 4.5 and DTBAS, our specified algorithm has better accuracy and lower computation cost. Listing of comparison report follows:

Table II : Tabular representation of accuracy comparison between ID3, C4.5, DTBAS, DTBLSD

Dataset	ID3	C4.5	DTBAS	DTBLSD
ZOO	0.899	0.901	0.911	0.914
Lymphography	0.75	0.791	0.765	0.767
Iris	0.932	0.936	0.943	0.947
Monk	0.637	0.6	0.821	0.824
Breas t cancer	0.9	0.9	0.912	0.916
Sick	0.931	0.937	0.955	0.959



(a) Line Chart of Comparison report



(b) Bar chart of comparison report

Fig 1: Accuracy comparison graphs

V. Conclusion

The paper first focuses on explaining the basic notion of label support degree and attribute support degree [10] and selecting it as a basic decisive factor on the basis of degree of involvement between condition attribute and decision attribute accordingly where a unique decision algorithm tree based on continuous label support degree and label level attribute support degree (DTBLSD) is recommended. Accordingly a suitable methodology is devised which is flexible enough to accommodate and provides lower complexity and high level of accuracy as compared to other algorithm generating methods. A disadvantage identified in [10] is issues pertaining to adjustment with adaptability of samples, which has been overcome successfully in our model.

References References Referencias

1. J. R. Quinlan, "Induction of Decision Trees", Machine Learning, voLl, 1986, pp. 81-106.
2. J. R. Quinlan, "Improved Use of Continuous Attributes in C4.5", Journal of Artificial Intelligence Research, volA, 1996, pp.77-90.
3. F. Seifi, H. Ahmadi, M. Kangavari, "Twins Decision Tree Classification: A Sophisticated Approach to Decision Tree
4. Z. Pawlak, "Rough Sets," International Journal of Information and Computer Science, voLII, 1982, pp. 341 -356.
5. T. Y. Lin, "Granular Computing II Infrastructure for AI-Engineering Examples, Intuitions and Modeling," Proc. 2006 IEEE International Conference on Granular Computing(GrC 2006), 2006, pp. 2-7.
6. J. W. Han, M. Kamber, Data Mining Concepts and Techniques[M). Morgan Kaufmann Publishers, 2001.
7. B. S. Ding, Y. Q .. Zheng, S.Y. Zang, "A New Decision Tree Algorithm Based on Rough Set Theory," Proc. Asia-Pacific Conference on Information Processing(APCIP 2009), 2009, pp. 326-329.
8. D. Q. Miao, G. Y. Wang, Granular Computing: Past, Now, Future[M), Science publishing house, 2007.
9. E. Frank, M. Hall, WEKA, http://www. cs. waikato.ac. nz/ml/weka
10. Qing Lin; Zongzhuan Ding; Jianping Yong; Jun Zhou; , "An algorithm of decision tree construction based on attribute support degree," Educational and Information Technology (ICEIT), 2010 International Conference on , vol.2, no., pp.V2-516-V2-519, 17-19 Sept. 2010.
11. Y.L.Chen,C.L. Hsu, and S.C. Chou, "Constructing a Multi-Valued and Multi-Labeled Decision Tree",Expert Systems with Apllications,vol.25,pp. 199-209, 2003.

This page is intentionally left blank