International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 2 Issue 6, June - 2013

Prodtu: A Novel Probabilistic Approach To Classify Uncertain Data Using Decision Tree Induction

Swapnil Andhariya, Khushali Mistry, Prof. Sahista Machchhar, Prof. Dhruv Dave
[1]*Student, Marwadi Education Foundation's Group of Institution, Rajkot,*
[2]*Student, Marwadi Education Foundation's Group of Institution, Rajkot,*
[3]*Assistant prof., Marwadi Education Foundation's Group of Institution, Rajkot*
[4]*Assistant prof., LDRP Institute of Technology and Research, Gandhinagar*

***Abstract*-- Classification is a data mining technique used to predict group membership for data instances. Traditional decision tree classifiers work with data whose values are known and precise. We extend such classifiers to handle data with uncertain information. Value uncertainty arises in many applications during the data collection process. Sources of uncertainty include measurement/quantization errors, data staleness, and multiple repeated measurements. A way to handle data uncertainty is to abstract probability distributions by summary statistics such as means and variances. This approach is known as averaging. Another approach is to consider the complete information carried by the probability distributions to build a decision tree. This approach is known as Distribution-based. In this thesis I have devised novel algorithm to build decision trees for classifying such uncertain data. I have applied suitable probability distribution due to that data uncertainty leads to decision trees with remarkably higher accuracies.**

Keywords- Classification, Data Mining, Decision Tree Induction, ProDTU, Uncertain Data.

1. Introduction

Data Mining refers to extracting or mining knowledge from large amounts of data. It is the process of sorting through large amounts of data and picking out relevant information through the use of certain sophisticated algorithms. Data mining is also known as "knowledge mining from data". There are many other terms similar to data mining such as knowledge extraction, data dredging, data or pattern analysis [13].

Classification is one type of data mining technique used to predict group membership for instances of data. Classification of data is a two-step process, the first step is known as training phase and second step is known as testing phase. In the training phase, a classifier is built by describing

a predetermined set of data classes[13]. This is also known as learning step. In testing phase, the model is used for classification [13]. One of the most popular classification techniques is the decision tree induction. Among this entire decision tree is very popular method because they are practical and easily understandable. Rules can be extracted easily from decision trees. Many algorithms like ID3 [2] and C4.5 [3], have been created for decision tree construction.

Classification and regression techniques used decision tree methodology. Its advantage lies in the fact that it is easy to understand; also, it can be used to predict patterns with missing values and categorical attributes [5]. Decision trees are used in many domains. For example, in database marketing, decision trees can be used to segment groups of customers and develop customer profiles to help marketers produce targeted promotions that achieve higher response rates. Decision tree is also robust and scalable. It performs well with large data in a short period of time [6].

In this paper section 2 describes the related work. Traditional decision trees are described in section 3. Section 4 discusses how to handle uncertain information. In section 5, we discuss proposed system. And performance analysis is to be described in section 6. In section 7 conclusions and future work is to be discussed. And finally section 8 describes references.

2. Related work

Perhaps most of common source of uncertainty comes from repeated measurements types of Uncertainty. For example a patient's body temperature can be measured multiple times within a day, another example is an anemometer can record wind speed once every minute [1].
Data uncertainty has been classified into two type of uncertainty. First one is existential uncertainty and second one is value uncertainty. When it is uncertain, whether a data tuple exists, Existential uncertainty appears. For example, relational database tuple can associate with a probability which represents like the confidence [8]. "Probabilistic Database" has been applied to XML and semi structured data [9], [10]. While Value uncertainty, is appears when one data tuple is known to exist, but that tuple's values are not known. A data with value uncertainty is generally represented by a probability distribution function over

bounded and a finite region of possible values [11], [12]. There are missing value appeared in the decision tree induction for uncertain data [2], [3]. These types of missing values are appears due to data entry errors or because of some attribute values which are not available during data collection. In probabilistic decision trees and C4.5 [3] missing values appeared in first phase of classification, which is training phase. Here data handled by using fractional tuples. Each and every missing value is replaced with multiple values with probabilistic based training tuples in testing phase of classification

3. Traditional Decision Tree

In the normal model, a data set has d training tuples which are known as $\{t_1, t2, \ldots t_d$. and k numerical real valued or feature attributes, which are known as $A_1 \ldots A_k$. The domain of attribute A_j is written as $Dom(A_j)$. here for each and every tuple ti is associated with a feature vector , which is written as $Vi = (v_{i,1}, v_{i,2} \ldots v_{i,k})$ and a class label, which is written as c_i, where $v_{i;j} \in dom(A_j)$ and $c_i \in C$, here C has the set of all class labels. The main problem of classification is to construct a model M that maps each feature vector $(v_{x,1}, v_{x,2} \ldots v_{x,k})$ to P_x,which is probability distribution on class label C so that a test tuple $t_0 = (v_{0,1}, v_{0,2} \ldots v_{0,k}, c_0)$ and probability distribution $P_0 = M (v_{0,1}, v_{0,2} \ldots v_{0,k})$ predicts the class label c_0 which have high accuracy. There are two types of data are to be considered numerical as well as categorical.

In this section we discuss about the binary decision tree for numerical value. In this binary tree each internal node of tree is associated with an A_{jn} and $z_n \in dom(A_{jn})$,where $v_{0,j} \leq z_n$ is satisfied. In this binary tree each and every internal node has two children, which are named as "left child" and "right child" respectively. Each terminal node or leaf node in the tree is associated with a discrete P_m over class lable C.

The class label of a test tuple $t_0 = (v_{0,1}, v_{0,2} \ldots v_{0,k}, ?)$, here the traversing of the tree is occurred from starting of the root node and going into depth until terminal node is to be reached. Whenever visit an non terminal or internal node n, executing the test $v_{0,j} \leq zn$ and proceed to the left and right child accordingly. Finally, reach a leaf node m. For a single result, probability distribution $P_m(C)$ maximizes class label $c \in C$.

4. Handling Uncertain Information

In this type of uncertainty model, a feature value is represented by a pdf, $f_{i;j}$ other than single value, $v_{i;j}$,. For this reason, we assume that $f_{i;j}$ contain nonzero value within a interval $[a_{i,j}, b_{i,j}]$. A $f_{i;j}$ can be programmed logically if it is specified in closed interval. More naturally, it could be implementing numerical value by storing a set of s points $x \in [a_{i,j}, b_{i,j}]$ with the value $f_{i,j}(x)$. Where s is a sample. In fact approximation of pdf $f_{i,j}$ by a discrete distribution with s(no of sample) possible values[1].

An uncertain decision tree model resembles that of the point data model. The only difference is that how the tree is to be employed to classify unseen testing tuples. Like the training tuples, a testing tuple t_0 contains uncertain attribute. So that feature vector of this pdf is written by $(f_{0,1}, \ldots f_{0,k})$. The model of classification is a function M that maps to feature vector (which is discussed above), to a probability distribution P over class lable C. Probabilities calculation is given below part. Here each intermediate tuple written as t_x , weight is written as w_x , where $w_x \in [0,1]$. After that, recursively define quantity is written as $\varphi_n(c; t_x, w_x)$ which is also known as conditional probability with class label c[1].

Each non terminal node n, we have to also include root node, to determine the quantity $\varphi_n(c; t_x, w_x)$, we have to check the A_{jn} and z_n of node n, where A_{jn} is attribute and z_n is split point. Since the pdf of t_x under attribute A_{jn} spans the interval $[a_{x,jn}, b_{x,jn}]$, we compute the probability of left child

$$P_L = \int_{ax,jn}^{Zn} fx, jn(t)dt \text{ . (Or } p_L = 0 \text{ in case } z_n < a_x; \text{ jn)[1]}$$

and the probability of right child" $p_R = 1 - p_L$. Then, we split t_x into two fractional tuples t_L and t_R. Tuples t_L and t_R inherit the class label of t_x as well as the pdfs of t_x for all attributes except A_{jn} . Tuple t_L is assigned a weight of $w_L = w_x * p_L$ and its pdf for A_{jn} is given by

$f_{L,jn}(x) = f_{x,jn}(x)/w_L$, if $x \in [a_{x,jn}, z_n]$
$f_{L,jn}(x) = 0$ Otherwise

Tuple t_R is assigned a weight and pdf analogously [1].

$$\varphi_n(c; t_x, w_x) = p_L * \varphi_{nL}(c; t_L, w_L) + p_R * \varphi_{nR}(c; t_R, w_R)$$

Where n_L = left child of node n.
And n_R = right child of node n.

5. Proposed System

There are basically 3 modes in this algorithm :

- GEN - Generate uncertain data.
- BUILD - Build decision tree.
- BUILDSAVE - Build decision tree and save it in a given tree.

Input: the training dataset D; the set of candidate attributes att-list
Output: An uncertain decision tree.

Begin
1: create a node N;
2: if (D are all of the same class, C) then

3: return N as a leaf node labeled with the class C;
4: else if (attribute-list is empty) then
5: return N as a leaf node labeled with the highest weight class in D;
6: end if ;
7: select a test-attribute with the highest probabilistic information gain ratio to label node N;
8: if (test-attribute is numeric or uncertain numeric) then
9: binary split the data from the selected position y;
10: for (each instance Rj) do
11: if (test-attribute <= y) then
12: put it into D_l with weight Rj .w;
13: else if (test-attribute > y) then
14: put it into D_r with weight Rj .w;
15: else
16: put it into D_l with weight Rj .w * $\int_{x1}^{y} f(x)dx$
17: put it into D_r with weight Rj .w * $\int_{y}^{x2} f(x)dx$
18 end if ;
19: end for;
20: else
21: for (each value a_i (i = 1… n) of the attribute) do
22: grow a branch Di for it;
23: end for;
24: for (each instance R_j) do
25: if (test-attribute is categorical or uncertain categorical) then
26: put it into D_i with R_j .a_i.w * R_j .w;
27: else
28: put it into a certain Di with weight R_j .w;
29: end if
30: end for;
31: end if ;
32: for each Di do
33: attach the node returned by (Di, att-list);
34: end for;
End

Here the steps 1 to 6 are in GEN mode. Steps 7 to 31 are in BUILD mode and steps 32 to 34 are in BUILDSAVE mode.

The tree starts as a single node representing the training samples. If the samples are all of the same class; then the node becomes a leaf. Otherwise, the algorithm uses a probabilistic information gain ratio, as the criteria for selecting the attribute that will best separate the samples into an individual class. This attribute becomes the ”test” attribute at the node. If the test attribute is numerical or uncertain numerical, we split for the data at the selected position y.

A branch is created for test-attribute ≤ y or test-attribute > y respectively.

- If an instance’s test attribute value [x1, x2] is less than or equal to y (x2 ≤ y), it is put into the left branch.
- If an instance’s test attribute value [x1, x2] is larger than y (x1 > y), it is put into the right branch.

If the test attribute is categorical or uncertain categorical, we split the data multiway. The algorithm recursively applies the same process to generate a decision tree for the samples.

The recursive partitioning process stops only when either of the following conditions becomes true:

- All samples for a given node belong to the same class.
- There are no remaining attributes on which the samples may be further partitioned.
-

6. Performance Analysis

In existing system the Glass dataset have accuracy of 0.6649 while proposed system has accuracy of 0.6820. Similarly the Ionosphere dataset have accuracy of 0.8869 while proposed system has accuracy of 0.9260. The Iris dataset have accuracy of 0.9473 while proposed system has accuracy of 0.96. The Breast cancer dataset have accuracy of 0.9352 while proposed system has accuracy of 0.9561. The Vehicle dataset have accuracy of 0.7103 while proposed system has accuracy of 0.7399. Below table show the accuracy of existing system and proposed system with various datasets.

Table 1. Compare result with existing system

Below Figure show the graph of accuracy on various datasets of existing as well as proposed system.

Datasets	Existing System(Accuracy)	Proposed System(Accuracy)
Glass	0.6649	**0.6820**
Ionosphere	0.8869	**0.9260**
Iris	0.9473	**0.9600**
Breast Cancer	0.9352	**0.9561**
Vehicle	0.7103	**0.7399**

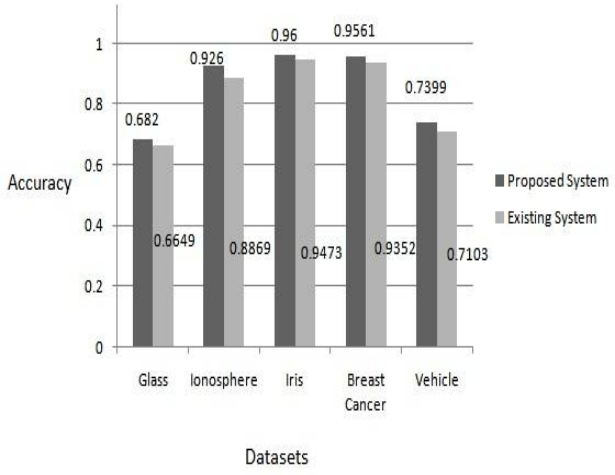

Figure 1. Accuracy on various datasets with two systems

7. Conclusion and Future work

In this paper, I have discussed about how to handle uncertain data in classification problem with decision tree classifier. Traditional decision tree classifiers works with data whose values are precise and known here extend such classifiers so that it can handle data with uncertain information.

In this, the experiment demonstrates that proposed algorithm can process both uncertain numerical data and uncertain categorical data. In experimental results chapter, there are detail discussion about how the algorithm is used for various mode. And finally generate the decision tree for uncertain dataset and also measure the accuracy. And from this we see that the proposed system have greater accuracy then existing ones. Finally we can say that when proper probability distribution functions are being used, data uncertainty leads to decision trees with greater accuracies. In future likewise numerical counterparts, the uncertainty can occur in categorical attributes because of repeated measurements, ambiguities and data staleness.

8. References

[1] S.Tsang,B.Kao,K.Y.Yip, Wai-Shing Ho, and Sau Dan Lee "Decision Tree for Uncertain Data" IEEE Transactions On Knowledge and Data Engineering, Vol. 23, No. 1, pp. 64-78, January 2011.

[2] J.R. Quinlan, "Induction of Decision Trees," Machine Learning, vol. 1, no. 1, pp. 81-106, 1986.

[3] J.R. Quinlan, C4.5: Programs for Machine Learning. Morgan kaufmann, 1993.

[4] C.L. Tsien,I.S. Kohane,and N. McIntosh,"Multiple Signal Integration by Decision Tree Induction to Detect Artifacts in the Neonatal Intensive Care Unit", Artificial Intelligence in Medicine, vol. 19, no. 3, pp. 189-202, 2000.

[5] Jason R. Beck, Maria E. Garcia, Mingyu Zhong, Michael Georgiopoulos, Georgios Anagnostopoulos "A Backward Adjusting Strategy for the C4.5 Decision Tree Classifier" AMALTHEA REU SITE, 2007.

[6] Biao Qin, Yuni Xia, Rakesh Sathyesh, Jiaqi Ge, Sunil Probhakar"DTU:Decision Tree for Uncertain Data" DASFAA 2011, Part II, LNCS 6588, pp. 454–457, 2011.

[7] Matthew N. Anyanwu and Sajjan G. Shiva,"Comparative Analysis Of Serial Decision Tree Classification Algorithms",International Journal of Computer Science and Security,Volume (3):Issue(3).

[8] N.N. Dalvi and D. Suciu,"Efficient Query Evaluation on Probabilistic Databases,"The VLDB J,vol. 16,no. 4,pp. 523-544,2007.

[9] E. Hung,L. Getoor,and V.S. Subrahmanian,"Probabilistic Interval XML,"ACM Trans. Computational Logic,vol. 8,no. 4, 2007.

[10] A.Nierman and H.V. Jagadish,"ProTDB:Probabilistic Data in XML,"Proc. Int'l Conf.Very Large Data Bases(VLDB),pp.646-657, Aug 2002.

[11] J. Chen and R. Cheng, "Efficient Evaluation of Imprecise Location Dependent Queries," Proc. Int'l Conf. Data Eng. (ICDE), pp. 586-595, Apr. 2007.

[12] M. Chau,R. Cheng,B. Kao, and J. Ng,"Uncertain Data Mining: An Example in Clustering Location Data",Proc. Pacific-Asia Conf.Knowledge Discovery and Data Mining (PAKDD),pp. 199-204,Apr.2006.

[13] Jiawei Han and Micheline Kamber "Data Mining Concepts And Techniques" , Morgan kaufman publishers, San Francisco, pp. 285-351, Elsevier, 2011.

[14] A. Asuncion and D. Newman, UCI Machine Learning Repository, http://www.ics.uci.edu/mlearn/MLRepository.html, 2007.