

# DTU: Decision Tree for Uncertain Data

Biao Qin<sup>1</sup>, Yuni Xia<sup>1</sup>, Rakesh Sathyesh<sup>1</sup>, Jiaqi Ge<sup>1</sup>, Sunil Probhakar<sup>2</sup>,

<sup>1</sup> Indiana University Purdue University Indianapolis  
{biaoqin, yxia, sathyesr, jiaqge}@cs.iupui.edu

<sup>2</sup> Purdue University West Lafayette  
sunil@cs.purdue.edu

**Abstract.** This demo presents a decision tree based classification system for uncertain data. Decision tree is a commonly used data classification technique. Tree learning algorithms can generate decision tree models from a training data set. When working on uncertain data or probabilistic data, the learning and prediction algorithms need handle the uncertainty cautiously, or else the decision tree could be unreliable and prediction results may be wrong. In this demo, we will present DTU, a new decision tree based classification and prediction system for uncertain data. This tool uses new measures for constructing, pruning and optimizing decision tree. These new measures are computed considering uncertain data probability distribution functions. Based on the new measures, the optimal splitting attributes and splitting values can be identified and used in the decision tree. We will show in this demo that DTU can process various types of uncertainties and it has satisfactory classification performance even when data is highly uncertain.

## 1 Introduction

Classification is one of the most important data mining techniques. It is used to predict group/class membership for data instances. In many applications, data contains inherent uncertainty. A number of factors contribute to the uncertainty, such as the random nature of the physical data generation and collection process, measurement and decision errors, and data staling. As data uncertainty widely exists, it is important to develop data mining techniques for uncertain and probabilistic data. In this demo, we will show a tool called DTU – Decision Tree for Uncertain Data [1,2], which generates decision-tree based classifier with uncertain data. In DTU, data uncertainty model is incorporated into every step of the tree learning and prediction procedure to achieve higher classification accuracy..

## 2 Demonstration

The DTU tool is implemented based on the open source data mining tool WEKA. We have also extended the Arff Viewer in Weka so that it can display uncertain data in a proper tabular format as shown in table 1. A dataset can have both uncertain numerical attributes and uncertain categorical attributes. Table I shows such an example. Among all the attributes,  $A_i$  is an Uncertain Numerical Attribute(UNA)

whose precise value is unavailable. We only know the range of the  $A_i$  of each tuple.  $A_i$  is an Uncertain Categorical Attribute (UGA). It can be either  $V_1$  or  $V_2$ , each with associated probability.

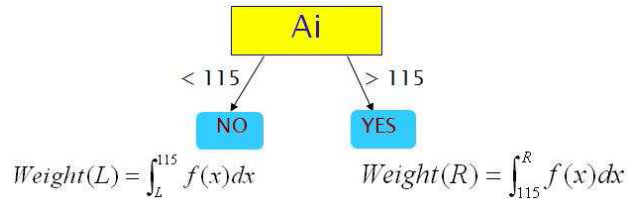
**Table 1.** An uncertain dataset

ID	$A_i$	...	$A_i$	class
1	110-120	...	( $V_1: 0.3; V_2:0.7$ )	No
2	100-120	...	( $V_1: 0.3; V_2:0.7$ )	No
3	60-85	...	( $V_1: 0.3; V_2:0.7$ )	No
4	110-145	...	( $V_1: 0.3; V_2:0.7$ )	No
5	110-120	...	( $V_1: 0.3; V_2:0.7$ )	Yes
6	50-80	...	( $V_1: 0.3; V_2:0.7$ )	No
7	170-250	...	( $V_1: 0.3; V_2:0.7$ )	No
8	85-100	...	( $V_1: 0.3; V_2:0.7$ )	Yes
9	80-100	...	( $V_1: 0.3; V_2:0.7$ )	No
10	120-145	...	( $V_1: 0.3; V_2:0.7$ )	Yes
11	105-125	...	( $V_1: 0.3; V_2:0.7$ )	Yes
12	80-95	...	( $V_1: 0.3; V_2:0.7$ )	No

## 2.1 Decision Tree for Uncertain Data

The core issue in a decision tree induction algorithm is to decide the method of records being split. Each step of the tree-grow process needs to select an attribute test condition to divide the records into smaller subsets. A decision tree algorithm must provide a method for specifying the test condition for different attribute types as well as an objective measure for evaluating the goodness of each test condition.

There are many measures that can be used to determine the best way to split the records. These measures are usually defined in terms of the class distribution of the records before and after splitting. Widely used splitting measures such as information entropy and Gini index are all based on the purity of the nodes and choose the split those results in the highest node purity. These measures are not applicable to uncertain data. For example, for data in Table I, if the splitting condition is  $A_i < 115$ , it cannot be determined whether instances 1, 2, 4, 5, 11 belong to the left or right node. Our approach is that when the cutting point of an attribute lies within the uncertain interval of an instance, the instance is split into both branches and the weights of being in both branches are calculated according to the probability distribution function  $f(x)$ . When the probability distribution function  $f(x)$  is unavailable, we can use domain knowledge to find the appropriate pdf or assume commonly used distribution such as uniform or Gaussian distribution. An example of such a split is shown in figure 1.



**Figure 1.** Uncertain numerical data split

Splitting based on an UCA  $A$  is an  $n$ -ary split, assume attribute  $A$  has  $n$  possible values  $a_i$ , ( $1 \leq i \leq n$ ), If an uncertain categorical attribute is identified as the best splitting attribute, a branch is created for each known value of the test attribute, and the data are partitioned accordingly. For each value  $a_i$  of the attribute, the instance is put into all of the branches with the weight equal to the probability of the attribute to be  $a_i$ , as shown in figure 2.

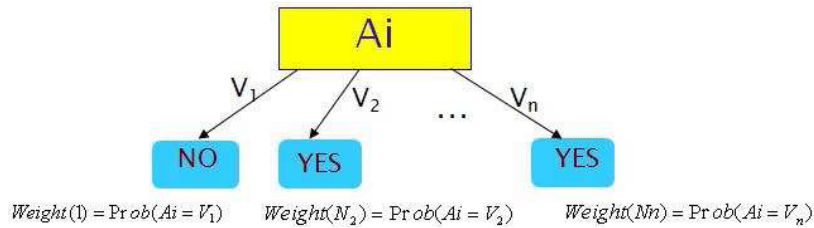


Figure 2. An uncertain categorical Data

Figure 3 shows the uncertain decision tree for an uncertain glass dataset and the decision tree determines the type of glass based on the oxide content. In case an uncertain data instance covers a test point, it is split into both branches according to the cumulative probability distributions; our visualization routine highlights those leaf nodes in red. The leaf nodes indicate the class type of the node  $C_i$ , followed by two real values  $x/y$ .  $x$  is the total probabilistic cardinality of the node, that is, the total number of instance fall in that node, and  $y$  is the number of false positives, that is, the number of instance fall in that node but not in class  $C_i$ . Since both  $x$  and  $y$  are calculated according to probability distribution function, they are floating-point numbers instead of integers, which is different from traditional decision tree. Detailed algorithm for uncertain decision tree can be found in [1, 2].

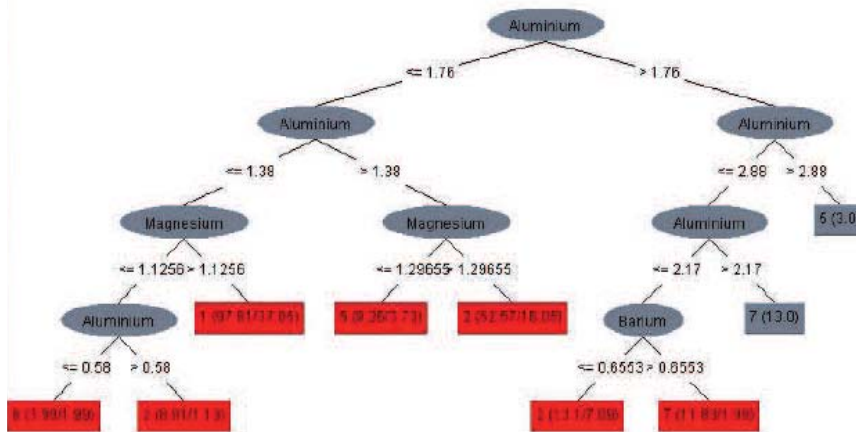


Figure 3. An uncertain decision tree

## 2.2 Comparison with Other Classifiers

In the demo, we will compare the DTU with the traditional decision tree. We will

show the difference in splitting condition selection and tree structures when applied to uncertain data. We will also demonstrate that although DTU takes slightly more time in training, it significantly outperforms the traditional decision tree in accuracy on uncertain data sets.

We will also compare DTU with a rule based uncertain classifier uRule [5]. uRule extracts a set of rules of the form  $R: Condition \Rightarrow y$  based on uncertain data. The rules show the relationships between the attributes of a dataset and the class label, as shown in figure 4. The red shade area highlights all the uncertain classification rules. We will compare the performance of DTU and uRule on various uncertain data sets with different distributions.

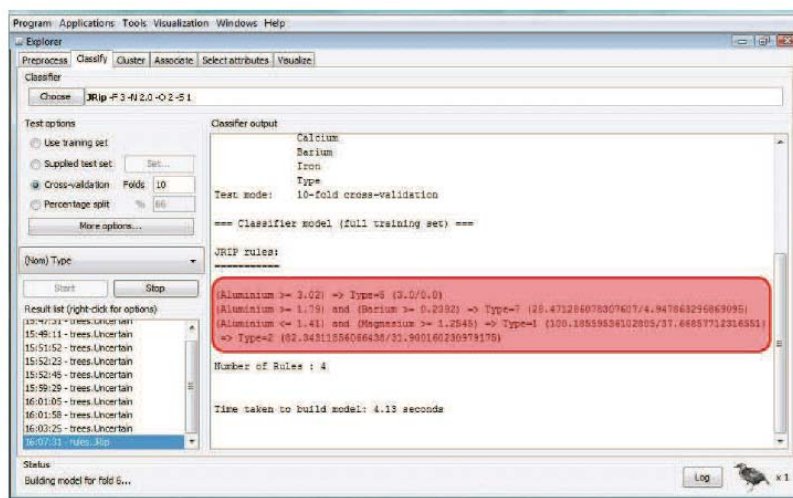


Figure 4: Rule-based Classifier for Uncertain Data

## References

1. B. Qin, Y. Xia, and F. Li, "Dtu: A decision tree for classifying uncertain data," in Proc. the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), 2009, pp. 4–15.
2. S. Tsang, B. Kao, K. Y. Yip, W.-S. Ho, and S. D. Lee, "Decision trees for uncertain data," in ICDE, 2009.
3. S. Singh, C. Mayfield, S. Prabhakar, R. Shah, and S. Hambrusch, "Indexing categorical data with uncertainty," in ICDE, 2007, pp. 616–625.
4. P. Agrawal, O. Benjelloun, A. D. Sarma, C. Hayworth, S. Nabar, T. Sugihara, , and J. Widom, "Trio: A system for data, uncertainty, and lineage," in VLDB, 2006.
5. B. Qin, Y. Xia, S. Prabhakar, and Y. Tu, "A rule-based classification algorithm for uncertain data," in Proc. the IEEE workshop on Management and Mining of Uncertain Data(MOUND), 2009.