Decision Trees for Uncertain Data Mining

Ms.Kiran Dhandore [#1], Dr.Lata Ragha [*2]

[#]Dept. Computer Engineering, Terna Engineering College
Nerul, Navi-Mumbai-400706, MS, India

[1]kdhandore@gmail.com
[2] lata.ragha@gmail.com

***Abstract*— Data uncertainty is an inherent property in various applications due to reasons such as outdated sources, transmission problems or imprecise measurement. When data mining techniques are applied to these data, their uncertainty has to be considered to obtain high quality results. Although much research effort has been directed towards the management of uncertain data in databases, few researchers have addressed the issue of mining uncertain data. We know that with uncertainty, data values are no longer atomic. To apply traditional data mining techniques, uncertain data have to be summarized into atomic values. Unfortunately, the discrepancy in the summarized recorded values and the actual values could seriously affect the quality of the mining results. The Decision tree is a widely used data classification technique. Traditional decision tree classifiers work with data whose values are known and precise. We propose a methodology to handle uncertain data using decision tree. The utility and robustness of the proposed algorithm and its prediction accuracy is discussed here. The efficiency of the algorithm can be verified based on execution time, pruning effectiveness etc.**

Keywords— **Decision Trees, Uncertain Data, Classification, Data Mining, Entropy**

I. INTRODUCTION

In computer science, uncertain data is the notion of data that contains specific uncertainty. Uncertain data is typically found in the area of sensor networks. When representing such data in a database, some indication of the probability of the various values. Data is often associated with uncertainty because of measurement inaccuracy, sampling discrepancy, outdated data sources, or other errors. In recent years, uncertain data have become ubiquitous because of new technologies for collecting data which can only measure and collect the data in an imprecise way. While many applications lead to data which contains errors, we refer to uncertain data sets as those in which the level of uncertainty can be quantified in some way. Some examples of applications which create uncertain data are as follows: [1]

- Many scientific measurement techniques are inherently imprecise. In such cases, the level of uncertainty may be derived from the errors in the underlying instrumentation.
- Many new hardware technologies such as sensors generate data which is imprecise. In such cases, the error in the sensor network readings can be modeled, and the resulting data can be modeled as imprecise data.
- In many applications such as the tracking of mobile objects, the future trajectory of the objects is modeled by forecasting techniques. Small errors in current readings can get magnified over the forecast into the distant future of the trajectory. This is frequently encountered in cosmological applications when one models the probability of encounters with Near-Earth-Objects (NEOs). Errors in forecasting are also encountered in non-spatial applications such as electronic commerce.
- Location-based services: in the scenario of moving objects (such as vehicles or people), it is impossible for the database to track the exact locations of all objects at all time instants. Therefore, the location of each object is associated with uncertainty between updates.

In recent years, there has been much research on the management of uncertain data in databases, such as the representation of uncertainty in databases and querying data with uncertainty. However, little research work has addressed the issue of mining uncertain data. We know that with uncertainty, data values are no longer atomic. To apply traditional data mining techniques, uncertain data have to be summarized into atomic values. Taking moving-object applications as an example again, the location of an object can be summarized either by its last recorded location or by an expected location (if the probability distribution of an object's location is taken into account). Unfortunately, the discrepancy in the summarized recorded values and the actual values could seriously affect the quality of the mining results.

Figure 1.1 illustrates this problem when a clustering algorithm is applied to moving objects with location uncertainty. Figure 1.1(a) shows the actual locations of a set of objects, and Figure 1.1(b) shows the recorded location of these objects, which are already outdated. The clusters

obtained from these outdated values could be significantly different from those obtained as if the actual locations where available (Figure 1.1(b)). If we solely rely on the recorded values, many objects could possibly be put into wrong clusters. Even worse, each member of a cluster would change the cluster centroids, thus resulting in more errors. By incorporating uncertainty information, such as the probability density functions (pdf) of uncertain data, into existing data mining methods, the mining results could resemble closer to the results obtained as if actual data were available and used in the mining process (Figure 1.1 (c)).

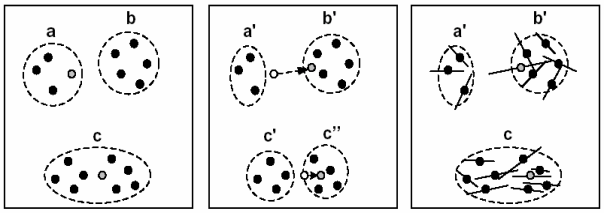

Figure 1 (a) The real-world data is partitioned into three clusters (a, b, c). (b) The recorded locations of some objects (shaded) are not the same as their true location, thus creating clusters a', b', c' and c''. Note that a' has one less object than a, and b' has one more object than b. Also, c is mistakenly split into c' and c''. (c) Line uncertainty is considered to produce clusters a', b' and c. The clustering result is closer to that of (a) than (b).

1.1 Decision Tree

A decision tree is a flowchart-like tree structure, where each internal node (non leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. In the simplest case, each test considers a single attribute, such that the instance space is partitioned according to the attribute's value. In case of numeric attributes, the condition refers to a range. Each leaf is assigned to one class representing the most appropriate target value. Alternatively, the leaf may hold a probability vector indicating the probability of the target attribute having a certain value.

Decision trees can easily be converted to classification rules. The construction of decision tree classifiers does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery. Decision trees can handle high dimensional data. Their representation of acquired knowledge in tree form is intuitive and generally easy to assimilate by humans. The learning and classification steps of decision tree induction are simple and fast. In general, decision tree classifiers have good accuracy. However, successful use may depend on the data at hand.

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. 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, unreliable data transmission and data staling. Decision tree induction algorithms have been used for classification in many application areas, such as medicine, manufacturing and production, financial analysis, astronomy, and molecular biology.

II. RELATED WORKS

Classification is a well-studied area in data mining. Many classification algorithms have been proposed in the literature, such as decision tree classifiers [2]. In spite of the numerous classification algorithms, building classification based on uncertain data has remained a great challenge. There are early works performed on developing decision trees when data contains missing or noisy values. Various strategies have been developed to predict or fill missing attribute values [3]. There is also some previous work performed on classifying uncertain data in various applications [4]. The methods try to solve specific classification tasks instead of developing a general algorithm for classifying uncertain data.

Recently, more research has been conducted in uncertain data mining. Most of them focus on clustering uncertain data [5] [6]. The key idea is that when computing the distance between two uncertain objects, the probability distributions of objects are used to compute the expected distance. Xia et al. [7] Introduce a new conceptual clustering algorithm for uncertain categorical data**.** Agarwal [8] proposes density based transforms for uncertain data mining. There is also some research on identifying frequent item sets and association mining [9] from uncertain datasets. The support of item sets and confidence of association rules are integrated with the existential probability of transactions and items. Burdicks [10] discuss OLAP computation on uncertain data. None of them address the issue of developing a general classification and prediction algorithm for uncertain data.

Decision trees are one of the most important aspects for "Decision-making". Classification is one of the most widespread data mining problems found in real life. Decision tree classification is one of the best-known solution approaches. There has been a growing interest in uncertain data mining. In [11], the well-known k-means clustering algorithm is extended to the UK-means algorithm for clustering uncertain data. Data uncertainty is usually captured by pdf's, which are generally represented by sets of sample values. Mining uncertain data is therefore computationally costly due to information explosion for sets of samples vs. single values. To improve the performance of UK-means, pruning techniques have been proposed. Examples include min-max dist pruning [12] and CK-means [13].
In C4.5 [1] and probabilistic decision trees [14], missing values in training data are handled by using fractional tuples. During testing, each missing value is replaced by multiple values with probabilities based on the training tuples, thus

allowing probabilistic classification results. They have adopted the technique of fractional tuple for splitting tuples into subsets when the domain of its pdf spans across the split point.

III. EXISTING SYSTEM

In traditional decision-tree classification, a feature or an attribute of a tuple is either categorical or numerical. For the latter, a precise and definite point value is usually assumed. In many applications, however, data uncertainty is common. The value of a feature/attribute is thus best captured not by a single point value, but by a range of values giving rise to a probability distribution. Although the previous techniques can improve the efficiency of means, they do not consider the spatial relationship among cluster representatives, nor make use of the proximity between groups of uncertain objects to perform pruning in batch. A simple way to handle data uncertainty is to abstract probability distributions by summary statistics such as means and variances called averaging approach. Another approach is to consider the complete information carried by the probability distributions to build a decision tree called Distribution-based approach.

IV. PROPOSED SYSTEM

We construct a decision tree classifier on data with uncertain numerical attributes. The main goals are:

(1) By using the basic algorithm to construct decision trees out of uncertain datasets.

(2) Find out whether the Distribution-based approach could lead to higher classification accuracy compared with the Averaging approach.

(3) Establish a theoretical foundation on which pruning techniques are derived that can significantly improve the computational efficiency of the Distribution-based algorithms.

The following Figure 2 shows the flow diagram of the proposed methodology.

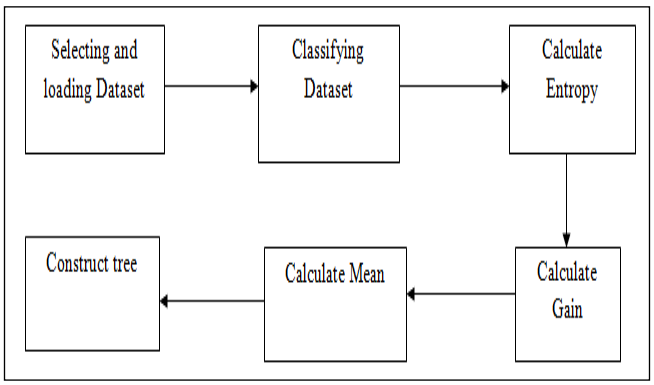

Figure 2: Proposed Methodology

The two approaches for handling uncertain data. The first approach, called "Averaging", transforms an uncertain dataset to a point-valued one by replacing each pdf with its mean value. To exploit the full information carried by the pdf's, the second approach, called "Distribution- based", considers all the sample points that constitute each pdf.

A. Averaging

A simple way to handle data uncertainty is to abstract probability distributions by summary statistics such as means and variances which is called as the averaging approach. A straightforward way to deal with the uncertain information is to replace each pdf with its expected value, thus effectively converting the data tuples to point-valued tuples. The algorithm starts with the root node and with S being the set of all training tuples. At each node n, we first check if all the tuples in S have the same class label.

B. Distribution Based

An approach is to consider the complete information carried by the probability distributions to build a decision tree which is called as Distribution-based approach. After an attribute Ajn and a split point zn has been chosen for a node n, we split the set of tuples S into two subsets L and R. The major difference from the point-data case lies in the way the set S is split. If the pdf properly contains the split point, i.e., ai,jn $\leqslant$ zn < bi,jn, we split ti into two fractional tuples[3] tL and tR and add them to L and R, respectively. The algorithm is called UDT (Uncertain Decision Tree).

V. CONCLUSION

We propose a new decision tree algorithm for classifying and predicting uncertain data. We extend the measures used in traditional decision trees, such as information entropy and information gain, for handling data uncertainty. The decision tree for uncertain data can process both uncertain numerical data and uncertain categorical data. It can achieve satisfactory classification and prediction accuracy even when data are highly uncertain. Our proposed method handles the uncertain data through "Averaging" by using means and variances. But in "Distribution-based" the accuracy of an uncertain data is detected through decision trees. Decision trees, calculate the entropy measure and enhance the information gain for better accuracy. Several procedures and algorithm handles data uncertainty. We expect higher accuracy for uncertain data using decision trees.

VI. ACKNOWLEDGMENT

We express our sincere thanks to my Project Guide
Dr. Lata Ragha HOD of Computer Department for her guidance and supervision, assisting with all kinds of support and inspiration, excellent guidance and valuable suggestions throughout this investigation and preparation of this project.

REFERENCES

[1] Michael Chau, Reynold Cheng, and Ben Kao" Uncertain Data Mining: A New Research Direction", published in Proceedings of the 10th Pacific-Asia conference on Advances in Knowledge discovery and Data Mining.

[2] Varsha Choudhary, Pranita Jain, "Classification: A Decision Tree for Uncertain Data Using CDF" International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 Vol. 3, Issue 1, January -February 2013, pp.1501-1506

[3] Hawarah L, Simonet A, Simonet M "Dealing with Missing Values in a Probabilistic Decision Tree during Classification", The Second Intern8ational Workshop on Mining Complex Data, pp. 325-329.

[4] Bi J, Zhang T (2004) Support Vector Classification with Input Data Uncertainty, Advances in Neural Information Processing Systems 17: 161-168.

[5] Ngai WK, Kao B, Chui CK, Cheng R, Chau M, Yip KY (2006) Efficient Clustering of Uncertain Data, In: Proceedings of IEEE International Conference on Data Mining'06, pp. 436-445.

[6] W. Street, W. Wolberg, and O. Mangasarian, "Nuclear feature extraction for breast tumor diagnosis," in SPIE, vol. 1905, San Jose, CA, U.S.A., 1993, pp. 861–870. [Online]. Available: http: //citeseer.ist.psu.edu/street93nuclear.html

[7] Xia Y, Xi B (2007) Conceptual clustering categorical data with uncertainty, In: Proceedings of international conference on tools with artificial intelligence, pp. 329-336.

[8] Aggarwal C "On density based transforms for uncertain data mining". In IEEE Conference on Data Mining, (2007) pp. 866-875.

[9] Chui C, Kao B, Hung E (2007) Mining Frequent Itemsets from Uncertain Data, In: Proceedings of the PAKDD'07, pp. 47-58.

[10] M. Chau, R. Cheng, B. Kao, and J. Ng, "Uncertain data mining: An example in clustering location data," in PAKDD, ser. Lecture Notes in Computer Science, vol. 3918. Singapore: Springer, 9–12 Apr. 2006, pp. 199–204.

[11] W. K. Ngai, B. Kao, C. K. Chui, R. Cheng, M. Chau, and K. Y. Yip, "Efficient clustering of uncertain data," in IEEE International Conference on Data Mining. Hong Kong, China: IEEE Computer Society, 18–22 Dec. 2006, pp. 436–445.