International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 11, November 2015
Available at http://internationaljournalofresearch.org

Extend Classical Decision Tree Building Algorithms to Handle Data with Uncertain Numerical Attributes using the Distribution-Based Approach

Shabanaunnisa Begum
Assistant Professor, Department of Computer Science, Malla Reddy Engineering College for Women, Maisammaguda, Hyderabad.
S. Venkata Ramana
Assistant Professor, Department of Computer Science, Malla Reddy Engineering College for Women, Maisammaguda, Hyderabad.

Abstract-
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. Example sources of uncertainty include measurement/quantization errors, data staleness, and multiple repeated measurements. With uncertainty, the value of a data item is often represented not by one single value, but by multiple values forming a probability distribution. Rather than abstracting uncertain data by statistical derivatives (such as mean and median), we discover that the accuracy of a decision tree classifier can be much improved if the "complete information" of a data item (taking into account the probability density function (pdf)) is utilized. We extend classical decision tree building algorithms to handle data tuples with uncertain values. Extensive experiments have been conducted that show that the resulting classifiers are more accurate than those using value averages. Since processing pdf's is computationally more costly than processing single values (e.g., averages), decision tree construction on uncertain data is more CPU demanding than that for certain data. To tackle this problem, we propose a series of pruning techniques that can greatly improve construction efficiency.
Keywords **–** Uncertain Data; Decision Tree; Classification; Data Mining

INTRODUCTION

Classification is a classical problem in machine learning and data mining. Given a set of training data tuples, each having a class label and being represented by a feature vector, the task is to algorithmically build a model that predicts the class label of an unseen test tuple based on the tuple's feature vector. [2]One of the most popular classification models is the decision tree model. Decision trees are popular because they are practical and easy to understand. Rules can also be extracted from decision trees easily. Many algorithms have been devised for decision tree construction. These algorithms are widely adopted and used in a wide range of applications such as image recognition, medical diagnosis, credit rating of loan applicants, scientific tests, fraud detection, and target marketing. Data uncertainty arises naturally in many applications due to arious reasons. We briefly discuss three categories here:

a) Measurement Errors: Data obtained from measurements by physical devices are often imprecise due to measurement errors. As an example, a tympanic (ear) thermometer measures body temperature by measuring the temperature of the ear drum via an infrared sensor. A typical ear thermometer has a quoted calibration error of +/- 0:20C, which is about 6.7% of the normal range of operation, noting that the human body temperature ranges from 370C (normal) and to 400C (severe fever) that along with other factors such as placement and technique, measurement error can be very high. For example, it is reported in [5] that about 24% of measurements are off by more than 0.50C, or about 17% of the operational range. Another source of error is quantisation errors introduced by the digitisation process. Such errors can be properly handled by assuming an appropriate error model, such as a Gaussian error distribution for random noise or a uniform error distribution for quantisation errors.

b) Data Staleness: In some applications, data values are continuously changing and recorded information is always stale. One example is location-based tracking system. The where about of a mobile device can only be approximated by imposing an uncertainty model on its last reported location[6]. typical uncertainty model requires knowledge about the moving speed of the device and whether its movement is restricted (such as a car moving on a road network) or unrestricted (such as an animal moving on plains). Typically a 2D probability density function is defined over a bounded region to model such uncertainty.

International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 11, November 2015
Available at http://internationaljournalofresearch.org

c) Repeated Measurements: Perhaps the most common source of uncertainty comes from repeated measurements. For example, a patient's body temperature could be taken multiple times during a day; an anemometer could record wind speed once every minute; the space shuttle has a large number of heat sensors installed all over its surface. When we inquire about a patient's temperature, or wind speed, or the temperature of a certain section of the shuttle, which values shall we use? Or, would it be better to utilize all the information by consider in the distribution given by the collected data values? As a more elaborate example, consider the "Breast Cancer" dataset reported in [7]. This dataset contains a number of tuples. Each tuple corresponds to a microscopic image of stained cell nuclei.

A typical image contains 10–40 nuclei. One of the features extracted from each image is the average radius of nuclei. We remark that such a radius measure contains a few sources of uncertainty: (1) an average is taken from a large number of nuclei from an image, (2) the radius of an (irregularly-shaped) nucleus is obtained by averaging the length of the radial line segments defined by the centroid of the nucleus and a large number of sample points on the nucleus' perimeter, and (3) a nucleus' perimeter was outlined by a user over a fuzzy 2D image. From (1) and (2), we see that a radius is computed from a large number of measurements with a wide range of values. The source data points thus form interesting distributions. From (3), the fuzziness of the 2D image can be modelled by allowing a radius measure be represented by a range instead of a concrete point-value. Yet another source of uncertainty comes from the limitation of the data collection process. For example, a survey may ask a question like, "How many hours of TV do you watch each week?" A typical respondent would not reply with an exact precise answer. Rather, a range (e.g., "6–8 hours") is usually replied, possibly because the respondent is not so sure about the answer himself. In this example, the survey can restrict an answer to fall into a few pre-set categories (such as "2–4 hours", "4–7 hours", etc.). However, this restriction unnecessarily limits the respondents' choices and adds noise to the data. Clustering of uncertain data has recently attracted interests from researchers. This is driven by the need of applying clustering techniques to data that are uncertain in nature, and a lack of clustering algorithms that can cope with the uncertainty. Uncertainty in data arises naturally due to random errors in physical measurements, data staling, as well as defects in the data collection models. For instance, when track locations with GPS devices, the reported location can have errors of a few meters. When attempting to cluster the location of objects tracked using GPS, the errors may affect the clustering result. Traditional clustering approaches model objects as having accurately known positions. This model does not cope well with uncertain data. It does not take into account the uncertainty inherent in the data, and may lead to undesirable clustering results because information on the uncertainty is dropped.[1] Owing to this shortcoming as well as the practical need to deal with data with uncertainty, there has been growing interest in developing problem models and algorithms to handle uncertain data. Rather than a single point in space, the location of each object is represented by a probability density function (pdf) over the space Rm being studied. Given a set of such objects, we want to divide them into k clusters, minimizing the total expected distance (ED) from the objects to their cluster centers. We focus on the case where ED is defined using MSE (mean squared error).

LITERATURE SURVEY

There has been a growing interest in uncertain data management in recent years. Data uncertainty has been classified into three types

1) Existential or tuple uncertainty
2) Value uncertainty or attribute uncertainty
3) Co-related uncertainty.

When uncertain object or the uncertain data tuples exists then existential uncertainty appears. It is also arises due to uncertain a feature of a data objects. Data uncertainty is a "probabilistic database", in that each data tuples is situated with a probability value which contains the confidence of its presence [6]. Value uncertainty appears when a tuple is known to exist, but the tuple value is not known precisely. In data item value uncertainty is usually represented by a PDF over a finite and the bounded region of possible values [12]. "Imprecise queries processing" is one well known topic on the value uncertainty. Such a query is associated with a probability that represent the guarantee on its correctness.

In co-related uncertainty value of multiple attributes describe by a joint- probability- distribution Significant research has been interested in uncertain data mining [8] in recent years. Goal of the data mining process is to extract information from data set and transform this data into an understandable format for further use. For the uncertain data mining, extend traditional data mining algorithm in such a way that the extended data mining algorithm are applicable for uncertain data. The extension process contains two main steps to allow traditional data mining algorithm computationally feasible for uncertain data [9].

To convert the traditional data mining algorithm to make it theoretically workable for uncertain data is the first step. For clustering uncertain data well-known k-means clustering algorithm is extended to the uk-means algorithm [10]. In traditional clustering algorithm, data is generally represented by points in the space. To improve the I/O time and computational efficiency of the modified algorithm is the second step. In clustering, data uncertainty is generally

International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 11, November 2015
Available at http://internationaljournalofresearch.org

captured by PDF and usually represented by sets of sample values. The uncertain data mining is therefore computationally costly due to the information explosion. To improve the efficiency of K-means and UK-means series of pruning technique have been proposed. Examples include ck-means [11] and min-max-dist. pruning. The decision tree classification has been addressed for several decades.

A series of pruning technique are introduced to deal with the problem of over fitting the data. Pruned decision tree are faster in classifying unseen test tuples because pruned decision tree are smaller than unprimed decision tree. There are two techniques of pruning a decision tree namely – pre-pruning and post-pruning. Preprinting technique stops the construction of decision tree earlier. On other hand post-pruning technique removes branches from fully constructed decision tree. Classification of uncertain data has been studied for decades in the form of missing values. Missing value appears when some attribute value is not available during data collection process or due to data entry error. For handing missing data methods are divided into three category containing, ignoring or discarding data, parameter estimation and imputation. Missing value affect the accuracy of the classifier, so that proper handling of it is important. One of the ways to handle the missing value is to ignore the tuple. Another way is to use the pattern of other data tuples to approximate missing values. Simplest way is to use the majority values or most commonly known values to approximate missing values [4].

EXISTING SYSTEM

In traditional decision-tree classification, a feature (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. We call this approach Averaging. Another approach is to consider the complete information carried by the probability distributions to build a decision tree. We call this approach Distribution-based.

RELATED WORK

a. Uncertain Data Classification:-

Numerous uncertain data classification algorithms have been proposed in the literature in recent years. Qin et al. (2009) proposed a rule-based classification algorithm for uncertain data. Rental. (2009) proposed to apply Naive Bays approach to uncertain data classification problem. Decision tree model is one of the most popular classification models because of its advantages (Tsang et al. 2009, Quinlan 1993). Several decision tree based classifiers for uncertain data are proposed by research community. The wellknown C4.5 classification algorithm was extended to the DTU (Qin et al. 2009a) and the UDT (Tsang et al. 2009) for classifying uncertain data. (Qin et al. 2009a) used probability vector and probability density function (pdf) to represent uncertain categorical attribute (Qin et al. 2009b) and uncertain numerical attribute (Cheng et al. 2003) respectively. They constructed a well performance decision tree for uncertain data (DTU). Tsang et al. (2009) used the ―complete information‖ of pdf to construct a uncertain decision tree(UDT) and proposed a series of pruning techniques to improve the efficiency. Both DTU and UDT algorithm are extensions of C4.5, they can only be used to deal with uncertain static data set, while our UCVFDT algorithm is capable of handling uncertain data stream, in which huge volume of data arrive at high speed. 210 Decision Tree for Dynamic and Uncertain Data Streams .

b. Data Stream Classification:

There are two main approaches for classifying data streams: single classifier based approach and ensemble based approach. For single classifier based approach, the most well-known classifiers are VFDT and CVFDT. The CVFDT improves the VFDT with ability to deal with concept drift. After that, many decision tree based algorithms were proposed for data stream classification (for example, Gametal. 2005, Gametal. 2006). For ensemble based approaches, the initial papers use static majority voting (for example, Street and Kim 2001, Wang et al. 2003), while the current trend is to use dynamic classifier ensembles (for example, Zhang and Jin 2006, Zhu et al. 2004). All of algorithms mentioned above can only handle certain data. Pan et al. (2009) proposed two types of ensemble based algorithms, Static Classifier Ensemble (SCE) and Dynamic Classifier Ensemble (DCE) for mining uncertain data streams. To the best of our knowledge, this is the only work devoted to classification of uncertain data streams. However, in (Pan et al. 2009), class value of a sample is assumed to be uncertain, while attributes is assumed to have precise value. Our UCVFDT

International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 11, November 2015
Available at http://internationaljournalofresearch.org

handles uncertain attributes, while class value is assumed to be precise.

c. Construction:-

Some premises guide this algorithm, such as the following

• If all cases are of the same class, the tree is a leaf and so the leaf is returned labeled with this class;

• For each attribute, calculate the potential information provided by a test on the attribute (based on the probabilities of each case having a particular value for the attribute). Also calculate the gain in information that would result from a test on the attribute (based on the probabilities of each case with a particular value for the attribute being of a particular class);

PROPOSED SYSTEM

We study the problem of constructing decision tree classifiers on data with uncertain numerical attributes. Our goals are

(1) To devise an algorithm for building decision trees from uncertain data using the Distribution-based approach;

(2) To investigate whether the Distribution-based approach could lead to a higher classification accuracy compared with the Averaging approach; and

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

IMPLEMENTATION

Implementation is the stage of the project when the theoretical design is turned out into a working system. Thus it can be considered to be the most critical stage in achieving a successful new system and in giving the user, confidence that the new system will work and be effective.

The implementation stage involves careful planning, investigation of the existing system and it's constraints on implementation, designing of methods to achieve changeover and evaluation of changeover methods.

Data Insertion:-

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. With uncertainty, the value of a data item is often represented not by one single value, but by multiple values forming a probability distribution. This uncertain data is inserted by user.

Generate Tree:-

Building a decision tree on tuples with numerical, point valued data is computationally demanding. A numerical attribute usually has a possibly infinite domain of real or integral numbers, inducing a large search space for the best "split point". Given a set of n training tuples with a numerical attribute, there are as many as n-1 binary split points or ways to partition the set of tuples into two non-empty groups. Finding the best split point is thus computationally expensive. To improve efficiency, many techniques have been proposed to reduce the number of candidate split points

Averaging:-

A simple way to handle data uncertainty is to abstract probability distributions by summary statistics such as means and variances. We call this approach Averaging. A straight-forward 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. This reduces the problem back to that for point-valued data. AVG is a greedy algorithm that builds a tree top-down. When processing a node, we examine a set of tuples S. 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.

Distribution Base:-

An approach is to consider the complete information carried by the probability distributions to build a decision tree. We call this approach Distribution-based. Our goals are,

(1) To devise an algorithm for building decision trees from uncertain data using the Distribution-based approach;

(2) To investigate whether the Distribution-based approach could lead to a higher classification accuracy compared with the Averaging approach;

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

RESULTS AND DISCUSSION

The accuracy of a decision tree classifier can be much improved if the "complete information" of a data item (taking into account the probability density function (pdf)) is utilised. Distribution based algorithm can improve classification accuracy because there are more choices of split points. The distribution approach has to examine k (ms-

1) split points whereas the AVG approach has to examine k(m-1) split points. Entropy calculations are the most computation intensive part of UDT. To explore the potential of achieving a higher classification accuracy by considering data uncertainty, we have implemented AVG and UDT and applied them to 04 datasets namely glass dataset, page block , Japanese Vowel, Breast Cancer dataset taken from the UCI Machine Learning Repository. These datasets are chosen because they contain mostly numerical attributes obtained from measurements. We model uncertainty information by fitting appropriate error models on to the point data. For each tuple ti and for each attribute Aj , the point value vi;j reported in a dataset is used as the mean of a pdf fi;j , defined over an interval [ai;j ; bi;j]. The range of values for Aj (over the whole data set) is noted and the width of [ai;j ; bi;j] is set to w _ jAj j, where jAj j denotes the width of the range for Aj and w is a controlled parameter. To generate the pdf fi;j , we consider two options. The first is uniform distribution, which implies fi;j(x) = (bi;j - ai;j)-1. The other option is Gaussian distribution, for which we use 1/4 (bi;j -ai;j) as the standard deviation. In both cases, the pdf is generated using s sample points in the interval. Using this method (with controllable parameters w and s, and a choice of Gaussian vs. uniform distribution), we transform a data set with point values into one with uncertainty. The reason that we choose Gaussian distribution and uniform distribution is that most physical measures involve random noise which follows Gaussian distribution, and that digitisation of the measured values introduces quantisation noise that is best described by a uniform distribution.

CONCLUSION

We have extended the model of decision-tree classification to accommodate data tuples having numerical attributes with uncertainty described by arbitrary pdf's. We have modified classical decision tree building algorithms to build decision trees for classifying such data. We have found empirically that when suitable pdf's are used, exploiting data uncertainty leads to decision trees with remarkably higher accuracies. We therefore advocate that data be collected and stored with the pdf information intact. Performance is an issue, though, because of the increased amount of information to be processed, as well as the more complicated entropy computations involved.

Therefore, we have devised a series of pruning techniques to improve tree construction efficiency. Our algorithms have been experimentally verified to be highly effective. Their execution times are of an order of magnitude comparable to classical algorithms. Some of these pruning techniques are generalisations of analogous techniques for handling point-valued data. Other techniques, namely pruning by bounding and end-point sampling are novel. Although our novel techniques are primarily designed to handle uncertain data, they are also useful for building decision trees using classical algorithms when there are tremendous amounts of data tuples.

REFERENCES

[1] R. Agrawal, T. Imielinski, and A. N. Swami, "Database mining: A performance perspective," IEEE Trans. Knowl. Data Eng., vol. 5, no. 6, pp. 914–925, 1993.

[2] J. R. Quinlan, "Induction of decision trees," Machine Learning, vol. 1, no. 1, pp. 81–106, 1986.

[3]C4.5: Programs for Machine Learning. Morgan Kaufmann, 1993, ISBN 1-55860-238-0.

[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] G. L. Freed and J. K. Fraley, "25% "error rate" in ear temperature sensing device," Pediatrics, vol. 87, no. 3, pp. 414–415, Mar. 1991.

[6] O. Wolfson and H. Yin, "Accuracy and resource consumption in tracking and location prediction," in SSTD, ser. Lecture Notes in Computer Science, vol. 2750. Santorini Island, Greece: Springer, 24-27 Jul. 2003, pp. 325–343.

[7] 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

[8] N. N. Dalvi and D. Suciu, "Efficient query evaluation on probabilisticdatabases," The VLDB Journal, vol. 16, no. 4, pp. 523–544, 2007.

[9] E. Hung, L. Getoor, and V. S. Subrahmanian, "Probabilistic intervalXML," ACM Transactions on Computational Logic (TOCL), vol. 8,no. 4, 2007.

[10] A. Nierman and H. V. Jagadish, "ProTDB: Probabilistic data in XML," in VLDB. Hong Kong, China: Morgan Kaufmann, 20-23 Aug. 2002, pp. 646–657.