Praise Worthy Prize

International Review on Computers and Software (IRECOS)

A Modified Decision Tree Algorithm for Uncertain Data Classification

S. Meenakshi[1*], V. Venkatachalam[2]

(*) *Corresponding author*

Authors' affiliations

DOI's assignment:
the author of the article can submit here a request for assignment of a DOI number to this resource!
Cost of the service: euros 10,00 (for a DOI)

Abstract

The classifications of uncertain data become one of the tedious processes in the data mining domain. The uncertain data are contains tuples with different probability distribution and thus to find similar class of tuples is a complex process. When we consider uncertain data, the feature vector will not be a single valued but a function. Recently, different methods are proposed on decision tree based uncertain data classification with binary based operation on the decision tree. When multiclass data are given to the decision tree, their algorithm has to give repeated calculation to produce the probability distribution matching the class labels, thus time and memory utilization will be high for the particular algorithm. In this paper, we have intended to propose a classification method for uncertain data based on the decision tree. The proposed approach concentrates on an adaptive averaging method, where we have incorporated mean and median of the tuple to produce the feature value that will be used in the decision tree for decision making. Then a probability calculation is executed to find the relevance of tuple with respect to a class. If the calculated probability value is similar to a particular probability distribution, then the tuple is marked to that particular class. Thus, we produce a decision tree with c number of leaf nodes, where c is the number of class labels in the training

Cropping Based Image Captcha Technology
K. Suresh Kumar et al.
3910 views since: 2014-01-31

Seamless and Secure Design for Subsequent Handover in Mobile WiMAX Networks
H. F. Zmezm et al.
2708 views since: 2014-08-31

An Efficient Approach for Cancer Prediction Using Genomic Signal Processing
T. Inbamalar et al.
2225 views since: 2014-03-31

Texture Pattern Based Lung Nodule Detection (TPLND) Technique in CT Images
T. Kumar et al.
2210 views since: 2014-03-31

Survey and Analysis of Visual Secret Sharing Techniques
L. Anbarasi et al.
2124 views since: 2014-09-30

data. The test data is subjected to the trained decision tree to obtain the classified data. . The experimental analysis are conducted for evaluating the performance of the proposed approach. The vehicle dataset and segment dataset from the UCI data repository is selected for the performance analysis. The results from the experimental analysis showed that the adaptive method has achieved a maximum average accuracy of 0.997 while the existing approach achieved only 0.985

Keywords

Uncertain Data; Probability Distribution; Decision Tree Algorithm; Classification

Full Text:

PDF

References

Smith Tsang, Ben Kao, Kevin Y. Yip, Wai-Shing Ho, Sau Dan Lee, "Decision Trees for Uncertain Data, TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, PP: 1-

JiangtaoRen, Sau Dan Lee, Xianlu Chen, Ben Kao, Reynold Cheng and David Cheung,"Naive Bayes Classification of Uncertain Data, "Ninth IEEE International Conference on Data Mining,pp.944-949,2009.

Fayyad U., Piatetsky-Shapiro G., Smyth P., From data mining to knowledge discovery: an overview, Advances in knowledge discovery and data mining, American Association for Artificial Intelligence, Menlo Park, CA, AAAI/MIT Press, 1996, pp: 1-36.

Romero C., Ventura S., Espejo P.G., and Hervas C., Data Mining Algorithms to Classify Students, proceedings of the 1st Int'l conference on educational data mining, Canada, 2008, pp: 8-17.

Zhang J., Mani I., kNN Approach to Unbalanced Data Distributions: A Case Study involving Information Extraction, In Proceedings of The Twentieth International Conference on Machine Learning (ICML-2003), Workshop on Learning from Imbalanced Data Sets II, August 21, 2003.

Charu C. Aggarwal and Philip S. Yu, "A Survey of Uncertain Data Algorithmsand Applications", IEEE transactions on knowledge and data engineering, Vol. 21, No. 5, MAY 2009.

Barbara, D., Garcia-Molina, H. and Porter, D. "The Management of Probabilistic Data," IEEE Transactionson Knowledge and Data Engineering, 4(5), 1992.

Cheng, R., Kalashnikov, D., and Prabhakar, S. "Evaluating Probabilistic Queries over Imprecise Data,"Proceedings of the ACM SIGMOD International Conference on Management of Data, June 2003.

Chau, M., Cheng, R., and Kao, B., "Uncertain Data Mining: A New Research Direction," in Proceedings of the Workshop on the Sciences of the Artificial, Hualien, Taiwan, December 7-8, 2005.

J. R. Quinlan, C4.5: Programs for Machine Learning. Morgan KaufmanPublishers, 1993.

W. W. Cohen, "Fast effective rule induction," in Proc. of the 12th Intl.Conf. on Machine Learning, 1995, pp. 115–123.

P. Langley, W. Iba, and K. Thompson, "An analysis of bayesianclassifiers," in National Conf. on Artigicial Intelligence, 1992, pp. 223-228.

A Modified Decision Tree Algorithm for Uncertain Data Classification

S.Meenakshi[1], V.Venkatachalam[2]

Abstract-- *The classifications of uncertain data become one of the tedious processes in the data mining domain. The uncertain data are contains tuples with different probability distribution and thus to find similar class of tuples is a complex process. When we consider uncertain data, the feature vector will not be a single valued but a function. Recently, different methods are proposed on decision tree based uncertain data classification with binary based operation on the decision tree. When multiclass data are given to the decision tree, their algorithm has to give repeated calculation to produce the probability distribution matching the class labels, thus time and memory utilization will be high for the particular algorithm. In this paper, we have intended to propose a classification method for uncertain data based on the decision tree. The proposed approach concentrates on an adaptive averaging method, where we have incorporated mean and median of the tuple to produce the feature value that will be used in the decision tree for decision making. Then a probability calculation is executed to find the relevance of tuple with respect to a class. If the calculated probability value is similar to a particular probability distribution, then the tuple is marked to that particular class. Thus, we produce a decision tree with c number of leaf nodes, where c is the number of class labels in the training data. The test data is subjected to the trained decision tree to obtain the classified data. . The experimental analysis are conducted for evaluating the performance of the proposed approach. The vehicle dataset and segment dataset from the UCI data repository is selected for the performance analysis. The results from the experimental analysis showed that the adaptive method has achieved a maximum average accuracy of 0.997 while the existing approach achieved only 0.985.*

Keywords: *uncertain data, probability distribution, decision tree algorithm, classification*

Nomenclature

D	*Dataset*
T	*Set of training tuples*
t	*tuples*
c	*class labels*
a	*attributes*
v	*feature vector*
p	*probability distribution*
f	*probability distribution function*

I. Introduction

Classification is a well-recognized Data Mining task and it has been studied extensively in the fields of statistics, pattern recognition, and decision theory, machine learning literature, neural networks and more. Classification operation usually uses supervised learning methods that induce a classification model from a database [3]. The task of classification is to assign a new object to a class from a given set of classes based on the attribute values of the object. [4]. the classification algorithm learns from the training set and builds a model. The model is used to classify new objects [5]. Numerous classification algorithms have been proposed in the literature, such as decision tree classifiers [10], rule-based classifiers [11], Bayesian classifiers [12], support vector machines (SVM) [13], artificial neural networks [14], Lazy Learners, and ensemble methods [15]. Decision tree induction is the learning of a decision tree from class-labeled training tuples A rule-based classifier is a technique for classifying records using a collection of "if ... then ..." rules. Bayesian classifiers are statistical classifiers and are based on Bayes theorem. SVM has its roots in statistical learning theory and has shown promising empirical results in many practical applications, from handwritten digit recognition to text categorization. An artificial neural network is a computational model based on biological neural networks. An ensemble method constructs a set of base classifiers from training data and performs classification by taking a vote on the predictions made by each base classifier [22].

In recent years, many advanced technologies have been developed to store and record large quantities of data continuously. In many cases, the data may contain errors or may only be partially complete. For example, sensor networks typically create large amounts of uncertain data sets. In other cases, the data points may correspond to objects which are only vaguely specified, and are therefore considered uncertain in their representation. Similarly, surveys and imputation techniques create data which is uncertain in nature [6].

Data uncertainty can be categorized into two types, namely *existential uncertainty* and *value uncertainty*. In the first type it is uncertain whether the object or data tuple exists or not. For example, a tuple in a relational database could be associated with a probability value that indicates the confidence of its presence [7]. In value uncertainty, a data item is modelled as a closed region which bounds its possible values, together with a probability density function (pdf) of its value [8, 9].

Classification of Uncertain data concerns with building classifiers based on uncertain data has remained a great challenge even the numerous classification algorithms [10-15] have been presented. There are early work performed on developing decision trees when data contains missing or noisy values [16], [17], [18]. Various strategies have been developed to predict or fill missing attribute values. However, the problem studied in this paper is different from before. Instead of assuming part of the data has missing or noisy values, we allow the whole dataset to be uncertain. Furthermore, the uncertainty is not shown as missing or erroneous values but represented as uncertain intervals and probability distribution functions. There are also some previous work performed on classifying uncertain data in various applications [19], [20], [21]. These methods try to solve specific classification tasks instead of developing a general algorithm for classifying uncertain data [22]. An intuitive way of handling uncertainty in classification is to represent the uncertain value by its expectation value and treat it as a certain data. Thus, conventional classification algorithms can be directly applied [23].

Thus, considering the methods and discussion over the methods resulted in designing an approach for classification of the uncertain data. As discussed in the above sections, the uncertain data has to be considered very specifically as to get it classified. The main objective is to design a modified algorithm based on the decision tree algorithm. The details of the proposed approach are listed in the section 5. The proposed approach is designed to deal with multiple classed and feature set of the uncertain data. The decision tree is modified such way that, it will be responsive with respect to the classes we provide.

The main contributions of the approach are,

- Analysis of different methods of uncertain data classification for improving performance.
- Developed a decision tree based algorithm for classifying uncertain data with the aid of adaptive averaging method
- The relevance of the proposed approach is evaluated through performance and comparative analysis

The rest of the paper is organized as, the 2nd section describes about recent researches on uncertain data. The 3rd section plots motivation behind the approach and 4th section plots the basic algorithms used in the proposed approach. The 5th and 6th section includes the detailed description about the proposed approach. In section 7, we plot the performance and comparative analysis and with section 8, we conclude the paper.

II. Review Literature

A handful of researches are available in the literature for uncertain data mining especially in classification. When reviewing the literature, initially,Smith Tsanget al [1] have extended classical decision tree building algorithms to handledata tuples with uncertain values. Extensive experiments havebeen conducted that show that the resulting classifiers are moreaccurate than those using value averages. Since processing pdf'sis computationally more costly than processing single values(e.g., averages), decision tree construction on uncertain data ismore CPU demanding than that for certain data. To tackle the problem, they proposed a series of pruning techniques that cangreatly improve construction efficiency.

Literature presents numerous research for uncertain data classification in which most of the works have modified the traditional' classifiers significantly for handling uncertain data. In various classification algorithms, Jinbo Bi and Tong Zhang [19] have adapted the traditional SVM classifier to the uncertain data. They presented a general statistical framework to tackle the problem of noisy data. Based on the statistical reasoning, they proposed a formulation of support vector classification, which allows uncertainty in input data. They derived an intuitive geometric interpretation of the proposed formulation, and develop algorithms to efficiently solve it. Empirical results were included to show that the newly formed method was superior to the standard SVM for problems with noisy input. After this work, researchers tried to modify the various classification algorithms for uncertain data. Accordingly, WlodzislawDuch [24] have discovered the relations between input uncertainties and fuzzy rules were systematically explored. Multi-layered perceptron (MLP) networks were shown to be a particular implementation of hierarchical sets of fuzzy threshold logic rules based on sigmoidal membership functions. They were equivalent to crisp logical networks applied to input data with uncertainty. Leaving fuzziness on the input side makes the networks or the rule systems easier to understand. Practical applications of those ideas were presented for analysis of questionnaire data and gene expression data.

In 2007, Jianqiang Yang and Steve Gunn [25] have proposed a approach of input uncertainty classification. The approach have developed a technique which extends the support vector classification (SVC) by incorporating input uncertainties. Kernel functions was used to generalize that proposed technique to non-linear models and the resulting optimization problem was a second order cone program with a unique solution. Then, again the same year, Biao Qin *et al* [22] 2007 have proposed a rule-based classification and prediction algorithm called uRule for classifying uncertain data. This algorithm introduced some measures for generating, pruning and optimizing rules. Those some measures were computed considering uncertain data interval and probability distribution function. Based on the measures, the optimal

splitting attribute and splitting value was identified and used for classification and prediction. The proposed uRule algorithm was process uncertainty in both numerical and categorical data.

In recent years, naive bayes and neural network were modified to handle the uncertain data. JiangtaoRen *et al* [2] have proposed a naive Bayes classification algorithm for uncertain data with a pdf. They addressed the problem of extending traditional naïve Bayes model to the classification of uncertain data. They have extended the kernel density estimation method to handle uncertain data. For particular kernel functions and probability distributions, the double integral was analytically evaluated to give a closed-form formula, allowing an efficient formula-based algorithm. Extensive experiments on several UCI datasets showed that the uncertain naive Bayes model considering the full pdf information of uncertain data was produced classifiers with higher accuracy than the traditional model using the mean as the representative value of uncertain data. Time complexity analysis and performance analysis based on experiments showed that the formula-based approach has great advantages over the sample-based approach.

In 2011, JiaqiGe *et al* [23] have proposed a neural network method for classifying uncertain data (UNN). They extended the conventional neural networks classifier so that it was taken not only certain data but also uncertain probability distribution as the input. They started with designing uncertain perceptron in linear classification, and analyze how neurons use the new activation function to process data distribution as inputs. They illustrated how perceptron generates classification principles upon the knowledge learned from uncertain training data. They also constructed a multilayer neural network as a general classifier, and proposed an optimization technique to accelerate the training process.

IV. Motivation behind the Approach

The classifications of uncertain data become one of the tedious processes in the data mining domain. The uncertain data are contains tuples with different probability distribution and thus to find similar class of tuples is a complex process. When we consider uncertain data, the feature vector will not be a single valued but a function. Recently, Smith Tsang *et al* [1] have proposed a decision tree based uncertain data classification. In the method they have utilized a binary based operation on the decision tree. When multiclass data are given to the decision tree, their algorithm has to give repeated calculation to produce the probability distribution matching the class labels, thus time and memory utilization will be high for the particular algorithm. Inspired from the research,we proposed a modified decision tree algorithm for handling uncertain data. The proposed is based on a mean and median approach. The decision tree make decision based on mean and median of the uncertain data. The averaging method presented in [1] is adopted in the proposed approach for handling uncertain data. The major change in decision tree algorithm is regarding the split function and it is defined based on the mean and median parameter.

III. Decision tree Algorithm

A decision tree is typically used for classification purposes among different types of data. Usually a decision tree algorithm consists of tuples and attributes. A tuple is a part of data defined by attributes. According to the decision tree algorithm, a dataset D contain T set of training tuples with set of attribute defined in set A, which can be represented as,

$$D \Rightarrow T = [t_1, t_1, ..., t_n]$$

Here n is the maximum number of tuples present in the dataset. Each tuple in the set T is associated with a feature vector v. The value v is associated with the attribute corresponding to the tuple and a class label c. Here, $t \Rightarrow [(a_1, a_2, ..., a_n), c_i]$ represents a tuple and with class label. The classification problem is to construct a model M that maps each feature vector $v \Rightarrow [(a_1, a_2, ..., a_n), c_i]$ to a probability distribution P_x on set of classes C such that given a test tuple $t = [(v_1, v_2, ..., v_n), c_i]$, probability distribution P over feature vector predicts the class label c with high accuracy. A binary decision tree is constructed in the proposed approach with split function z_i. The split function is derived from the values of attributes. The split function is used to construct nodes on decision tree to either left or right. A test $v_i < z_i$ predicts the tuple goes to left or right on the decision tree. This will be associated with the distributionof feature vector v_i.

A class label c_i of a given test tuple $t_{test} \Rightarrow [(a_1, a_2, ..., a_n), c = ?]$, we traverse the tree starting from therootnode until a leaf node is reached. When we visit an internalnode n, we execute the test $v_{test} \leq z_n$ and proceed to the leftchild or the right child accordingly. Eventually, we reach aleaf node m. Theprobability distribution Pm associated withm gives the probabilities that t_{test} belongs to each class label $c \in C$. For a single result, we return the class label $c \in C$ that maximizes Pm(c). Consider the following example,

$$T \Rightarrow [(t_1 : v_1 = 3), (t_2 : v_2 = 1)(t_3 : v_3 = 2)],$$

class➔ A, B, Z_n=2,

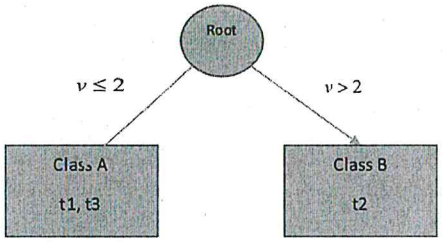

Fig.1. decision tree

A sample decision tree is represented in the Fig 1, which shows how the tuples t_1,t_2 and t_3 are classified into class A and B based on the split function. The split function is defined based on the context of classification, usually a user defined constant value.

V. Uncertain data and Uncertain Data Handling

The main objective of the proposed approach is concerned with the classification of uncertain data using the decision tree algorithm. As per the discussion in the above section, a feature vector is usually represented by a single value. On the other hand, in uncertain data, the feature vector is represented in a specific range or is known as probability distribution function (pdf). The pdf is defined over range [p, q] over the data and is in a closed form. The decision tree algorithm considered in the proposed approach cannot process feature in the above described form. So the possible way available is to convert the pdf into single values. The process need complex calculation effectively reduce the pdf into a single value that can be processed with the decision tree algorithm.

The process would be implemented numerically by storing a set of sample points, x, which is a values in the range [p, q] and will be stored on a set S. The value x is associated to the value $f_i(x)$, which is the pdf. The sample points created are used to effectivelyapproximating $f_i(x)$ by a discrete distribution with s possiblevalues. Considering the decision tree scenario to handle the uncertain data, a decision tree considered in our uncertainty model is of the point-data model. The difference lies in the way the tree is employed to classify test tuples with no class labels. Similar to the training tuples, a test tuple t_{test} contains uncertain attributes. So the feature vector is thus a vector of pdf's $v = [f_1, f_2, \ldots, f_n]$. A classification model is thus a function M that maps such a feature vector to a probability distribution P over C. The probabilities for P are calculated as follows. During these calculations, we associate each tuple t_i weight w in the range 0 and 1. The idea is interpret the conditional probability of the tuple under test with weight x, to identify the class label. The class label that possess highest probability for tuple will be the class of that particular tuple. There have been numerous approaches are proposed for handling and classifying the uncertain data, here, we proposed an adaptive averaging based method to classify the uncertain data.

VI. Adaptive Averaging Method

In this section, an approach known as averaging is user for handling the uncertain data. The averaging is the process of replacing the pdf with point value. The expected point value is constructed from mean values of all pdfs. In other words, for a tuple t_i with attributes A_j, we calculate the mean value of pdf f_i and the mean value is considered as the feature vector.Averaging 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, and hence traditional decision tree algorithms can be applied. In the proposed approach, we adopt an adaptive averaging method because the averaging method in [1]uses mean value alone as the parameter for deciding the class label. This can lead to lack of precision, as probability of tuples with particular mean value is high, then that particular tuple will be given a class label regarding that mean value, for example,

TABLE.1.
TUPLE AND MEAN VALUES

Tuples	mean	class
1	2	A
2	2	B
3	1	B
4	1	B

In table 1, we can see that, tuple 2 is actually belong to class B, but since its mean value is 2 and according to the split function defined on the averaging method, the class label of tuple 2 will be assigned as B. So in order to resolve this problem, we have to design a precise decision parameter from the uncertain data. The proposed averaging method uses two parameters for deciding the decision parameter or the feature vector, mean and median.

VI.1. Uncertain Data Preparation

To start with the training phase of the decision tree algorithm, we need to have the uncertain data to formulate the input feature vector. The proposed approach uses a non-uncertain data for the creating uncertain data. The data set D, contains n number of tuples and k number of attributes. The proposed approach select each attribute from the tuple and create a series of data from it. The range of the created data will [p,q]. The series of values for particular attribute will follow a distribution function as per the weight defined for tuple. The weight of each tuple will be defined in between 0

and 1. As per the averaging [1] method, a single attribute is selected for the uncertain data classification. So we select an attribute from the set of attributes A_i of the tuple t_i in the data set D,

$$D = [t_1, t_2, ..., t_n]$$

$$t_i = [a_1, a_2, ..., a_k], a_i \in A$$

From, the set of attributes a_1 is selected for the uncertain data formation. As of now a_1 will represent as single value. Now, we define a range over the a_1 to produce the uncertain data. Thus a_1 will be represented with in a particular range. There will also be a number of value for a, which will be used for calculation of mean and median.

$$a_i = [a_{i1}, a_{i2}, a_{i3}, ..., a_{ij}]$$

The seta$_i$ shows a sample data generated for the attribute a_i from the tuple t_i selected from the dataset D. There are a total j elements in the sample data prepared. This data is then subjected as the pdf of the tuple t_i. Then, we subject the mean and median calculation on the defined pdf. All the tuples in the dataset are also processed similarly. Consider the following example,

$$tuple = t_i; attribute = 51,$$

$$pdf(t_i : a_i) = [50,51,52,52,53]$$

The above sample represent the pdf of t_i with respect to attribute a_i. Now, we have probability distribution if all tuples with respect of a single attribute. As per the averaging, we have to convert the pdf into a point value or into representative of the pdf. We use two parameters for converting the pdf into representatives.

1. Mean

The mean value calculation is adopted from the averaging method represented in [1]. Each tuple is assigned weight in the range 0 and 1. The mean is calculated based on the number of sample points used to represent the tuple in a particular weight. The mean value is calculated based on the basic mean calculation formula defined as,

$$Mean(t_i) = \frac{\sum_{i=1}^{k} a_i}{N}$$

Here, $a_i$represents the values from the pdf and N represent the total number of elements present in the pdf. The mean is rather considered as an expected value for the tuple from the attribute value defined in set a_i.

2. Median

The reason of considering the median value is that, if have five elements in the pdf and which can be represented as, [2,3,2,2,10]. The mean value of this particular tuple will be obtained as, 3.8, but it will be too high with respect to the majority of elements represented in pdf. This happened because of the single value 10. So in such circumstances we can select the representatives as the median of values in the set. Thus we get 2 as the expected value for the tuple, as the most frequent value in the set is 2. The feasibility of considering the median value for all the tuple will not be a wise decision. So, the proposed approach incorporated median and mean in the averaging method to obtain precise result

3. Adaptive averaging based on mean and median.

The proposed approach deals with a combined method of mean and median to extract the feature value for the given tuples.Initially,we select all the tuples and their newly formed probability distribution functions. The distribution functions are then subjected to mean calculation and median calculation. Now, assessing the total number of units present in the pdfs, we set a threshold value *th*for selection in between mean and medium. The adaptive averaging method state that, by considering the mean value and median value with respect to the threshold, we can precisely define the feature value. The method defines that, if the average obtained from pdf is higher than the threshold *th*, then we use median value instead of mean otherwise mean is used. This process will help in identifying the most appropriate feature value for decision making in the decision tree. Sometimes the mean value can be affected by calculation error or with small number of elements getting high value. In order to resolve this problem median can be performed and which leads to a conclusion that an optimized feature value can be obtained through adaptive averaging method.

Algorithm 1.

Step1. Accept tuple set T,

$T = [t_1, t_2, t_3, ..., t_n]$

Step2. Select attribute a_i of t_i

Step3. Select samples from a_i

Step4. Compute mean,

$$Mean(t_i) = \frac{\sum_{i=1}^{k} a_i}{N}$$

Step4. Calculate median distribution $f(a_i)$

$$meaian, m = P(ai \le m) \le \frac{1}{2}$$

Where, m is any real number

Step5. Calculate Mean and m for all tuples in the set T

Step6. Set threshold *th*

Step7. Define v, feature value for each t_i

Step8. $if\ (mean > th), assign\ v = m$

$else, v = mean(t_i)$

Step9. Repeat procedure for all tuples

Step10. Store v values in V, set of feature values.

Step11. End

VII. Decision tree for Uncertain Data Classification

The main objective discussed in the proposed approach is to design and develop a method based on decision tree to classify the uncertain data. The decision tree used in the proposed approach is a binary decision tree and it uses single valued data for classification. As discussed in the section 4, a decision tree has two phases training phase and testing phase. In the training phase of decision tree, we provide the data that is obtained after the adaptive averaging method. That is, the data will be a known data with feature values obtained from the pdf of each tuple and the class label corresponding to each label. The data will be like,

$$training\ data = [v_1, v_2, ..., v_n]$$

Where, v is the feature value of a tuple t_i and it contains three parameters particularly,

$$v_i = [t_i, m(t_i) / mean(t_i), c_i]$$

Here, t_i represents the tuple, $m(t_i)$ represents the median of the tuple and $mean(t_i)$ represents the mean, finally c_i represents the class label of the particular tuple t_i. Initially, the decision tree algorithm checks whether all the tuples possess same class label, then all the tuples will be grouped into one leaf node. If the classes are different, a split function is utilized to put the tuples into left of right of the leaf node. The split function, z_i, uses the mean or median value of the corresponding tuple to plot it into either left or right of the leaf nodes. We can consider the mean or median as the decision parameter x, then,

$$if\ (x > z)$$
$$put\ it\ left\ node$$
$$else$$
$$put\ \ it\ right\ node$$

Now, we calculate the probability of the nodes in left and right with respect to the class labels. The probability can be subjected as $P(c_i)$ and the tuples with higher probability regarding a class label then the tuple is assigned to that class. Thus we obtain the probability range of all the tuples corresponding to that particular input. We set an average probability of classes $p(c_i)_{avg}$ for each class c_i in set C. This will be used to classify an unknown uncertain data.

In the testing phase of the decision tree, an unknown test tuple is given to the trained decision tree. The test tuple will be in the similar format like a training tuple but with the class label field empty or unknown. The test tuple t_{test} is given to the decision tree algorithm and the split function plot it into either left of right. Then correspond probability is compared with the probability of test tuple. As per the obtained probability value, the test tuple is plotted to the corresponding class. In similar way we can calculate and classify any kind of uncertain data with unknown class labels. The adaptive averaging method is utilized to give more precise x value to determine each classes clearly.

VIII. Experimental Results and Discussion

The experimental analysis discuss about the performance of the proposed approach in classifying the uncertain data. The performance is analyzed based on the modified method using the decision tree algorithm. The experiment is conducted according to different dataset with normal data and intruded data or the unwanted data. The experiments will evaluate, how efficiently the proposed approach will classify the clean data and the unwanted data. Later on the analysis is studied in detail and the relevance of the proposed approach is stated according to the result analysis.

VIII.1 Experimental Setup

The experiment is conducted on a system with Intel core i5 processor, running on 4 GB of RAM and 500GB of hard disk. The experiment utilized most of the memory form the RAM and dataset are stored in the hard disk space. The programs for the experimental analysis are developed using the JAVA programing language under JDK 1.7.0.

VIII.2. Dataset Description

The dataset used for the proposed classification technique on uncertain data are based on noon- uncertain datasets. In order to analyze the feasibility of the proposed approach, we have generated uncertain data from the known da. Thus the data generated can help in identifying the responses of the proposed approach, whentreated with actual uncertain data. The proposed approach uses two dataset for the processing uncertain data classification, which are mainly,

Vehicle dataset [26]: This data was originally gathered at the TI in 1986-87 by JP Siebert. It was partially financed by Barr and Stroud Ltd. The original purpose was to find a method of distinguishing 3D objects within a 2D image by application of an ensemble of shape feature extractors to the 2D silhouettes of the objects. Four "Corgie" model vehicles were used for the experiment, a double decker bus, Chevrolet van, Saab 9000 and an Opel Manta 400. This particular combination of vehicles was chosen with the expectation that the bus, van and either one of the cars would be readily distinguishable, but it would be more difficult to distinguish between the cars. There are total of 4 classes and 946 sample or attributes.

Segment dataset[26]: The dataset contains the instances drawn randomly from a database of 7 outdoor images. The images were handsegmented to create a classificationfor every pixel. There are total of 210 trained data and 2100 test data. The number of attributes possessed by the dataset is 19 and there are 7 classes.

VIII.3. Performance analysis

The performance of the proposed approach is conducted based on the evaluation parameter accuracy. Initially we select all the elements from the two dataset and a column containing an attribute is extracted from them. Then the attribute is converted into uncertain data attributes by processing based on the step 6.a. In the performance evaluation process, we have selected sample of 100, 200 and 300 groups. The three set of uncertain data samples are processed with the proposed classification algorithm. The result of the analysis are plotted in the following section.

Accuracy Analysis based on Vehicle Data

In this section we plot the analysis on accuracy based on the vehicle dataset. The dataset is initially selected as three sets with 100, 200 and 300 elements respectively. The analysis is conducted by varying the weights on the sample ranging from 0.3 to 0.9 with an interval of 0.2.

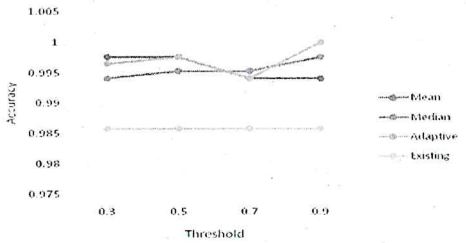

Fig.2. Accuracy analysis over 100 samples

The Fig 2 represents the analysis of the proposed approach on accuracy by considering 100 samples of the vehicle dataset. In the analysis we have calculated accuracy based on three other parameters also. The parameters are mainly mean, median and existing averaging method [1]. The analysis from the Fig shows that, when calculating mean alone, we have achieved an average accuracy of 0.995. The average accuracy is 0.994 for the median based analysis. The accuracy is increased to 0.997 for the adaptive based method. The analysis based on 200 and 300 samples are given in the following graphs,

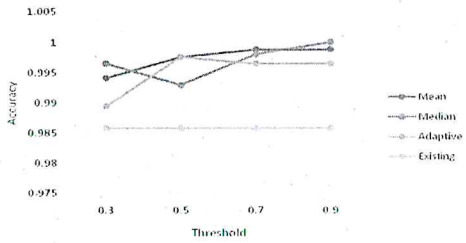

Fig.3. Accuracy analysis based on 200 samples

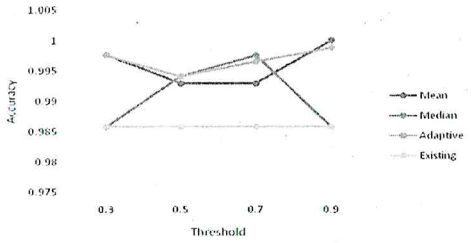

Fig.4. Accuracy analysis based on 300 samples

The Figs 3 and 4 represents the accuracy analysis based on 200 and 300 samples extracted from the vehicle data. Here also we can see that the proposed adaptive method has better accuracy over the other method. In all the cases, we can see one thing that, the existing averaging method has achieved an average accuracy rate of 0.985. As the number of samples increases there are slight deviations in the accuracies of all the other methods including the proposed method. Though, the adaptive method possess an average accuracy more than 0.996 compared to the mean and median based methods.

Accuracy Analysis based on Segment data

Here, we plot the analysis on accuracy based on the segment dataset. The dataset is grouped into three sets with 100, 200 and 300 elements respectively. The analysis is conducted by varying the weights on the sample ranging from 0.3 to 0.9 with an interval of 0.2.

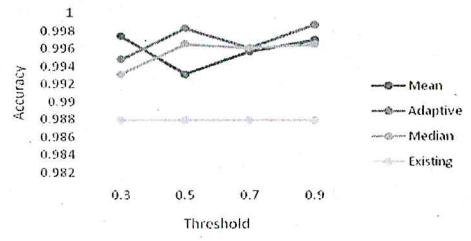

Fig.5. Analysis based on 100 samples

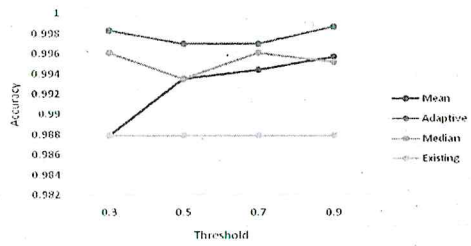

Fig.6. Analysis based on 200 samples

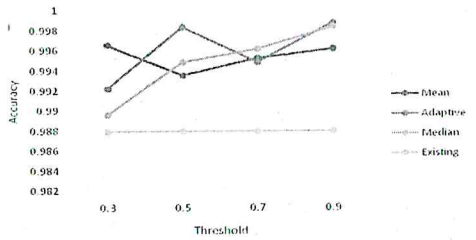

Fig.7. Analysis based on 300 samples

The Figs 5, 6 and 7 represents the accuracy analysis of the proposed approach with respect 100, 200 and 300 data samples collected from the segment dataset. The analysis showed similar result like the vehicle data set. In this analysis also, the existing method based on averaging has achieved average accuracy of 0.987 on all the three samples of data. On the other hand, the adaptive method achievedvaryingaverage accuracies on the three set of samples, which can listed as 0.995, 0.995 and 0.994 respectively for 100, 200 and 300. The accuracy values obtained for mean based and median based methods are considerable low compared to the adaptive method. So the analysis from the two dataset indicates that, the adaptive method is efficient in classifying the uncertain data compared to the other three methods.

IX. Conclusion

Classification of Uncertain data concerns with building classifiers based on uncertain data and has remained a great challenge even the numerous classification algorithms have been presented in the prior sections. In this paper, we have presented a modified averaging based methodology for uncertain data classification. The various steps included in the proposed approach are uncertain data preparation from normal data, mean calculation, median calculation and adaptive feature value selection. The adaptive feature value calculation is a method to select the best of mean or median regarding a tuple. The z function on the decision tree will be designed based on the adaptive method. The experimental analysis are conducted for evaluating the performance of the proposed approach. The vehicle dataset and segment dataset from the UCI data repository is selected fro the performance analysis. The results from the experimental analysis showed that the adaptive method has achieved a maximum average accuracy of 0.997 while the existing approach achieved only 0.985. The analysis indicate that the adaptive method used for selecting the feature value enhances the accuracy of classification of uncertain data.

References

[1] Smith Tsang, Ben Kao, Kevin Y. Yip, Wai-Shing Ho, Sau Dan Lee, "Decision Trees for Uncertain Data, *TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, PP: 1-

[2] JiangtaoRen, Sau Dan Lee, Xianlu Chen, Ben Kao, Reynold Cheng and David Cheung,"Naive Bayes Classification of Uncertain Data, *"Ninth IEEE International Conference on Data Mining*,pp.944-949,2009.

[3] Fayyad U., Piatetsky-Shapiro G., Smyth P., From data mining to knowledge discovery: an overview, Advances in knowledge discovery and data mining, *American Association for Artificial Intelligence, Menlo Park, CA, AAAI/MIT Press*, 1996, pp: 1-36.

[4] Romero C., Ventura S., Espejo P.G., and Hervas C., Data Mining Algorithms to Classify Students, *proceedings of the 1st Int'l conference on educational data mining, Canada*, 2008, pp: 8-17.

[5] Zhang J., Mani I., kNN Approach to Unbalanced Data Distributions: A Case Study involving Information Extraction, *In Proceedings of The Twentieth International Conference on Machine Learning (ICML-2003), Workshop on Learning from Imbalanced Data Sets II,* August 21, 2003.

[6] Charu C. Aggarwal and Philip S. Yu, "A Survey of Uncertain Data Algorithmsand Applications", *IEEE transactions on knowledge and data engineering*, Vol. 21, No. 5, MAY 2009.

[7] Barbara, D., Garcia-Molina, H. and Porter, D. "The Management of Probabilistic Data," *IEEE Transactionson Knowledge and Data Engineering*, 4(5), 1992.

[8] Cheng, R., Kalashnikov, D., and Prabhakar, S. "Evaluating Probabilistic Queries over Imprecise Data,"*Proceedings of the ACM SIGMOD International Conference on Management of Data*, June 2003.

[9] Chau, M., Cheng, R., and Kao, B., "Uncertain Data Mining: A New Research Direction," in Proceedings of the Workshop on the Sciences of the Artificial, Hualien, Taiwan, December 7-8, 2005.

[10] J. R. Quinlan, *C4.5: Programs for Machine Learning*. Morgan KaufmanPublishers, 1993.

[11] W. W. Cohen, "Fast effective rule induction," in *Proc. of the 12th Intl.Conf. on Machine Learning*, 1995, pp. 115–123.

[12] P. Langley, W. Iba, and K. Thompson, "An analysis of bayesianclassifiers," in *National Conf. on Artigicial Intelligence*, 1992, pp. 223-228.

[13] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer Verlag,1995.

[14] R. Andrews, J. Diederich, and A. Tickle, "A survey and critique oftechniques for extracting rules from trained artificial neural networks,"*Knowledge Based Systems*, vol. 8, no. 6, pp. 373–389, 1995.

[15] T. G. Dietterich, "Ensemble methods in machine learning," *LectureNotes in Computer Science*, vol. 1857, pp. 1–15, 2000.

[16] Q. JR, *Probabilistic decision trees, in Machine Learning:An ArtificialIntelligence Approach*. Morgan Kaufmann, 1990.

[17] L. O and N. M, "Ordered estimation of missing values," in *PAKDD*,1999, pp. 499–503.

[18] H. L, S. A, and S. M, "Dealing with missing values in a probabilistic decisiontree during classification," in *The Second International Workshopon Mining Complex Data*, 2006, pp. 325–329.

[19] B. Jinbo and Z. Tong, "Support vector classification with input data uncertainty,"in *Advances in Neural Information Processing Systems*, 2004, pp. 161–168.

[20] E. V. Gonzalez, I. A. E. Broitman, E. E. Vallejo, and C. E. Taylor,"Targeting input data for acoustic bird species recognition using datamining and hmms," in *ICDMW 2007*, pp. 513–518.

[21] C. Jebari and H. Ounelli, "Genre categorization of web pages," in*ICDMW 2007*, pp. 455–464.

[22] Biao Qin, Yuni Xia, Sunil Prabhakar and YichengTu, " A Rule-Based Classification Algorithm forUncertain Data", *IEEE 25th International Conference on Data Engineering*,pp.1633 - 1640,2009.

[23] JiaqiGe, and Yuni Xia, "UNN: A Neural Network for uncertain data classification",*In Proceedings of PAKDD*, vol.1, pp.449-460, 2010.

[24] WłodzisławDuch,"Uncertainty of data, fuzzy membership functions, and multi-layer perceptrons,"*IEEE Transactions on Neural Networks*,vol.16,no.1,pp.10-23,2005.

[25] Jianqiang Yang and Steve Gunn,"Exploiting Uncertain Data in Support VectorClassification,"Springer-Verlag Berlin Heidelberg, pp. 148–155, 2007.

[26] Web reference http://archive.ics.uci.edu/ml/datasets.html.

S. Meenakshi (Sankarasubramanian Meenakshi) obtained her Bachelor's degree from Madurai Kamaraj University, Tamilnadu, India. Then she obtained her Master's degree in Computer Applications and pursuing her PhD in Computer Science in Anna University majoring in Data Mining, She has also obtained Sun Certified Java professional (SCJP) qualifications. Currently, she is a lecturer at the Faculty of Computer Applications, The Kavery Engineering College, Affiliated to ANNA University Chennai, approved by AICTE. Her research interests are Artificial Intelligence, Social Network Analysis, Game Theory, and Virtual Reality.

Dr.V. Venkatachalam (Varadharajan Venkatachalam) received the B.E. degree in Electronics and Communication from Bharathiyar University in 1989 and M.S. degree in software systems from Birla Institute of Technology in 1996. He received M.Tech. Degree in Computer Science from National Institute of Technology, Trichy in 2004 and the Ph.D. degree from Anna University, India in 2009. From 1990 to 2000, he worked as a Lecturer at Kongu Engineering College, Perundurai, Erode. He worked as an Assistant Professor at Erode Sengunthar Engineering College from 2000 to 2008. He is currently the PRINCIPAL of "The Kavery Engineering College", Mecheri, Salem. His research interests are Network Security and Cryptography, Data Mining, Artificial Intelligence and Wireless Networks. Dr.Venkatachalam is a life member of the ISTE.(Indian Society of Technical Education)