Handling Uncertainty Information through Extended Classifiers

[1]Poonam.Khaparde,[2]Farhana Zareen ,[3]Dr.R.V.Krishnaiah

[1]Department of SE, JNTU H, DRK Institute of Science and Technology
Hyderabad, Andhra Pradesh, India

[2]Department of CSE, JNTU H, DRK College of Engineering and Technology
Hyderabad, Andhra Pradesh, India

[3] Principal, Department of CSE, JNTU H, DRK Institute of Science and Technology
Hyderabad, Andhra Pradesh, India

Abstract

The data whose values are precise is known as certain data whereas uncertain data is the data whose values are not precise. It does mean that value of a data item is represented by multiple values. The traditional data mining algorithms, especially classifiers work on certain data. They can't handle uncertain data. This paper extends traditional decision tree classifiers to handle such data. We understood that the simple mean and median of uncertain values can't give accurate results. For this reason this paper considers Probability Distribution Function (PDF) to improve the accuracy of decision tree classifier. It also proposes pruning techniques to improve the performance of the classifier. Empirical results show that, when compared to algorithms that use averages of uncertain values our algorithm is more accurate. However, it is computationally more expensive as it has to compute PDFs. Our pruning techniques help in reducing the computational cost.

Keywords-*Data mining, uncertain data, decision tree classifiers, and pruning.*

1. Introduction

Data mining is a process of extracting trends from historical data. These trends or patterns form business intelligence that leads to well informed business decisions.In data mining domain classification is one of the algorithms. It is also part of machine learning [1]. Classification algorithm takes a dataset containing training tuples and programmatically predicts the class labels of tuples including unknown ones based on the feature vector of tuple. Decision tree model is one famous classification model. Decision trees show the practical information that is useful in taking decisions. From decision trees, it is easy to extract rules and follow them. Based on decision tree models many algorithms came into existence. They include C4.5 [2], ID3 [3] etc. Due to the usefulness of these algorithms they are widely used. They are practically used in many real time applications. The applications include fraud detection, scientific tests, medical diagnosis, target marketing etc. A feature or attribute of a record is used in traditional classification. The type of the attribute might be categorical or numerical. In case of numerical data some point value is expected. There is no problem when there is single value for an attribute. However, there is problem when a numeric attribute has multiple values. Such data item with multiple values is known as uncertain data. Then such data has to be handled differently. Probability distribution function has to be used to handle such data. However, a simple way to solve it is to obtain abstract probability distributions by summary values such as variances and means. This procedure is known as averaging. In another approach known as "Distributed-based" considers complete information for classification.

In this paper, the problem of constructing decision tree classifiers for uncertain data which is numerical in nature is carried out. The algorithms converted uncertain values and point values and processed further. We proposed two algorithms known as averaging and also distribution – based. When compared with the averaging method, the distribution – based approach is computationally expensive. Averaging is based on the summary statistics; it is simple and does not cause expensive operations. In case of distribution – based algorithm, it has to compute PDFs which involves extensive data processing for generating decision trees for uncertain data. Therefore, it is essential to minimize the computational cost in case of the second algorithm. This is achieved by using a series of pruning techniques.

2. Related Work

In recent years, the usage of data mining in real time applications has been increased. Huge amount of data is being processed for making business decisions. The trends or patterns that are extracted from historical data are used to make well informed decisions as they form business intelligence. However, the data mining algorithm such as classification is widely used. Decision trees are the result of classification algorithm. For instance ID3 algorithm is one of the examples of decision tree algorithms. It is widely used as it can produce a tree of decisions suitable for business decisions. These algorithms work fine with data with certain values. However, there are attributes that may have numerical data that is uncertain. It does mean that the attribute which has multiple values is known to be uncertain. The traditional classification algorithms can't work on the uncertain data. In that case probability distribution of values is to be considered. Semi structured data and XML [4], [5] are subjected to probabilistic databases. Uncertainty of values exists when the values of an attributes are not known. Any data item whose values are uncertain is represented by a pdf. It is computed from the possible values [6]. Imprecise query processing is well suited to value uncertainty. Probabilistic guaranty of correctness influences the quality of answer to such query. For the purpose of solving range queries indexing solutions can be used on uncertain data [7]. Indexing solutions for also help in aggregate queries [8] such as NN queries. Indexing solutions also help in solutions for location – dependent queries [9]. Uncertain data mining has been in the research circles recently. Well known K-means algorithm is improved and named UK-Means algorithm [9] that can handle uncertain data. Afterwards pruning techniques came into existence for improving the performance of UK-means algorithm. The pruning algorithms are namely CK-means [10] and min-max-dist pruning [11]. Other algorithms that came into existence for handling uncertain data are density-based classification [12]; frequent item set mining [13], etc. Each data point is assigned an error model in [12]. Each attribute is operated independently and uncertainty is handled effectively.

In the form of missing values [3], [2], decision tree for uncertain data has been addressed for many years. When values are not available for certain attributes missing values come. Solutions of that kind include approximating missing values using a classifier [14]. Fractional tuples are also used to handle missing values in training data. There is another related approach known as fuzzy decision tree that is based on fuzzy information models which can also handle uncertain data [15]. Our work is based on the distribution. It does mean that it gives classification results as distribution. Many variations are available fuzzy extension to [16], [15] and [17]. In all these models an attribute works as an important attribute that can be used for classification and thus a decision tree is generated. It is computationally demanding to build decision trees on tuples with point values data and numerical data [18]. However, it is much more expensive when such data is of type uncertain data. For best "split point" a numerical column can have large search space. In this case finding the best "split point" itself is computationally expensive. In [19], and [20], candidate split points are reduced by using many techniques for efficiency. Well known evaluation functions like Gini Index [21] and Gain [3] are also used by these techniques.

To overcome the problems of the previous works, this paper presents a set of algorithms and also pruning techniques that help in handling uncertain data. The result of these algorithms is to produce a decision tree that helps in taking well informed decisions in real world applications. Enterprises use these techniques to handle uncertain data. Especially we developed two algorithms namely averaging and distribution – based. These two algorithms can handle uncertain data. The first algorithm is based on simple summary statistics and thus less expensive as it involves no computations further. However, in case of distribution-based algorithm, it computes pdfs for each and every attribute and finally produces decision tree. Computing PDFs is an expensive operation and thus it consumes more processing power. To overcome this problem we introduced a series of pruning techniques that can effectively reduce the computational cost of the operations.

3. Problem Description

This section describes problem of classifying uncertain data. It discusses both classical decision trees and decision trees for uncertain data.

3.1 Classical Decision Trees

Traditional decision trees work on data which is precise. The data here is containing number of tuples. For each and every domain or attribute, the values are single values and precise values. They are certain values. The classical decision trees take such dataset and perform data mining operation such as classification. The result of this operation is a decision tree that helps in taking well informed decisions.

3.2 Handling Uncertain Information

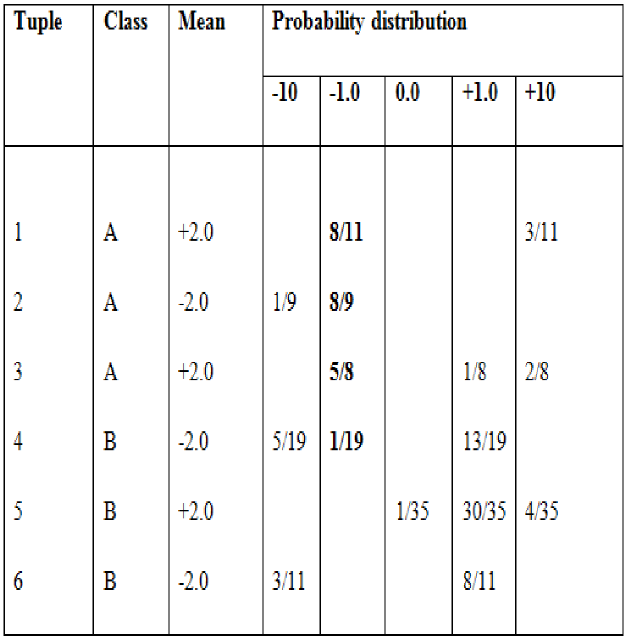
The uncertainty model devised by us does not take a single value for a feature. The feature value is represented by a set of values. From such data PDFs can be computed analytically. This representation helps the amount of data is exploded. The richer information, the better is the classification model. The drawback of this is, of course, computational cost is high as computing PDFs involve large amount of data to be processed. We discovered a fact that simple averaging of uncertainty data can't improve classification accuracy. For this reason we opted computing PDFs from uncertain data that improves accuracy dramatically. The resultant decision tree looks like the point data model. The difference is found in the way the tree is employed. A test tuple values are uncertain and the feature vector is nothing but PDFs computed. Therefore a classification model is a function represented by M that maps the feature vector to a P (probability distribution) over C. For given tuple t, and attribute Ajn the PDF is computed as

$$f_{L,j_n}(x) = \begin{cases} f_{x,j_n}(x)/w_L & \text{if } x \in [a_{x,j_i} \\ 0, & \text{otherwise.} \end{cases}$$

It is very challenging to construct a decision tree for uncertain data. It needs finding a testing attribute suitable for decision making. The algorithms for constructing decision trees for uncertain data are provided in the next section.

4. Algorithms for Handling Uncertain Data

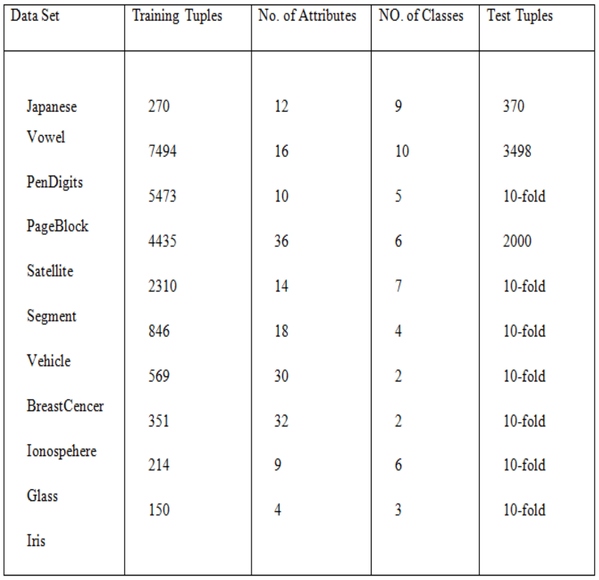
In this section two approaches are discussed that are meant for handling uncertain data. The first approach is known as "Averaging" while the second approach is named "Distribution – based". The first approach transforms uncertain data into point valued type. It is achieved by replacing each PDF with corresponding mean value. The mean values and probability distribution values are presented in Table 1.



Tuple	Class	Mean	Probability distribution				
			-10	-1.0	0.0	+1.0	+10
1	A	+2.0		8/11			3/11
2	A	-2.0	1/9	8/9			
3	A	+2.0		5/8		1/8	2/8
4	B	-2.0	5/19	1/19		13/19	
5	B	+2.0			1/35	30/35	4/35
6	B	-2.0	3/11			8/11	

Table 1

As can be seen in table 1, the mean and probability distribution values for given tuples are presented. The feature vector of tiis transformed into (vi, 1, …., vi, k). Afterwards a traditional decision tree construction algorithm can be used to build decision tree. The second approach is used to fully exploit the pdfs. The datasets used for the experiments are shown in table 2.



Data Set	Training Tuples	No. of Attributes	NO. of Classes	Test Tuples
Japanese	270	12	9	370
Vowel	7494	16	10	3498
PenDigits	5473	10	5	10-fold
PageBlock	4435	36	6	2000
Satellite	2310	14	7	10-fold
Segment	846	18	4	10-fold
Vehicle	569	30	2	10-fold
BreastCencer	351	32	2	10-fold
Ionospehere	214	9	6	10-fold
Glass	150	4	3	10-fold
Iris				

Table 2 – Datasets collected from UCI machine repository

In the following sections the algorithms used under the two approaches are described.

4.1 Averaging

It is one of the ways to deal with uncertain data. In this approach each pdf is replaced by its corresponding value. This results in conversion of data tuples into point-valued ones. Once the data is converted into point-valued data, then traditional algorithms such as ID3 [22] can be used. This approach is known as averaging. When an attribute value is not certain, a range of values is to be considered. In this case probability distributed is calculated. By using summary statistics such as variance and means it can be achieved. This is known as averaging.

4.2 Distribution-Based Approach

In this approach also same procedure is used as described above for converting uncertain data into point values. An attribute such as Ajn and a split point zn are chosen.

Afterwards, the whole set of tuples S is divided into two subsets named L and R.

$$\text{If } b_{i,j_n} \leq z_n$$

L = ti

$$\text{if } z_n < a_{i,j_n}$$

R = ti

Then ti is split into two fractional tuples named tL and tR. This algorithm is known as Uncertain Decision Tree (UDT).

4.3 PRUNING ALGORITHMS

A more accurate decision tree can be built using UDT when compared to averaging. However, averaging is much faster than UDT and computationally less expensive. As UDT is computationally more expensive, pruning techniques are required to overcome this drawback. The pruning techniques used are taken from [23]. They are pruning empty and homogenous intervals; pruning by bounding and end point sampling.

4.4 PROTOTYPE APPLICATION

The application developed for demonstrating the effectiveness of the proposed algorithms is with Graphical User Interface to be user friendly. The main screen of the application is as shown in fig. 1. This application with synthetic data set is executed and the results are shown in fig. 2, 3, and 4.

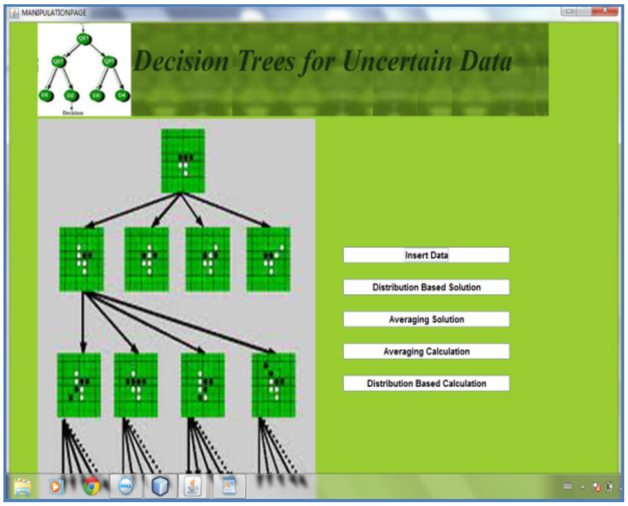
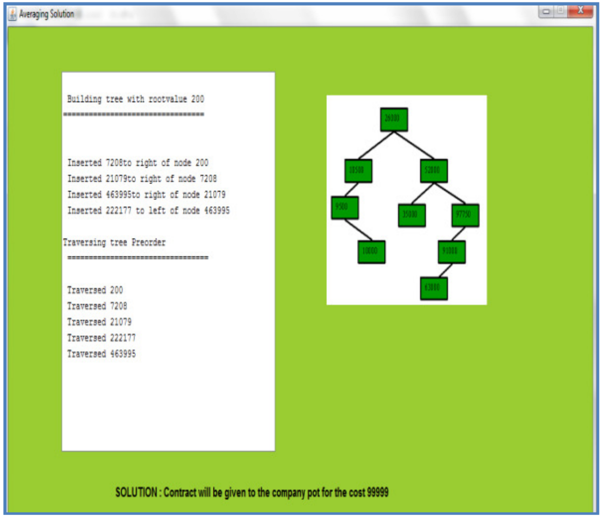

MANIPULATIONPAGE
Decision Trees for Uncertain Data
Decision
Insert Data
Distribution Based Solution
Averaging Solution
Averaging Calculation
Distribution Based Calculation

Fig. 1 –The main screen of the application

As can be seen in fig. 1, the application allows data insertions to get synthetic data, distribution based solution, and distribution based calculations, averaging solution and averaging calculations.

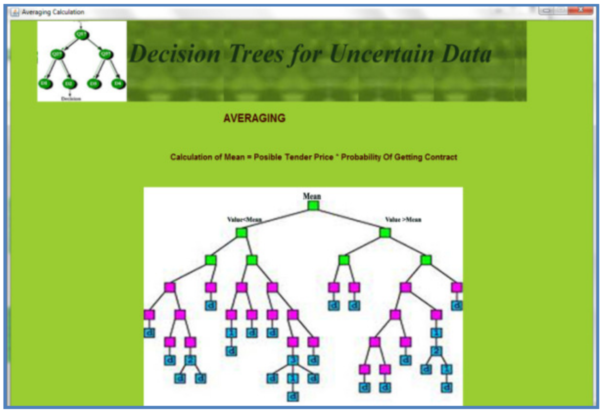


```
Building tree with rootvalue 200
================================


Inserted 7208to right of node 200
Inserted 21079to right of node 7208
Inserted 463995to right of node 21079
Inserted 222177 to left of node 463995

Traversing tree Preorder
=================================


Traversed 200
Traversed 7208
Traversed 21079
Traversed 222177
Traversed 463995
```

SOLUTION : Contract will be given to the company pot for the cost 99999

Fig. 2 –Averaging solution

As can be seen in fig. 2, averaging solution is presented.

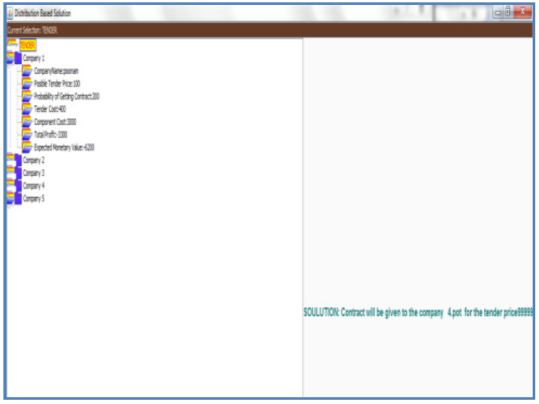


AVERAGING

Calculation of Mean = Posible Tender Price * Probability Of Getting Contract

Fig. 3 – Averaging calculation

As can be seen in fig. 3, shows calculation criteria of averaging.



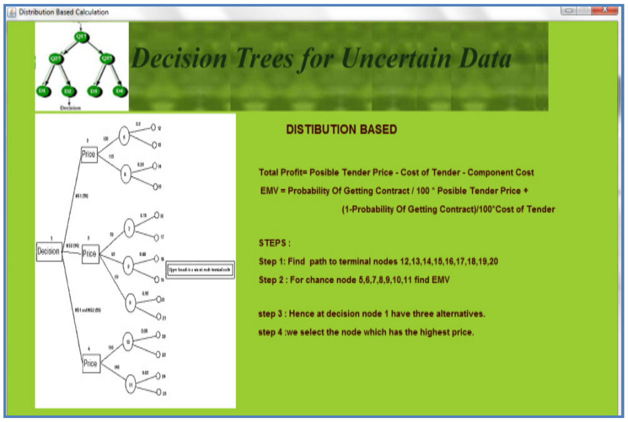
Current Selection: TENDER

- TENDER
  - Company 1
    - CompanyName:poonam
    - Posible Tender Price:100
    - Probability of Getting Contract:200
    - Tender Cost:400
    - Component Cost:3000
    - Total Profit:-3300
    - Expected Monetary Value:-6200
  - Company 2
  - Company 3
  - Company 4
  - Company 5

SOULUTION: Contract will be given to the company 4.pot for the tender price99999

Fig. 4 – Distribution based solution

As can be seen in fig. 4, distribution – based solution is presented.



Decision Trees for Uncertain Data

DISTIBUTION BASED

Total Profit= Posible Tender Price - Cost of Tender - Component Cost
EMV = Probability Of Getting Contract / 100 * Posible Tender Price +
(1-Probability Of Getting Contract)/100*Cost of Tender

STEPS :
Step 1: Find path to terminal nodes 12,13,14,15,16,17,18,19,20
Step 2 : For chance node 5,6,7,8,9,10,11 find EMV

step 3 : Hence at decision node 1 have three alternatives.
step 4 :we select the node which has the highest price.

Fig. 5 – Distribution based calculation

As can be seen in fig. 5, the distribution – based calculation criteria is presented.

5. Experimental Results

The environment used for experiments include a PC with 2 GB RAM and 2.9x MHz processor. The software used for development areJDK 1.6 (Java Standard Edition), and Net Beans IDE. The data sets are taken from UCI public repository.

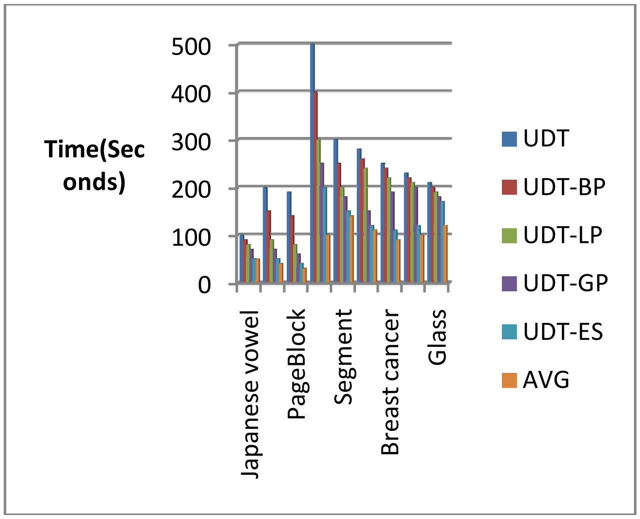


Fig. 5 – Execution time

Execution time for various data sets on different algorithms is shown graphically in fig. 5. For each data set six columns are drawn. Execution time is plotted in Y axis while the datasets are presented in X axis. Execution time is also given for AVG algorithm. The ascending order of efficiency is with algorithms such as UDT, UDT-BP, UDT-LP, UDT-GP, and UDTES. This reveals the success of pruning techniques described in section 5.

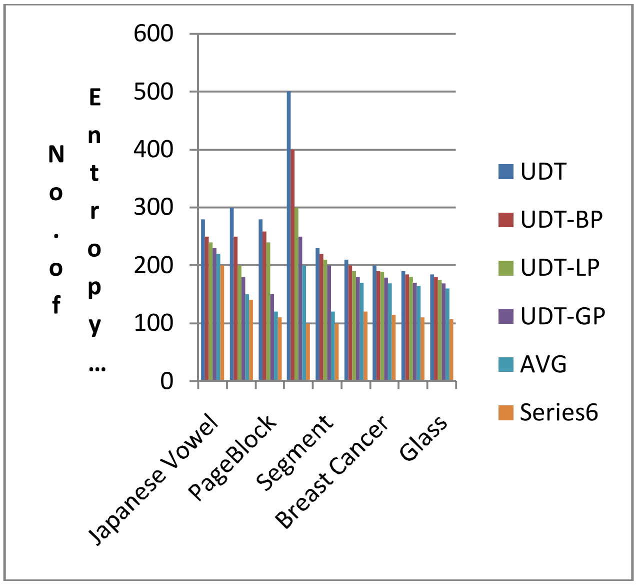


Fig. 6 – Pruning effectiveness

Pruning techniques improve efficiency of decision tree algorithms. The pruning techniques used in this paper are described in section 5.

The effectiveness of pruning techniques for various algorithms on given data sets is presented in fig. 6. In horizontal axis datasets are presented while the vertical axis represents the number of entropy calculations required.

6. Conclusion

This paper extends the existing decision tree classifiers to make them work for uncertain data. Uncertain data is the data that is not precise. For instance the data of a column is represented by multiple values. The traditional decision tree classifiers are modified to obtain decision trees for uncertain data. We have discovered a fact that using averages or mean values does not yield in accuracy of decision trees. However, the computation of PDFs makes the classifier more accurate. However, calculating PDFs is computationally expensive as it happens to process large amount of data. To overcome this problem we have use many pruning techniques that reduce computational cost. The experimental results revealed that our algorithms are highly effective and decision trees obtained from uncertain data are highly accurate.

References

[1] R. Agrawal, T. Imielinski, and A.N. Swami, "Database Mining: APerformance Perspective," IEEE Trans. Knowledge and Data Eng.,vol. 5, no. 6, pp. 914-925, Dec. 1993.

[2] J.R. Quinlan, C4.5: Programs for Machine Learning. MorganKaufmann, 1993.

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

[4] E. Hung, L. Getoor, and V.S. Subrahmanian, "ProbabilisticInterval XML," ACM Trans. Computational Logic (TOCL), vol. 8,no. 4, 2007.TSANG ET AL.: DECISION TREES FOR UNCERTAIN DATA 77.

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

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

[7] R. Cheng, Y. Xia, S. Prabhakar, R. Shah, and J.S. Vitter, "EfficientIndexing Methods for Probabilistic Threshold Queries overUncertain Data," Proc. Int'l Conf. Very Large Data Bases (VLDB),pp. 876-887, Aug./Sept. 2004.

[8] R. Cheng, D.V. Kalashnikov, and S. Prabhakar, "QueryingImprecise Data in Moving Object Environments," IEEE Trans.Knowledge and Data Eng., vol. 16, no. 9, pp. 1112-1127, Sept. 2004.

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

[10] S.D. Lee, B. Kao, and R. Cheng, "Reducing UK-Means to KMeans,"Proc. First Workshop Data Mining of Uncertain Data(DUNE), in conjunction with the Seventh IEEE Int'l Conf. Data Mining (ICDM), Oct. 2007.

[11] W.K. Ngai, B. Kao, C.K. Chui, R. Cheng, M. Chau, and K.Y. Yip,"Efficient Clustering of Uncertain Data," Proc. Int'l Conf. DataMining (ICDM), pp. 436-445, Dec. 2006.

[12] C.C. Aggarwal, "On Density Based Transforms for Uncertain DataMining," Proc. Int'l Conf. Data Eng. (ICDE), pp. 866-875, Apr. 2007.

[13] C.K. Chui, B. Kao, and E. Hung, "Mining Frequent Itemsets fromUncertain Data," Proc. Pacific-Asia Conf. Knowledge Discovery andData Mining (PAKDD), pp. 47-58, May 2007.

[14] O.O. Lobo and M. Numao, "Ordered Estimation of MissingValues," Proc. Pacific-Asia Conf. Knowledge Discovery and DataMining (PAKDD), pp. 499-503, Apr. 1999.

[15] C.Z. Janikow, "Fuzzy Decision Trees: Issues and Methods," IEEETrans. Systems, Man, and Cybernetics, Part B, vol. 28, no. 1, pp. 1-14,Feb. 1998.

[16] Y. Yuan and M.J. Shaw, "Induction of Fuzzy Decision Trees,"Fuzzy Sets and Systems, vol. 69, no. 2, pp. 125-139, 1995.

[17] C. Olaru and L. Wehenkel, "A Complete Fuzzy Decision TreeTechnique," Fuzzy Sets and Systems, vol. 138, no. 2, pp. 221-254,2003.

[18] T. Elomaa and J. Rousu, "General and Efficient Multisplitting ofNumerical Attributes," Machine Learning, vol. 36, no. 3, pp. 201-244, 1999.

[19] U.M. Fayyad and K.B. Irani, "On the Handling of Continuous-Valued Attributes in Decision Tree Generation," Machine Learning,vol. 8, pp. 87-102, 1992.

[20] T. Elomaa and J. Rousu, "Efficient Multisplitting Revisited:Optima-Preserving Elimination of Partition Candidates," DataMining and Knowledge Discovery, vol. 8, no. 2, pp. 97-126, 2004.

[21] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone, Classificationand Regression Trees. Wadsworth, 1984.

[22] M. Umanol, H. Okamoto, I. Hatono, H. Tamur a, F. Kawachi, S.Umedzu, and J. Kinoshita, "Fuzzy Decision Trees by Fuzzy ID3Algorithm and Its Application to Diagnosis Systems," Proc. IEEEConf. Fuzzy Systems, IEEE World Congress Computational.

[23] Smith Tsang, Ben Kao, Kevin Y. Yip, Wai-Shing Ho, and Sau Dan Lee, "Decision Trees for Uncertain Data", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 1, JANUARY 2011.