A Framework On: Decision Tree for Dynamic Uncertain Data

Megha Pimpalkar[1], **Garima Singh**[2]

[1]Student, Department of Computer Technology, WCEM, Nagpur, India

[2]Assistance Professor, Department of Computer Technology, WCEM, Nagpur, India

Abstract: ***Current research on data stream classification mainly focuses on certain data, in which precise and definite value is usually assumed. However, data with uncertainty is quite natural in real-world application due to various causes, including imprecise measurement, repeated sampling and network errors. In this paper, a new approach is proposed to construct a fuzzy decision tree (FDT) when the training set is built incrementally and when training examples are provided temporally. In this paper, we focus on uncertain data stream classification. Based on DTDU, we propose our DTDU (Decision Tree for Dynamic Uncertain Data) algorithm. Experimental study shows that the proposed DTDU algorithm is efficient in classifying dynamic data stream with uncertain numerical attribute and it is computationally efficient.***

Keywords: Uncertain data streams, Decision Tree, Classification, Fuzzy decision tree, Fractional samples

1. Introduction

In real-world applications, like credit fraud detection, network intrusion detection, vast volume of knowledge arrives ceaselessly with high speed. Such applications may well be sculptured as knowledge stream classification issues. Currently, attribute worth is sometimes assumed to be precise and definite by the analysis community of knowledge stream classification analysis. However, knowledge uncertainty arises naturally in several applications owing to varied reasons, as well as imprecise measuring, missing values, privacy protection, continual sampling and network errors, thus ancient knowledge stream learning algorithms aren't applicable to real-life applications.

In this paper, our aim is to provide a method to focus the training of the fuzzy decision tree on actual information (topics, words, values of attributes,...) and to let the tree “forget” old information. Thus, training examples are not equivalent in the training set: old examples could be considered as less important than actual and new examples. We face here dynamic data that arrive over time to increase the training set and that should be taken into account in a more important way that older data present in the training set.

2. Literature Survey

Classification

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. 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

Decision Tree

Decision trees may be a straightforward however wide used methodology for classification and predictive modelling. a decision tree partitions information into smaller segments referred to as terminal nodes. every terminal node is appointed a category label. The non-terminal nodes, which embrace the foundation and different internal nodes, contain attribute take a look at conditions to separate records that have totally different characteristics. The partitioning method terminates once the subsets can not be partitioned off any more victimisation predefined criteria. Decision trees are utilized in several domains. as an example, in information promoting, decision trees may be wont to phase teams clients|of consumers|of shoppers} and develop customer profiles to assist marketers turn out targeted promotions that reach higher response rates.

Data Uncertainty

This paper studies decision tree based classification methods for uncertain data. In many applications, data contains inherent uncertainty. A number of factors contribute to the uncertainty, such as the random nature of the physical data generation and collection process, measurement and decision errors, unreliable data transmission and data staling. For example, there are massive amounts of uncertain data in sensor networks, such as temperature, humidity, and pressure. Uncertainty can also arise in categorical data.

Here we focus on the attributes uncertainty and assume the class type is certain. When the value of a numerical attribute is called an uncertain numerical attribute (UNA), denoted by $A_i^{u_n}$. Further, we use $A_{i_j}^{u_n}$ to denote the jth instance of $A_i^{u_n}$. The value of $A_i^{u_n}$ is represented as a range or interval and the probability distribution function (PDF) over this range. Note that $A_i^{u_n}$is treated as a continuous random variable. The PDF f(x) can be related to an attribute if all instances have the same distribution, or related to each instance if each instance has a different distribution. Since data uncertainty is ubiquitous, it is important to develop classification models for uncertain data. We choose the decision tree because of its numerous positive features. Decision tree is simple to understand and interpret. It requires little data preparation.

Fuzzy Decision Tree

A popular extension of decision trees is the fuzzy decision trees (FDT) that proposes the use of fuzzy sets theory to handle
numerical or fuzzy data in the description of examples from the training set. However, few FDT related works have been done to handle streams of data when the training set of examples is provided not globally but sequentially. That kind of temporal or dynamic data are very popular nowadays in a lot of Big data applications and research on new approaches to handle them is a very hot topic nowadays.

Fractional Sample

Note that uncertain tree growing is a recursive process of partitioning the training samples into fractional samples. In this paper, we also adopt the technique of fractional sample to handle uncertainty.
Let split attribute at node N is denoted by $X_i^{u_c}$.Sample S_t is split into new samples $\{S_{t1,}S_{t2}, \ldots .S_{t_m}\}$by $X_i^{u_c}$.Thus S_{t_j} is fractional sample of S_t on attribute $X_i^{u_c}$.

Uncertain Information Gain

Information Gain (IG) is widely used in decision tree algorithms to decide which attribute being selected as the next splitting attribute. Base on DTU, for a data set S denoted by node N and an attribute set X^{u_c}= $\{X_1^{u_c}, X_2^{u_c}, \ldots X_d^{u_c}\}$.

UIG(S, $X_i^{u_c}$)=Entropy(S)-$\sum_{j=1}^{m} \frac{PC\ (s_{tj})}{PC(S)}$*Entropy($s_{tj}$)

Here m=|Dom$X_i^{u_c}$)|

Entropy(S)= -$\sum_{k=1}^{|C|} Ps(y_k) \log Ps(y_k)$

Here $PS(yk)$=the ratio of the sum of probabilities of each sample in *yk* to the sum of probabilities of each sample.

3. Proposed Algorithm

DTDU

1. The new coming sample is associated with ID and a weight, then it is saved to the sliding window
2. An outdated sample is removed from the tree, it is also deleted from the slide window
3. Tree growing is performed, DTDUGrow
4. The split validity of an internal node is checked periodically

DTDUGrow

For growing uncertain decision tree
1. Sufficient statistics are collected from the fractional samples
2. A sample is split into fractional samples
3. Uncertain information gain is calculated using sufficient statistics at leaf nodes
4. Split attribute is chosen and leaf node is split into an internal node

Classify Sample

HT: a decision tree for uncertain data
St: a test sample
dist: probability distribution over c
1. if l is leaf node of HT then dist=dist+l.dist
2. else split St into fractional sample E and for each Stj in E, let lj be the branch child for Stj, ClassifySample(lj,Stj)

4. Result and Analysis

The following information shows various results of this paper. The DTDU algorithm is run for three different databases, and the result and the accuracy of the algorithm is as follows:

```
Classification Started
Total Records Tested : 214
Correctly Classify : 201
Classification Accuracy (%): 93.92523364485982
Time Elapsed For Classification (ms) : 605

Classification Started
Total Records Tested : 768
Correctly Classify : 733
Classification Accuracy (%): 95.44270833333334
Time Elapsed For Classification (ms) : 12

Classification Started
Total Records Tested : 178
Correctly Classify : 166
Classification Accuracy (%): 93.25842696629213
Time Elapsed For Classification (ms) : 3959
```

5. Conclusion

In this paper, we focused on uncertain data stream classification. Based on DTDU, we propose our DTDU (Decision Tree for Dynamic Uncertain Data) algorithm. Experimental study shows that the proposed DTDU algorithm is efficient in classifying dynamic data stream with uncertain numerical attribute and it is computationally efficient.

6. Future Scope

We can use ensemble learning in future. An ensemble of classifiers is a set of classifiers whose individual decisions are combined in some way (typically by weighted or unweighted voting) to classify new examples.

References

[1] Christophe Marsala,” Fuzzy Decision Trees for Dynamic Data” IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS),2013

[2] Smith Tsang, Ben Kao, Kevin Y. Yip, Wai-ShingHo(2011), AndSau Dan Lee “Decision Trees For Uncertain Data”, IEEE Transactions On Knowledge And Data Engineering, Vol. 23, No. 1, January.

[3] Varshachoudhary,PranitaJain”Classification: A Decision Tree For Uncertain Data Using CDF”, International Journal Of Engineering Research And

Applications (IJERA) ISSN: 2248-9622 Vol. 3, Issue 1, Pp.1501-1506,2013

[4] M. Suresh Krishna Reddy1, R. Jayasree" Extending Decision Tree Clasifiers For Uncertain Data", International Journal Of Engineering Science & Advanced Technology ISSN: 2250–3676 Volume-2, Issue-4, 1030 – 1034,2011

[5] C. Marsala and B. Bouchon-Meunier, "An adaptable system to construct fuzzy decision trees," in Proc. of the NAFIPS'99 (North American Fuzzy Information Processing Society), New York, USA, pp. 223–227,1911

[6] Pragati Pandey , Miss Prateeksha Pandey, Mrs. MriduSahu," Mining Uncertain Data Using Classification Feature Decision Trees", ISSN: 2277 – 9043 International Journal Of Advanced Research In Computer Science And Electronics Engineering Volume 1, Issue 3,2012

[7] Chunquan Liang, Yang Zhang "Decision Tree For Dynamic And Uncertain Data Streams" JMLR: Workshop And Conference Proceedings 13: 209-224 2nd Asian Conference On Machine Learning (ACML2010), Tokyo, Japan, Nov. 2010.

[8] Swapnil Andhariya, Khushali Mistry, Prof.SahistaMachchhar, Prof. Dhruv Dave "Prodtu: A Novel Probabilistic Approach To Classify Uncertain Data Using decision Tree Induction" International Journal Of Engineering Research & Technology (IJERT) ISSN: 2278-0181Vol. 2 Issue 6, June – 2013

[9] Charu C. Aggarwal, Philip S. Yu "A Survey Of Uncertain Data Algorithms And Applications" Ieee Transactions On Knowledge And Data Engineering, Vol. 21, No. 5, May 2009

[10] Margaret H. Dunham,"Data Mining-Introductory And Advanced Topics" PearsonEducation,SixtyhImnpression,2009.