© 2018 IJSRST | Volume 4 | Issue 2 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X
Themed Section: Science and Technology

Decision Tree Analysis Tool with the Design Approach of Probability Density Function towards Uncertain Data Classification

Siripuri Kiran
Assistant Professor Kakatiya Institute Of Technology and Sciences. Warangal, Telangana, India

ABSTRACT

Traditional decision tree classifiers are built utilizing certain or point data as it were. Be that as it may, in numerous genuine applications innately data is constantly uncertain. Quality or esteem uncertainty is characteristically connected with data esteems amid data gathering process. Traits in the preparation data sets are of two kinds – numerical (constant) and clear cut (discrete) characteristics. Data uncertainty exists in both numerical and all out characteristics. Data uncertainty in numerical qualities implies scope of qualities and data uncertainty in all out traits implies set or accumulation of qualities. In this paper we propose a technique for taking care of data uncertainty in numerical properties. One of the least difficult and most straightforward techniques for taking care of data uncertainty in numerical properties is finding the mean or normal or agent estimation of the arrangement of unique estimations of each estimation of a characteristic. With data uncertainty the estimation of a property is generally spoken to by an arrangement of qualities. Decision tree classification precision is tremendously enhanced when property estimations are spoken to by sets of esteems as opposed to one single delegate esteem. Probability density function with equal probabilities is one compelling data uncertainty demonstrating system to speak to each estimation of a property as an arrangement of qualities. Here the principle presumption is that genuine esteems gave in the preparation data sets are found the middle value of or delegate estimations of initially gathered esteems through data accumulation process. For every illustrative estimation of each numerical characteristic in the preparation data set, approximated values relating to the initially gathered esteems are created by utilizing probability density function with equal probabilities and these recently produced sets of qualities are utilized as a part of developing another decision tree classifier.

Keywords : Probability Density Function, Data Mining, Classification, Uncertain Data, Decision Tree, Machine Learning.

I. INTRODUCTION

Classification is a data investigation method. Decision tree is a capable and well known instrument for classification and forecast yet decision trees are predominantly utilized for classification [1]. Primary favorable position of decision tree is its interpretability – the decision tree can without much of a stretch be changed over to an arrangement of IF-THEN decides that are effortlessly justifiable [2].

Cases wellsprings of data uncertainty incorporate estimation/quantization blunders, data staleness, and various rehashed estimations [3]. Data mining applications for uncertain data are – classification of uncertain data, bunching of uncertain data, visit design mining, exception discovery and so forth. Cases for specific data are – areas of colleges, structures, schools, universities, eateries, railroad stations and transport stands and so on. Data uncertainty normally emerges in an extensive

number of genuine applications including logical data, web data joining, machine learning, and data recovery.

Data uncertainty in databases is comprehensively grouped into three kinds:

1. Attribute or esteem uncertainty
2. Correlated uncertainty and
3. Existential or Tuple uncertainty

In characteristic or esteem uncertainty, the estimation of each trait is spoken to by an autonomous probability dissemination. In related uncertainty, estimations of numerous properties might be portrayed by a joint probability dispersion. For instance, a data tuple in a social database could be related with a probability that speaks to the certainty of its essence [3]. On account of existential uncertainty, a data tuple could possibly exist in the social database. For instance, expect we have the accompanying tuple from a probabilistic database [(a,0.4),(b,0.5)] the tuple has 10% shot of not existing in the social database. Initially gathered right esteems are around recovered by utilizing probability density function demonstrating procedure with equal probabilities. Subsequently, probability density function displaying system models data uncertainty suitably.

At the point when data mining is performed on uncertain data, distinctive composes uncertain data demonstrating procedures must be considered keeping in mind the end goal to acquire brilliant data mining comes about. One of the present difficulties in the field of data mining is to grow great data mining strategies to break down uncertain data. That is, one of the present difficulties with respect to the advancement of data mining procedures is the capacity to oversee data uncertainty.

II. ESTABLISHMENT OF UNCERTAIN DATA

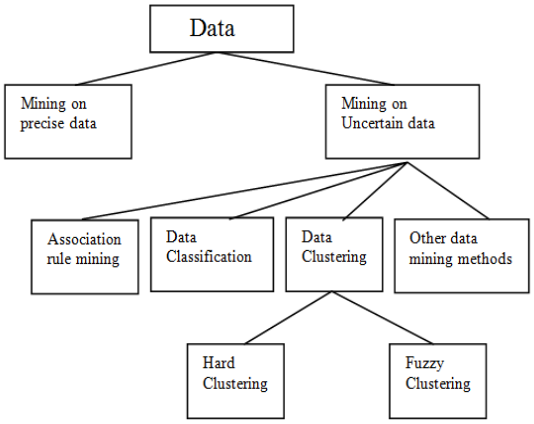

Figure 1. Taxonomy of Uncertain Data Mining

Numerous genuine applications contain uncertain data. With data uncertainty data esteems are not any more nuclear or certain. Data is regularly connected with uncertainty due to estimation mistakes, testing blunders, rehashed estimations, and obsolete data sources [3]. For safeguarding security some of the time certain data esteems are expressly changed to scope of qualities. For instance, for protecting security the specific estimation of the genuine age of a man is spoken to as a range [16, 26] or 16 – 26.

Table 1. Example on Categorical Uncertain and Numerical Uncertain attributes.

Tuple No.	Marks	Result	Class Label
1	550 – 600	(0.8,0.1,0.1,0.0)	(0.8,0.2)
2	222 – 444	(0.6,0.2,0.1,0.1)	(0.5,0.5)
3	470 – 580	(0.7,0.2,0.1,0.0)	(0.9,0.1)
4	123- 290	(0.4,0.2,0.3,0.1)	(0.7,0.3)
5	345 – 456	(0.6,0.2,0.1,0.1)	(0.8,0.2)
6	111 – 333	(0.3,0.3,0.2,0.2)	(0.9,0.1)
7	200 – 280	(0.3,0.3,0.2,0.2)	(0.7,0.3)
8	500 – 580	(0.7,0.2,0.1,0.0)	(0.5,0.5)
9	530 – 590	(0.7,0.3,0.0,0.0)	(0.6,0.4)
10	450 – 550	(0.7,0.2,0.1,0.0)	(0.4,0.6)
11	150 – 250	(0.3,0.3,0.2,0.2)	(0.2,0.8)
12	180 – 260	(0.4,0.2,0.2,0.2)	(0.1,0.9)

Imprints characteristic is a numerical uncertain attribute (NUA) and Result quality is a clear categorical uncertain attribute(CUA). Class name can likewise be either numerical or unmitigated.

III. PROBLEM DEFINITION

In numerous genuine applications data can't be in a perfect world spoke to by point data as it were. All decision tree calculations so far created depended on specific data esteems display in the numerical qualities of the preparation data sets. These estimations of the preparation data sets are the agents of the initially gathered data esteems. Data uncertainty was not considered amid the advancement of numerous data mining calculations including decision tree classification system. In this way, there is no classification system which handles the uncertain data. This is the issue with the current certain (conventional or established) decision tree classifiers. Data uncertainty is generally demonstrated by a probability density function and probability density function is spoken to by an arrangement of qualities as opposed to one single agent or normal or total esteem. Subsequently, in uncertain data administration, preparing data tuples are regularly spoken to by probability conveyances instead of deterministic esteems.

At present existing decision tree classifiers consider estimations of qualities in the tuples with known and exact point data esteems as it were. In actuality, the data esteems characteristically experience the ill effects of significant worth uncertainty (or trait uncertainty). Thus, certain (customary or established) decision tree classifiers create mistaken or less exact data mining comes about. A preparation data set can have both Uncertain Numerical Attributes (UNAs) and Uncertain Categorical Attributes (UCAs) and both preparing tuples contain uncertain data. As data uncertainty generally exists, all things considered, it is imperative to create exact and more effective data mining methods for uncertain data administration.
The present examination proposes a calculation called Decision Tree classifier development on Uncertain Data (DTU) to enhance execution of Certain Decision Tree (CDT). DTU utilizes probability density function with equal probabilities in demonstrating data uncertainty in the estimations of numerical qualities of preparing data sets. The execution of these two calculations is analyzed tentatively through reenactment. The execution of DTU is ends up being better.

IV. EXISTING ALGORITHM

The specific decision tree (CDT) calculation develops a decision tree classifier by part every node into left and right nodes. At first, the root node contains all the preparation tuples. The way toward dividing the preparation data tuples in a node into two nodes in view of the best split point esteem of the best split and putting away the subsequent tuples in its left and right nodes is alluded to as part. At whatever point additionally split of a node isn't required then it turns into a leaf node alluded to as an outside node. The part procedure at each inner node is done recursively until the point that no further split is required. Constant esteemed qualities must be discretized before property selection [7]. Additionally part of an inward node is halted if all the tuples in an inside node have a similar class name or part does not come about nonempty left and right nodes. Amid decision tree development within each inside node just fresh and deterministic tests are connected. Entropy is a metric or function that is utilized to discover the level of scattering of preparing data tuples in a node. In decision tree development the decency of a split is evaluated by a contamination measure [2]. One conceivable function to gauge polluting influence is entropy [2]. Entropy is a data based measure and it is construct just in light of the extents of tuples of each class in the preparation data set. Entropy is taken as

scattering measure since it is transcendently utilized for building decision trees. In a large portion of the cases, entropy finds the best split and adjusted node sizes after split such that both left and right nodes are however much unadulterated as could reasonably be expected. Exactness and execution time of certain decision tree (CDT) calculation for 9 data sets are appeared in Table 2

V. DECISION TREE CLASSIFICATION ON UNCERTAIN DATA (DTU) ALGORITHM(PSEUDO CODE)

Uncertain_Data_Decision_Tree(T)

1. If all the tuples in node T have the
2. same class label then
3. set
4. return(T)
5. If tuples in node T will have more than one class then
6. Find_Best_Split(T)
7. For i ← 1 to datasize[T] do
8. If split_atribute_value[t_i] <= split_point[T] then
9. Add tuple t_i to left[T]
10. Else
11. Add tuple t_i to right[T]
12. If left[T] = NIL or right[T] = NIL then
13. Create empirical probability distribution of the node T
14. return(T)
15. If left[T] != NIL and right[T] != NIL then
16. UNCERTAIN_DATA_DECISION_TREE(left[T])
17. UNCERTAIN_DATA_DECISION_TREE(right[T])
18. return(T)

DTU can fabricate more exact decision tree classifiers however computational many-sided quality of DTU is 'n' times costly than CDT. Henceforth, DTU isn't proficient as that of CDT. To the extent precision is viewed as DTU classifier is more exact however to the extent productivity is viewed as certain decision tree (CDT) classifier is more effective. To decrease the computational multifaceted nature DTU classifier we have proposed another pruning strategy with the goal that entropy is ascertained just thinking optimistically point for every interim. Subsequently, the new approach, pruned form, PDTU, for decision treeconstruction is more precise with roughly same computational unpredictability as that of CDT. Exactness and execution time of certain decision tree (CDT) classifier calculation for 9 preparing data sets are appeared in Table 6.2 and precision and execution time of decision tree classification on uncertain data (DTU) classifier calculation for 9 preparing data sets are appeared in Table 6.3 and correlation of execution time and precision for certain decision tree (CDT) and DTU calculations for 9 preparing data sets are appeared in Table 6.4 and diagrammed in Figure 6.1 and Figure 2 separately.

VI. EXPERIMENTAL RESULTS

A simulation display is created for assessing the execution of two calculations – Certain Decision Tree (CDT) classifier and Decision Tree classification on Uncertain Data (DTU) classifier tentatively. The preparation data sets appeared in Table 6.1 from University of California (UCI) Machine Learning Repository are utilized for assessing the execution and exactness of the above said calculations.

Table 2. Data Sets from the UCI Machine Learning Repository

No	Data Set Name	Training Tuples	No. Of Attributes	No. Of Classes
1	Iris	150	4	3
2	Glass	214	9	6
3	Ionosphere	351	32	2
4	Breast	569	30	2
5	Vehicle	846	18	4
6	Segment	2310	14	7
7	Satellite	4435	36	6
8	Page	5473	10	5
9	Pen Digits	7494	16	10

In every one of our examinations we have utilized preparing data sets from the UCI Machine Learning Repository [6]. The simulation demonstrate is executed in Java 1.7 on a Personal Computer with 3.22 GHz Pentium Dual Core processor (CPU), and 2GB of fundamental memory (RAM). The execution measures, precision and execution time, for the above said calculations are introduced in Table 6.2 to Table 6.4 and Figure 6.1 to Figure 6.2.

Table 6.2. Certain Decision Tree (CDT) Accuracy and Execution Time

No	Data Set Name	Total Tuples	Accuracy	Execution Time
1	Iris	150	97.4422	1.1
2	Glass	214	88.4215	1.3
3	Ionosphere	351	84.4529	1.47
4	Breast	569	96.9614	2.5678
5	Vehicle	846	78.9476	6.9
6	Segment	2310	97.0121	29.4567
7	Satellite	4435	83.94	153.234
8	Page	5473	97.8762	36.4526
9	Pen Digits	7494	90.2496	656.164

Table 6.3. Uncertain Decision Tree (DTU) Accuracy and Execution Time

No	Data Set Name	Total Tuples	Accuracy	Execution Time
1	Iris	150	98.5666	1.1
2	Glass	214	95.96	1.2
3	Ionosphere	351	98.128	16.504
4	Breast	569	97.345	24.223
5	Vehicle	846	96.01281	35.365
6	Segment	2310	98.122	212.879
7	Satellite	4435	85.891	294.96
8	Page	5473	98.8765	289.232
9	Pen Digits	7494	91.996	899.3491

Table 6.4. Comparison of accuracy and execution times of CDT and DTU

No	Data Set Name	CDT Accuracy	DTUAccuracy	CDT Execution Time	DTU Execution Time
1	Iris	97.4422	98.566	1.1	1.1
2	Glass	88.4215	95.96	1.3	1.2
3	Ionosphere	84.4529	98.128	1.47	16.504
4	Breast	96.9614	97.345	2.5678	24.223
5	Vehicle	78.9476	96.281	6.9	35.365
6	Segment	97.0121	98.122	29.4567	212.879
7	Satellite	83.94	85.891	153.234	294.96
8	Page	97.8762	98.847	36.4526	289.232
9	Pen Digits	90.2496	91.996	656.164	899.434

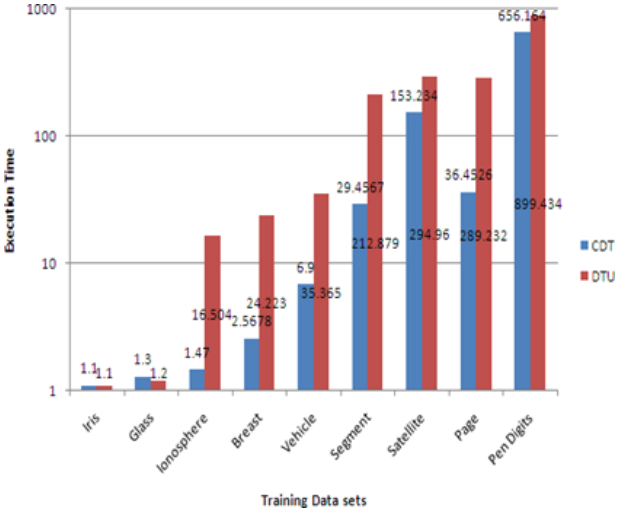

Figure 6.1. Correlation of execution times of DTU and CDT

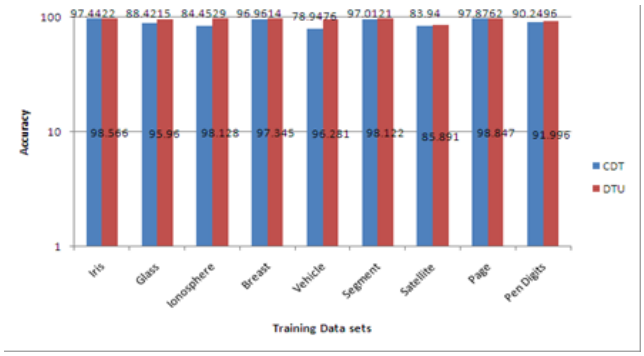

Figure 6.2. Comparison of Classification Accuracies of CDT and DTU

VII. CONCLUSION

7.1 Contributions

The execution of existing traditional or established or certain decision tree (CDT) is confirmed tentatively through simulation. Another decision tree classifier construction calculation called Decision Tree Classification on Uncertain Data (DTU) is proposed and contrasted and the current Certain Decision Tree classifier (CDT). Tentatively it is discovered that the classification precision of proposed calculation DTU is greatly improved than CDT calculation.

7.2. Suggestions for future work

Exceptional methods or thoughts or plans are expected to deal with various sorts of data uncertainties display in the preparation data sets. Unique techniques are expected to deal with data uncertainty in straight out properties moreover. Unique pruning systems are expected to diminish execution time of Decision Tree Classification on Uncertain Data (DTU). Likewise unique systems are expected to discover and adjust arbitrary commotion and different blunders in the all out properties.

VIII. REFERENCES

[1]. Jiawei Han, Micheline Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, second edition,2006. pp.285–292

[2]. Introduction to Machine Learning Ethem Alpaydin PHI MIT Press, second edition. pp. 185–188

[3]. 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

[4]. Hsiao-Wei Hu, Yen-Liang Chen, and Kwei Tang "A Dynamic Discretization Approach for Constructing Decision Trees with a Continuous Label" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,VOL.21,No.11,NOVEMBER 2009

[5]. R.E. Walpole and R.H. Myers, Probability and Statistics for Engineers and Scientists. Macmillan Publishing Company, 1993.

[6]. Shoban Babu Sriramoju, "Allocated Greater Order Organization of Rule Mining utilizing Information Produced Through Textual facts" in "International Journal of Information Technology and management" Vol-I, Issue-I, August 2011 [ISSN : 2249-4510]

[7]. A. Asuncion and D. Newman, UCI Machine Learning Repository, http://www.ics.uci.edu/mlearn/MLRepository.html, 2007.

[8]. 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, 1996.

[9]. Shoban Babu Sriramoju, "Analysis and Comparison of Anonymous Techniques for Privacy Preserving in Big Data" in "International Journal of Advanced Research in Computer and Communication Engineering", Vol 6, Issue 12, December 2017, DOI 10.17148/IJARCCE.2017.61212 [ISSN(online) : 2278-1021, ISSN(print) : 2319-5940]

[10]. Guguloth Vijaya, A. Devaki, Dr. Shoban Babu Sriramoju, "A Framework for Solving Identity Disclosure Problem in Collaborative Data Publishing" in "International Journal of Research and Applications" (Apr-Jun © 2015 Transactions), Vol 2, Issue 6, 292-295

[11]. Prof. Mangesh Ingle, Prof. Ashish Mahalle, Dr. Shoban Babu, "HLA Based solution for Packet Loss Detection in Mobile Ad Hoc Networks" in "International Journal of Research in Science and Engineering" Vol 3, Issue 4,July-August 2017 [ISSN : 2394-8299].

[12]. Sriramoju Ajay Babu, Dr. S. Shoban Babu, "Improving Quality of Content Based Image Retrieval with Graph Based Ranking" in "International Journal of Research and Applications" Vol 1, Issue 1,Jan-Mar 2014 [ISSN : 2349-0020].

[13]. Mounika Reddy, Avula Deepak, Ekkati Kalyani Dharavath, Kranthi Gande, Shoban Sriramoju, "Risk-Aware Response Answer for Mitigating Painter Routing Attacks" in "International Journal of Information Technology and Management" Vol VI, Issue I, Feb 2014 [ISSN : 2249-4510]

[14]. Mounica Doosetty, Keerthi Kodakandla, Ashok R, Shoban Babu Sriramoju, "Extensive Secure Cloud Storage System Supporting Privacy-Preserving Public Auditing" in "International Journal of Information Technology and Management" Vol VI, Issue I, Feb 2012 [ISSN : 2249-4510]

[15]. Shoban Babu Sriramoju, "An Application for Annotating Web Search Results" in "International Journal of Innovative Research in Computer and Communication Engineering" Vol 2,Issue 3,March 2014

[16]. [ISSN(online) : 2320-9801, ISSN(print) : 2320-9798]

[17]. Shoban Babu Sriramoju, "Multi View Point Measure for Achieving Highest Intra-Cluster Similarity" in "International Journal of Innovative Research in Computer and Communication Engineering" Vol 2,Issue 3,March 2014

[18]. [ISSN(online) : 2320-9801, ISSN(print) : 2320-9798]

[19]. Dr. Atul Kumar , Shoban Babu Sriramoju, "An Analysis around the study of Distributed Data Mining Method in the Grid Environment : Technique, Algorithms and Services" in "Journal of Advances in Science and Technology" Vol-IV, Issue No-VII, November 2012 [ISSN : 2230-9659]

[20]. Dr. Atul Kumar, Shoban Babu Sriramoju, "An Analysis on Effective, Precise and Privacy Preserving Data Mining Association Rules with Partitioning on Distributed Databases" in "International Journal of Information Technology and management" Vol-III, Issue-I, August 2012 [ISSN : 2249-4510]

[21]. Azmera Chandu Naik, N.Samba Siva Rao , Shoban Babu Sriramoju, "Predicting The Misusability Of Data From Malicious Insiders" in "International Journal of Computer Engineering and Applications" Vol V,Issue II,Febrauary 2014 [ISSN : 2321-3469]

[22]. Ajay Babu Sriramoju, Dr. S. Shoban Babu, "Study of Multiplexing Space and Focal Surfaces and Automultiscopic Displays for Image Processing" in "International Journal of

Information Technology and Management" Vol V, Issue I, August 2013 [ISSN : 2249-4510]

[23]. Shoban Babu Sriramoju, Dr. Atul Kumar, "A Competent Strategy Regarding Relationship of Rule Mining on Distributed Database Algorithm" in "Journal of Advances in Science and Technology" Vol-II, Issue No-II, November 2011 [ISSN : 2230-9659]