International Journal of Multidisciplinary Research and Publications
ISSN (Online): 2581-6187

Implementation with Performance Evaluation of Decision Tree Classifier for Uncertain Data: Literature Review

Abdarrouf Abdalati Abdalgader Saed[1], Adam Amril Jaharadak[1]
[1]School of Graduate Studies, Post Graduate Centre, Management and Science University, University Drive, Off Persiaran Olahraga, Section 13, 40100, Selangor, Malaysia
Email address: [1]asra1abdo@gmail.com

***Abstract*—** *To extract meaningful and non-negligible facts from large amounts of data for the extraction of patterns, anomalies, and correspondence information from large databases, data mining is used. Uncertain Data Implementation and Decision Tree Classifier Performance Evaluation. The study's goal is to build a decision tree from uncertain data, and existing systems have a number of limitations that need to be investigated further and resolved. Measurement errors, stale data, and repeated measurements all contribute to data uncertainty. There are numerous problems with classification, and this applies across a wide range of data mining applications. Data classification using decision trees is very popular because of their simple and robust structure. The accuracy of the decision tree for the uncertain data used is high because appropriate pdfs have been used. Improve the efficiency of a constructed tree by employing various pruning techniques. In comparison to other techniques, the proposed decision tree for uncertain data achieves higher efficiency. For the construction of the decision tree, this method uses classical algorithms that generate enormous numbers of data tuples (one for each decision). The proposed method achieves a better result because the execution time is shorter, and the system's efficiency is higher. The proposed work will be extended in the future to improve the data classifiers' pruning efficiency when building decision trees. This lays the groundwork for the rest of the research project.*

***Keywords*—** *Implementation, Performance, Evaluation, Decision Tree Classifier, Uncertain Data, Literature Review*.

I. INTRODUCTION

Data mining can be defined as the process of mining a large database for patterns, anomalies, and correspondence information. From huge amounts of data, extract the important, non-negligible facts in various fields such as business, communications, and engineering. Analyze the likelihood of future events and separate the data using a sophisticated mathematical algorithm. Discovery of Knowledge in Database is the process of validly finding new and interesting patterns in large, complex data sets (KDD). Knowledge Discovery in Data is another name for data mining (KDD). The data mining block diagram is shown in Figure. The data mining process can be broken down into five distinct phases. Preprocessing refers to gathering and preparing raw data for further processing such as data mining and data transformation. The knowledge-based discovery was made possible by the analysed data. In the decision-making process, patterns found in the dataset are used to predict and classify new data. There are several tasks involved in finding hidden patterns in the database, including frequent pattern mining, weighted pattern mining, and pattern mining with high utility. Most of these methods are used for transactional databases, but they can also be applied to streaming databases and other types of databases. Tools that analyse unknown patterns are used to support various applications like banking, customer relationship management (CRM), targeted marketing (TM), fraud detection (FDD), pharmaceuticals (PDUs), and web assortment schemes in DM techniques and algorithms.

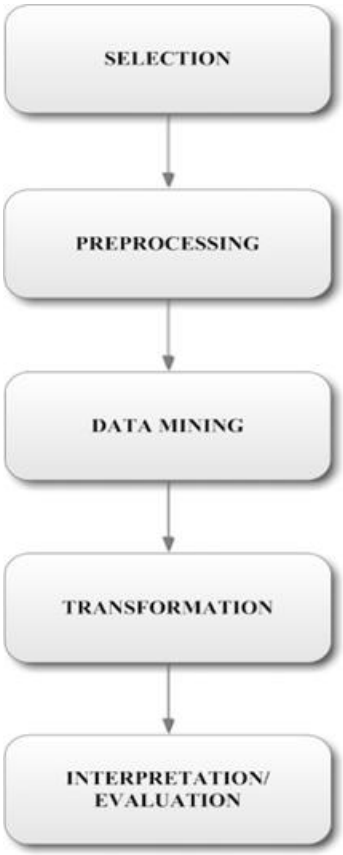

Fig. 1. Block diagram of Data Mining

Data classification techniques are widely used in data mining to sort the data into different categories. Classification techniques introduced to a tuple can now be quickly identified by their types and groups, which is important for data mining.

Abdarrouf Abdalati Abdalgader Saed and Adam Amril Jaharadak, "Implementation with Performance Evaluation of Decision Tree Classifier for Uncertain Data: Literature Review," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 5, Issue 5, pp. 125-132, 2022.

classification of data is the name given to the machine learning algorithm used to predict which data instances belong to which group. classification techniques are used, but some are summarised to classify data (Archana and Elangovan 2014), The Bayesian classifier is a graph-based model that uses a set of variable features to generate a probability relation between them. Based on Bayes's theorem, it uses a network structure that's a one-to-one correspondence between the required features and the nodes, which are represented by a directed acyclic graph (DAG). Known structure and unknown structure are examples of two network scenarios found in general frameworks. With its simple design and high computational efficiency, the Bayesian classifier is quick to learn and implement. The classifier's drawbacks are the need for a large number of data sets and the production of low precision in the results. The computational model communicates one signal with another signal via a large number of weighted connections, each of which is composed of a few simple processing units. Predicting new observations based on existing data that include neurological brain functions and the cognitive system of learning processes are neural network analytic techniques. It's a technique used in data mining. There is no need to reprogram this network because it learns as it goes. It can be applied to real-world problems and is simple to use and implement. The major drawbacks are that it takes longer to process, it's difficult to identify the neurons and layers, and it learns slowly. In recent days, decision tree classifiers can handle the uncertain numerical and categorical data produces many problems in the data mining. In the existing techniques provided many problems in the construction of decision tree because the tree satisfied either in numerical data or categorical data. Therefore, it creates tree with uncertain numerical and categorical data. In order to overcome the major limitations of traditional studies and the discussion of results to confirm the effectiveness of the proposed technique over the existing methods.

Effective decision-making and data analysis play important roles in data mining. Using database technology and uncertain numerical and categorical data, decide tree. Numerical data and categorical data are two types of data domains found in real-world applications. Uncertain data for decision tree making was discussed in this study. Many issues were solved by removing unreliable data from files such as video, mp3, and text. In the uncertain numerical and categorical data, think about extracting an attribute or a feature to think about. There are a variety of techniques used to improve classification accuracy and decision tree efficiency, such as fuzzy decision trees, probabilistic databases, the UK-means algorithm, and the processing of imprecise queries, which have been demonstrated in research studies. Tuple splitting is provided by the fuzzy decision tree technique, but it is inefficient. While a probabilistic database can predict the values, it only provides access to known tuples. The processing of imprecise queries takes longer because of the extra time required to solve the requestor's queries. Mining uncertain numerical or categorical data is more difficult with the extended UK-mean algorithm because it performs worse and costs more money. As a result of these existing techniques, we were inspired to develop a decision tree based on the averaging and distribution-based technique for accessing uncertain data in data mining. The study's goal is to build a decision tree from uncertain data, and existing systems have a number of limitations that need to be investigated further and resolved.

II. Literature Review

Machine learning algorithms and data mining both face the challenge of classification as a classic problem. The decision tree model, which has remained both practical and easy to understand, is an important classification model. Medical investigation, image recognition, fraud detection, systematic medical diagnosis, and target advertising are just a few uses for the algorithms. Probability distributions supported an active method of building decision trees, and decision tree classifiers relied on numerical attributes that were uncertain. Traditionally, decision-tree classification maintained a tuple's characteristic feature along categorical and numerical axes at the same time. One-point value capturing the value of an attribute or feature and postulating the range of possible values that increases probability (Pei, et al. 2014). Probability distributions, such as mean and variance, are used to abstract away data uncertainty.

Decision Tree Learning

One of the best hierarchical models for making decisions was the decision tree (DT), which used decision rules that recursively divided variables into homogeneous zones. With the DT building, we wanted to find out what decision rules were being used to predict the outcome from the input data. It was referred to as a classification tree when dealing with discrete variables, or a regression tree when dealing with continuous ones. Real-world applications, as well as situations like classification and prediction, were successfully explained. The classification and regression trees, as well as other algorithms, were used to build the decision tree model. Chi-square statistics were achieved using a CHAID as one of the decision tree techniques. These specific input and output variables were fed into the regressive decision tree method in order to get the best results. Various decision tree models saved as plain text files were converted and loaded into the database for use in the analysis (Celona 2017).

Agrawal and Gupta (2013) looked at a variety of decision tree classification methods, including ID3 and C4.5 algorithms as well as improved algorithms. Because of the high memory requirements and low efficiency, these classification algorithms were employed in the data processing. Better results are now possible due to the new algorithm, which constructs the decision tree more clearly and efficiently while preserving all previous decisions and inputs. The ability to efficiently sort and classify data was greatly improved. Binary trees produced only by a few decision tree algorithms were produced by others. To verify different large datasets that are publicly available on the UCI machine learning repository, we will use the less far algorithm generated by C4.5. The improved algorithm produces faster and more effective results without changing the final decision, and the presented algorithm helps to make the decision tree more understandable and less muddled to look at. There has been a significant improvement in terms of efficiency and classification. When working with large datasets, decision

Abdarrouf Abdalati Abdalgader Saed and Adam Amril Jaharadak, “Implementation with Performance Evaluation of Decision Tree Classifier for Uncertain Data: Literature Review,” *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 5, Issue 5, pp. 125-132, 2022.

tree generation efficiency suffers. After carefully reviewing the design process, all antilogarithms present in logarithmic calculation were often insignificant, so the procedure was streamlined using the L'Hospital Rule.

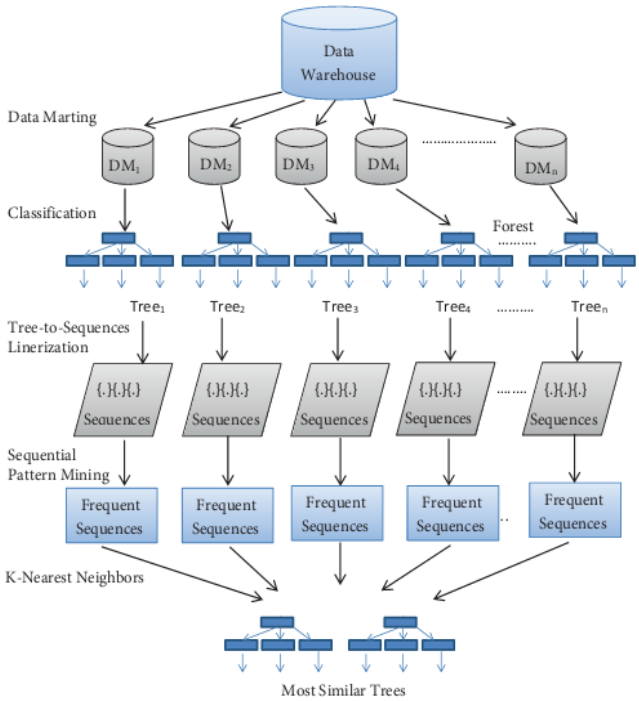

Fig. 2. Causal decision trees

To find the discriminant behaviour present in adult drug-dependent patients, Valero et al. (2014) used a data mining procedure called decision tree learning, which contradicted the results of a cross-validation technique. Neuroticism and impulsivity were found to be the most important personality traits to consider when conceptualising drug dependence, according to the ground-breaking findings. Despite this, they are still relevant when you consider their position in the organisation. When high levels of neuroticism are present in drug-dependent individuals, impulsivity becomes a key differentiating factor. When it comes to figuring out complex relationships, the decision tree learning method has proven to be an efficient and simple solution. This method improved upon previously obtained results, not only by giving clinicians a useful tool for making important clinical decisions, but also by limiting and prioritising important clinical variables when a specific therapeutic intervention was implemented.

It was found that the operating conditions (OC) variations and possible topology deviations of power systems occur throughout the operation of horizon when using an online dynamic security assessment data- mining framework by He, et al. (2013). A solid structure was predicted using adaptive co-operative decision tree learning, and it did not disappoint. When practising the classification model off-line, boosting algorithms were used to practise polling numerous unpruned small-height DTs. Small-height DT voting weights were also modernised to account for OC variations or possible changes in system topology, and new-fangled training cases were periodically incorporated into small-height DT voting weights. OC to safety of federation choices were plotted using the efficient cataloguing model in online DSA. To demonstrate the proposed system, IEEE 39-bus test system and the Western Electricity Coordinating Council's regional grid were used. As a result of the case studies, the proposed scheme's effectiveness was demonstrated by dealing with OC variation and changing the scheme's topology.

III. Materials and Methods

The current system handles valued data tuples with a traditional decision tree algorithm. A tuple's feature can be categorical or numerical in traditional decision-tree classification. In many cases, data uncertainty occurs before a precise point value is determined. A feature/value attributes isn't summed up in a single number, but rather in a set of numbers with a probability distribution attached to them. Measurement errors, stale data, and repeated measurements all contribute to data uncertainty. When constructing a decision tree, an average approach is typically employed. When calculating the mean and variance of a dataset, simple statistical derivations such as Classification makes use of ID3 (Iterative Dichotomiser 3). For decades, missing value has been used to address the decision tree's missing data. Pruning decision is the most critical algorithm in tree construction. Tree averaging is a classification method for classifying new objects in a forest. A new method of uncertain data pruning based on a decision tree is proposed to overcome the drawbacks. To deal with data tuples that have uncertain values, the proposed scheme uses a classical decision tree algorithm. It is possible to construct decision trees from probability distributions using the distribution-based approach. Classifiers using decision trees have a high degree of accuracy. The data item's complete information (probability density function) is used. Bounding and end point is a new technique for increasing computational efficiency.

IV. Discussion

The decision tree is the most widely used method of classifying various types of data. Root node of the decision tree has no incoming edges, while all other nodes have exactly one incoming edge. Most people are familiar with C4.5, which extends the ID3 decision tree algorithm in a significant way. There are advantages to using the C4.5, such as the ability to choose between splitting and continuous attributes, and the ability to handle data sets that contain missing values. There is a problem with the traditional decision tree when trying to classify data that is uncertain. This is an issue with traditional algorithms because the target attribute will be restricted to having discrete values in the new approach. Irrelevant attributes are overly sensitive to the training set. It's difficult to prepare a large decision tree with many branches when using a decision tree. The traditional algorithm is having a hard time keeping up with the rapid growth of cloud computing and big data. The explicit representation of the structure in a dataset is embodied in decision trees and lists, making them potentially powerful predictors. The learning algorithm's ability to summarise this structure succinctly determines their accuracy and understandability. The final model should exclude patterns that

Abdarrouf Abdalati Abdalgader Saed and Adam Amril Jaharadak, “Implementation with Performance Evaluation of Decision Tree Classifier for Uncertain Data: Literature Review,” *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 5, Issue 5, pp. 125-132, 2022.

aren't part of the underlying domain from spurious effects. Pruning mechanisms require a sensitive instrument that uses data to detect whether there is a genuine relationship between the components of a model and the domain. Pruning mechanisms require an efficient mechanism for determining when a particular effect is due solely to chance. For precisely this purpose, statistical significance tests are well-established tools with a solid theoretical foundation.

Determine the test tuple's class by traversing the tree from the root node until a leaf node is reached to determine t0 = (v0, 1,..., v 0, k). To move to the left child, go to internal node n and run the test v 0, jn zn. At some point, you'll come to node m, the leaf node. For each class label c C, the probability distribution Pm assigns a probability value to t0. Return the c C class label that maximises Pm for a single result (c). The decision tree algorithm is proposed to put an end to this. Nominal and numerical attributes can both be handled by the proposed decision tree. A single discrete-value classifier can be represented using this solitary example. In addition, it can deal with datasets with errors and datasets that could have errors. In the decision tree, the distribution of space and classifier structure is not assumed. Constructing a decision tree from tuples of uncertain values is the most difficult task. Finding an appropriate probability distribution Pm over C for each leaf node m requires finding a good testing attribute Ajn and a good split point zn for each internal node. There are d training tuples, A1...Ak, in the proposed system, each of which has three numerical attribute values: A1, A2, and A3. There was also an Aj is dom domain attribute (Aj). There is a feature vector Vi = (vi,1,vi,2,....) and a class label ci associated with each tuple ti, with the exception of tuple ti, which is associated with the set of all class labels.

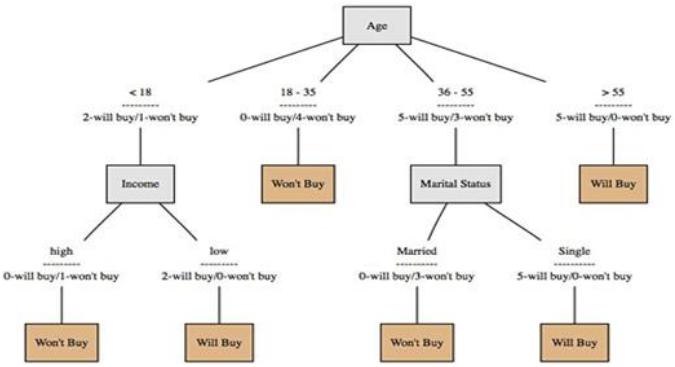

Fig. 3. Simple decision tree algorithm

Basically, the classification problem is to build an algorithm M that maps each feature vector (vx,1,...vx,k) to a probability distribution Px on C, such that given the tuple (v0,1,... v0k) and the classifier (P0), M = M (v0-1,1,... v0-k). In a decision tree, each internal node n has an attribute Ajn, which is linked to a split point zn (Ajn), which is used to generate the binary test. The "left" and "right" children of an internal node are labelled as such: "left" and "right." The discrete probability distribution Pm over C is attached to each decision tree leaf node m. When a tuple with a class label of C is assigned to leaf node m, Pm(c) estimates the probability that the tuple has a class label of c. Examine the curve for u=0 to get an estimate of the value. According to common sense, the highest point should be given a high estimate. As a result of the curve's wide plateau, it's difficult to nail down a single to estimate. To estimate a 95% confidence interval for each data point, use the accuracy values measured from the repeated experiments to estimate the interval, and then discover the set of points whose confidence interval overlaps with the interval of the most accurate point.

Ross Quinlan created the C4.5 algorithm, which generates a decision tree. Quinlan's earlier ID3 algorithm has been extended with C4.5. C4.5's decision trees can be used to classify data, which is why the programme is sometimes referred to as a statistical classifier instead. Like ID3, C4.5 uses the concept of information entropy to construct decision trees from a set of training data. There are already classified samples in the training data. In each sample, there is an array of p-dimensional vectors, each of which represents an attribute or feature of the sample, as well as the class in which Si belongs. These xi vectors are used to represent the attributes or features of the samples. First, the classifier must classify unknown data, for which a decision tree is generated. Algorithm C4.5 is based on algorithm ID3 but with some modifications. Similarly, the C5 algorithm adheres to the same principles as the C4.5 algorithm. The following are a few of the algorithm's many features. One way to look at the massive decision tree is as a simple set of rules. The noise and missing data are reduced by the C4.5 algorithm. The C4.5 algorithm resolves the issue of overfitting and error pruning. The C4.5 classifier predicts which attributes are relevant and which are not relevant in classification when using classification technique.

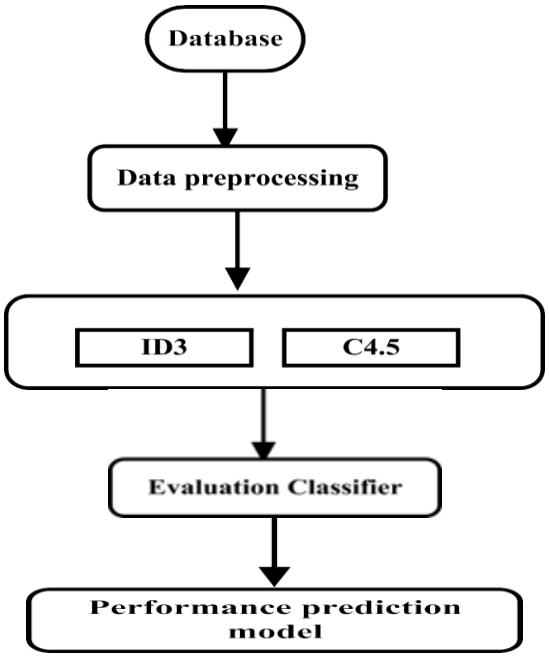

Fig. 4. Processing model

In the ID3 classification algorithm, all examples are mapped to different categories based on different values of the condition attribute set. The ID3 algorithm's core goal is to determine the best classification attribute from condition attribute sets.

Abdarrouf Abdalati Abdalgader Saed and Adam Amril Jaharadak, "Implementation with Performance Evaluation of Decision Tree Classifier for Uncertain Data: Literature Review," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 5, Issue 5, pp. 125-132, 2022.

Usually, the attribute with the highest information gain is selected as the splitting attribute of current node in order to make entropy information that the divided subsets need the smallest. The algorithm selects attribute information gain as selection criteria. Several new features have been added to ID3, such as the ability to classify continuous attributes, deal with missing value exceptions, prune decision trees, and derive business rules.

Techniques for data mining produce a wide variety of ways to represent the information. Classification schemes are commonly organised using decision tree structures. Decision trees show the steps taken to arrive at a classification when classifying tasks. The root node of every decision tree is referred to as the "parent" of all other nodes. In the tree, each node determines which path it should take based on an attribute in the data. Decision tests are frequently based on comparing one value to another. Routing from the root node to the leaf node is used for classification using a decision tree. D3 selects the splitting attribute with the highest information gain, where information gain is defined as the difference between how much information is required after the split. D3 chooses the highest information gain. Using the difference between each subdivided dataset's weighted entropies and the original dataset's entropies, this is calculated. A classic example of data mining is shown here, which involves deciding whether or not to play a game based on weather conditions. Outlook is the node at the very top of the tree. There are attribute values in the node's degrees. The child nodes in this example are used to measure humidity and wind speed, and the leaf nodes are used to classify the results.

Decision trees can represent diverse types of data. The simplest and most familiar is numerical data. It is often desirable to organize nominal data as well. Nominal quantities are formally described by a discrete set of symbols. For example, weather can be described in either numeric or nominal fashion. It can quantify the temperature by saying that it is 11 degrees Celsius or 52 degrees Fahrenheit. It could also say that it is cold, cool, mild, warm, or hot. The former is an example of numeric and the latter is a kind of nominal information. Data that isn't meaningful, like cold or mild, are referred to as nominal. In ordinal data, the values are assumed to have some sort of ordered relationship. Assuming we're still talking about weather, we could use terms like sunny, overcast, and rainy instead of the more descriptive adjectives. There are no connections between these numbers, and there is no way to know how far apart they are. Understanding how a tree works at the node level necessitates an understanding of the type of data it organises. Keeping in mind that each node is a test, numerical data is frequently analysed in terms of simple mathematical inequality. Numeric weather data, for example, could be tested to see if it exceeds 10 degrees Fahrenheit. To determine whether or not nominal data has a specific value, Boolean logic is used. Both types of tests are depicted in the illustration. Outlook is an example of a nominal data type in the context of weather. An attribute value is identified by asking the test which attribute value it represents. The humidity node reflects numerical tests, with a difference between 70 and greater than 70 being less than or equal to.

Recursive algorithms for decision tree induction are used. To begin, a root node must be assigned to an attribute. The root node must effectively split the data if the goal is to create the most efficient (i.e., smallest) tree. As a result of each split, a smaller and smaller subset of instances (the underlying data) is analysed until they are all classified the same way. When it comes to deciding on a split, choose the one that offers the most information gain. Claude Shannon, the father of modern information theory, introduced the concept of entropy in his work. The word "information" has many meanings, but in mathematics, it refers to the degree of certainty with which decisions can be made. Each branch of the decision tree should, in theory, bring us closer to a classification. To put it another way, think of it as removing randomness or entropy as you move up the tree. This is reflected mathematically in information. Using this as an example, let's create a decision tree to decide yes or no on the basis of some data. This is exactly how the decision tree depicts the situation. There will be a certain number of yes/no classifications for each attribute value. If the number of yeses and noes is equal, then the value has a lot of entropy. The amount of information available is at its most in this scenario. If the responses are all yeses or all noes, then the information is zero as well. Attribute value has a low entropy and is extremely useful for decision-making

V. CONCLUSION

In the decision tree building, known and precise values are used widely in data mining as a data classification technique. There are numerous problems with classification, and this applies across a wide range of data mining applications. Data classification using decision trees is very popular because of their simple and robust structure. Data loss in the decision tree because the traditional decision tree uses point-valued data tuples. As the amount of information increases, the decision tree's performance will be impacted, and the calculation of entropy will become more complex. For this reason, decision tree classification methods were extended to include knowledge tuples to ensure that numerical attributes with uncertainty data were defined in terms of discretional pdfs. They worked well. created call trees with data classification based on the modification present in existing classical call tree building algorithm using C4.5 algorithm. After that, the accuracy of the decision tree for the uncertain data used is high because appropriate pdfs have been used. Improve the efficiency of a constructed tree by employing various pruning techniques. In comparison to other techniques, the proposed decision tree for uncertain data achieves higher efficiency. As a result, a large number of experiments were considered. For the construction of the decision tree, this method uses classical algorithms that generate enormous numbers of data tuples (one for each decision). The proposed method achieves a better result because the execution time is shorter, and the system's efficiency is higher. The proposed work will be extended in the future to improve the data classifiers' pruning efficiency when building decision trees. This lays the groundwork for the rest of the research project.

Abdarrouf Abdalati Abdalgader Saed and Adam Amril Jaharadak, “Implementation with Performance Evaluation of Decision Tree Classifier for Uncertain Data: Literature Review,” *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 5, Issue 5, pp. 125-132, 2022.

REFERENCES

[1]. Agrawal GL & Gupta H 2013, 'Optimization of C4. 5 Decision Tree Algorithm for Data Mining Application', International Journal of Emerging Technology and Advanced Engineering, vol. 3, no. 3, pp. 341-345.

[2]. Aizpurua J, Papadopoulos Y, Muxika E, Chiacchio F & Manno G 2017, 'On Cost-effective Reuse of Components in the Design of Complex Reconfigurable Systems', Quality and Reliability Engineering International

[3]. Alkasem A, Liu H & Zuo D 2017, 'Utility Cloud: A Novel Approach for Diagnosis and Self-healing Based on the Uncertainty in Anomalous Metrics', Proceedings of the 2017 International Conference on Management Engineering, Software Engineering and Service Sciences, pp. 99-107.

[4]. Amagata D, Sasaki Y, Hara T & Nishio S 2016, 'Probabilistic nearest neighbour query processing on distributed uncertain data', Distributed and Parallel Databases, vol. 34, no. 2, pp. 259-287.

[5]. Anitha MBR, Dhakshayani L & Kavitha V 2017, 'Adaptive Processing for Distributed Skyline Queries over Uncertain Data',

[6]. Appel R, Fuchs TJ, Dollár P & Perona P 2013, 'Quickly Boosting Decision Trees-Pruning Underachieving Features Early', ICML (3), pp. 594-602.

[7]. Archana S & Elangovan K 2014, 'Survey of classification techniques in data mining', International Journal of Computer Science and Mobile Applications, vol. 2, no. 2, pp. 65-71.

[8]. BAKIRLI G & Birant D 2017, 'DTreeSim: A new approach to compute decision tree similarity using re-mining', Turkish Journal of Electrical Engineering & Computer Sciences, vol. 25, no. 1, pp. 108-125.

[9]. Balan S & Ponmuthuramalingam P, 'A Study on Pruning Techniques in Web Content Mining',

[10]. Bounhas M, Hamed MG, Prade H, Serrurier M & Mellouli K 2014, 'Naive possibilistic classifiers for imprecise or uncertain numerical data', Fuzzy Sets and Systems, vol. 239, pp. 137-156.

[11]. Brown N, Kroer C & Sandholm T 2017, 'Dynamic Thresholding and Pruning for Regret Minimization',

[12]. Bui N, Vo B, Huynh V-N, Lin C-W & Nguyen LT 2016, 'Mining closed high utility item sets in uncertain databases', Proceedings of the Seventh Symposium on Information and Communication Technology, pp. 7-14.

[13]. Cagliero L & Garza P 2014, 'Infrequent weighted itemset mining using frequent pattern growth', IEEE transactions on knowledge and data engineering, vol. 26, no. 4, pp. 903-915.

[14]. Cao K, Wang G, Han D, Bai M & Li S 2016, 'An algorithm for classification over uncertain data based on extreme learning machine', Neurocomputing, vol. 174, pp. 194-202.

[15]. Celona J 2017, Applying Decision Analysis to Human Factors in Decision Making at Stanford University Medical Center, in Advances in Human Factors, Business Management, Training and Education, Springer, pp. 433-445.

[16]. Chen L, Li X, Yang Y, Kurniawati H, Sheng QZ, Hu H-Y, et al. 2016, 'Personal health indexing based on medical examinations: a data mining approach', Decision Support Systems, vol. 81, pp. 54-65.

[17]. Cheng R, Emrich T, Kriegel H-P, Mamoulis N, Renz M, Trajcevski G, et al. 2014, 'Managing uncertainty in spatial and spatio-temporal data', Data Engineering (ICDE), 2014 IEEE 30th International Conference on, pp. 1302- 1305.

[18]. Ciceri E, Fraternali P, Martinenghi D & Tagliasacchi M 2016, 'Crowdsourcing for top-k query processing over uncertain data', IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 1, pp. 41-53.

[19]. Cuzzocrea A, Leung CK-S & MacKinnon RK 2014, 'Mining constrained frequent itemsets from distributed uncertain data', Future Generation Computer Systems, vol. 37, pp. 117-126.

[20]. Denoeux T 2013, 'Maximum likelihood estimation from uncertain data in the belief function framework', IEEE Transactions on knowledge and data engineering, vol. 25, no. 1, pp. 119-130.

[21]. Ding S, Wu F, Qian J, Jia H & Jin F 2015, 'Research on data stream clustering algorithms', Artificial Intelligence Review, vol. 43, no. 4, pp. 593-600.

[22]. Dralle DN, Karst NJ, Charalampous K, Veenstra A & Thompson SE 2017, 'Event-scale power law recession analysis: quantifying methodological uncertainty', Hydrology and Earth System Sciences, vol. 21, no. 1, p. 65.

[23]. Fournier-Viger P, Lin JC-W, Kiran RU & Koh YS 2017, 'A Survey of Sequential Pattern Mining', Data Science and Pattern Recognition, vol. 1, no. 1, pp. 54-77.

[24]. Galbrun E & Miettinen P 2012, 'A case of visual and interactive data analysis: geospatial redescription mining', Instant Interactive Data Mining Workshop at ECML-PKDD.

[25]. Gan W, Lin JC-W, Fournier-Viger P, Chao H-C, Wu JM-T & Zhan J 2017, 'Extracting recent weighted-based patterns from uncertain temporal databases', Engineering Applications of Artificial Intelligence, vol. 61, pp. 161-172.

[26]. Ge J & Xia Y 2016, 'Distributed Sequential Pattern Mining in Large Scale Uncertain Databases', Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 17-29.

[27]. Ghattas B, Michel P & Boyer L 2017, 'Clustering nominal data using Unsupervised Binary decision Trees: Comparisons with the state-of-the-art methods', Pattern Recognition

[28]. Goodkind A, Brizan DG & Rosenberg A 2017, 'Utilizing overt and latent linguistic structure to improve keystroke-based authentication', Image and Vision Computing, vol. 58, pp. 230-238.

[29]. Goyal N & Jain S 2016, 'A comparative study of different frequent pattern mining algorithm for uncertain data: A survey', Computing, Communication and Automation (ICCCA), 2016 International Conference on, pp. 183-187.

[30]. Guha S, Rastogi R & Shim K 2000, 'ROCK: A robust clustering algorithm for categorical attributes', Information systems, vol. 25, no. 5, pp. 345-366.

[31]. Gullo F, Ponti G, Tagarelli A & Greco S 2017, 'An information-theoretic approach to hierarchical clustering of uncertain data', Information Sciences.

[32]. Han D, Giraud-Carrier C & Li S 2015, 'Efficient mining of high-speed uncertain data streams', Applied Intelligence, vol. 43, no. 4, pp. 773-785.

[33]. He M, Zhang J & Vittal V 2013, 'Robust online dynamic security assessment using adaptive ensemble decision-tree learning', IEEE Transactions on Power Systems, vol. 28, no. 4, pp. 4089-4098.

[34]. Henriques D, Villaverde AF, Rocha M, Saez-Rodriguez J & Banga JR 2017, 'Data-driven reverse engineering of signaling pathways using ensembles of dynamic models', PLoS computational biology, vol. 13, no. 2, p. e1005379.

[35]. Jadhav A, Pandita A, Pawar A & Singh V 2016, 'Classification of Unstructured Data Using Naïve Bayes Classifier and Predictive Anaysis for RTI Application', An International Journal of Engineering & Technology, vol. 3, no. 6

[36]. Jiang B, Pei J, Tao Y & Lin X 2013, 'Clustering uncertain data based on probability distribution similarity', IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 4, pp. 751-763.

[37]. Jiang Z, Shekhar S, Zhou X, Knight J & Corcoran J 2015, 'Focal-test-based spatial decision tree learning', IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 6, pp. 1547-1559.

[38]. Jin C, Yu JX, Zhou A & Cao F 2014, 'Efficient clustering of uncertain data streams', Knowledge and information systems, vol. 40, no. 3, pp. 509-539.

[39]. Kabir S 2017, 'An overview of fault tree analysis and its application in model-based dependability analysis', Expert Systems with Applications

[40]. Kamadi VV, Allam AR & Thummala SM 2016, 'A computational intelligence technique for the effective diagnosis of diabetic patients using principal component analysis (PCA) and modified fuzzy SLIQ decision tree approach', Applied Soft Computing, vol. 49, pp. 137-145.

[41]. Karem F, Dhibi M, Martin A & Bouhlel MS 2017, 'Credal Fusion of Classifications for Noisy and Uncertain Data', International Journal of Electrical and Computer Engineering (IJECE), vol. 7, no. 2

[42]. Kellogg R 2014, 'The effect of uncertainty on investment: evidence from Texas oil drilling', The American Economic Review, vol. 104, no. 6, pp. 1698-1734.

[43]. Krishna S, Puhrsch C & Wies T 2015, 'Learning invariants using decision trees', arXiv preprint arXiv:1501.04725

[44]. Lee G & Yun U 2017, 'A new efficient approach for mining uncertain frequent patterns using minimum data structure without false positives', Future Generation Computer Systems, vol. 68, pp. 89-110.

[45]. Leung CK-S, MacKinnon RK & Jiang F 2016, 'Finding efficiencies in frequent pattern mining from big uncertain data', World Wide Web, pp. 1-24.

[46]. Li C 2017, 'Wearable Computing: Accelerometer-Based Human Activity Classification Using Decision Tree',

[47]. Li J, Le TD, Liu L, Liu J, Jin Z, Sun B, et al. 2016, 'From observational studies to causal rule mining', ACM Transactions on Intelligent Systems and Technology (TIST), vol. 7, no. 2, p. 14.
[48]. Li J, Ma S, Le T, Liu L & Liu J 2017, 'Causal decision trees', IEEE Transactions on Knowledge and Data Engineering, vol. 29, no. 2, pp. 257-271.
[49]. Li S, Dragicevic S, Castro FA, Sester M, Winter S, Coltekin A, et al. 2016, 'Geospatial big data handling theory and methods: A review and research challenges', ISPRS Journal of Photogrammetry and Remote Sensing, vol. 115, pp. 119-133.
[50]. Liao K-T & Liu C-M 2016, 'An Effective Clustering Mechanism for Uncertain Data Mining Using Centroid Boundary in UKmeans', Computer Symposium (ICS), 2016 International, pp. 300-305.
[51]. Lin JC-W, Gan W, Fournier-Viger P, Hong T-P & Tseng VS 2016, 'Efficient algorithms for mining high-utility itemsets in uncertain databases', Knowledge- Based Systems, vol. 96, pp. 171-187.
[52]. Liu B, Xiao Y, Cao L, Hao Z & Deng F 2013, 'SVDD-based outlier detection on uncertain data', Knowledge and information systems, vol. 34, no. 3, pp. 597- 618.
[53]. Liu H, Gegov A & Cocea M 2017, Complexity control in rule-based models for classification in machine learning context, in Advances in Computational Intelligence Systems, Springer, pp. 125-143.
[54]. Liu H, Zhang X, Zhang X & Cui Y 2017, 'Self-adapted mixture distance measure for clustering uncertain data', Knowledge-Based Systems
[55]. Lu H, Setiono R & Liu H 2017, 'Neurorule: A connectionist approach to data mining', arXiv preprint arXiv:1701.01358
[56]. Luo Q, Peng Y, Peng X & Saddik AE 2014, 'Uncertain data clustering-based distance estimation in wireless sensor networks', Sensors, vol. 14, no. 4, pp. 6584-6605.
[57]. Luo Q, Yan X, Li J & Peng Y 2014, 'DDEUDSC: a dynamic distance estimation using uncertain data stream clustering in mobile wireless sensor networks', Measurement, vol. 55, pp. 423-433.
[58]. Lyall-Wilson JR 2013, 'Automatic Concept-Based Query Expansion Using Term Relational Pathways Built from a Collection-Specific Association Thesaurus',
[59]. Ma J, Sun L, Wang H, Zhang Y & Aickelin U 2016, 'Supervised anomaly detection in uncertain pseudoperiodic data streams', ACM Transactions on Internet Technology (TOIT), vol. 16, no. 1, p. 4.
[60]. MacKinnon RK, Leung CK-S & Tanbeer SK 2014, 'A scalable data analytics algorithm for mining frequent patterns from uncertain data', Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 404-416.
[61]. Mansha S, Babar Z, Kamiran F & Karim A 2016, 'Neural Network Based Association Rule Mining from Uncertain Data', International Conference on Neural Information Processing, pp. 129-136.
[62]. Mantovani RG, Horváth T, Cerri R, Carvalho A & Vanschoren J 2016, 'Hyper- parameter Tuning of a Decision Tree Induction Algorithm', Brazilian Conference on Intelligent Systems (BRACIS 2016).
[63]. Mihelčić M, Džeroski S, Lavrač N & Šmuc T 2017, 'A framework for redescription set construction', Expert Systems with Applications, vol. 68, pp. 196-215.
[64]. Nabilah RM, Othman Z & Azuraliza BA 2016, 'Approaches of Handling Uncertain Time Series Data towards Prediction', International Journal of Future Computer and Communication, vol. 5, no. 6, p. 233.
[65]. Nguyen T-L, Vo B & Snasel V 2017, 'Efficient algorithms for mining colossal patterns in high dimensional databases', Knowledge-Based Systems, vol. 122, pp. 75-89.
[66]. OBRACZKA D 2017, Active Learning of Link Specifications using Decision Tree Learning, degree, UNIVERSITY OF LEIPZIG.
[67]. Oliver JJ & Hand DJ 2016, 'On pruning and averaging decision trees', Machine Learning: Proceedings of the Twelfth International Conference, pp. 430-437.
[68]. Pandya RPaJ 2015, 'C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning', International Journal of Computer Applications (0975 – 8887), Volume 117 – No. 16, May 2015
[69]. Patel RR & Aluvalu R 2014, 'A reduced error pruning technique for improving accuracy of decision tree learning', International Journal of Engineering and Advanced Technology (IJEAT), pp. 2249-8958.
[70]. Pei B, Zhao T, Zhao S & Chen H 2014, Fuzzy associative classifier for probabilistic numerical data, in Foundations and Applications of Intelligent Systems, Springer, pp. 563-578.
[71]. Phadatare MM & Nandgaonkar SS 2014, 'Uncertain Data Mining using Decision Tree and Bagging Technique', International Journal of Computer Science and Information Technologies, vol. 5, no. 3, pp. 3069-3073.
[72]. Pradhan B 2013, 'A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS', Computers & Geosciences, vol. 51, pp. 350-365.
[73]. Rind A, Lammarsch T, Aigner W, Alsallakh B & Miksch S 2013, 'Timebench: A data model and software library for visual analytics of time-oriented data', IEEE Transactions on Visualization and Computer Graphics, vol. 19, no. 12, pp. 2247-2256.
[74]. Schmid H 2013, 'Probabilistic part-ofispeech tagging using decision trees', New methods in language processing, p. 154.
[75]. Segatori A, Marcelloni F & Pedrycz W 2017, 'On Distributed Fuzzy Decision Trees for Big Data', IEEE Transactions on Fuzzy Systems
[76]. Shaikh SA & Kitagawa H 2014, 'Efficient distance-based outlier detection on uncertain datasets of Gaussian distribution', World Wide Web, vol. 17, no. 4, pp. 511-538.
[77]. Shajib MB-U-Z, Samiullah M, Ahmed CF, Leung CK & Pazdor AG 2016, 'An Efficient Approach for Mining Frequent Patterns Over Uncertain Data Streams', Tools with Artificial Intelligence (ICTAI), 2016 IEEE 28th International Conference on, pp. 980-984.
[78]. Shehzad K 2013, 'Simple hybrid and incremental postpruning techniques for rule induction', IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 2, pp. 476-480.
[79]. Sutton-Charani N, Destercke S & Denœux T 2014, 'Training and Evaluating Classifiers from Evidential Data: Application to E 2 M Decision Tree Pruning', International Conference on Belief Functions, pp. 87-94.
[80]. Thankachan SV, Patil M, Shah R & Biswas S 2015, 'Probabilistic threshold indexing for uncertain strings', arXiv preprint arXiv:1509.08608
[81]. Tomar D & Sathappan S 2016, 'A method for handling clusturing of uncertain data', Advances in Human Machine Interaction (HMI), 2016 International Conference on, pp. 1-5.
[82]. Valero S, Daigre C, Rodríguez-Cintas L, Barral C, Gomà-i-Freixanet M, Ferrer M, et al. 2014, 'Neuroticism and impulsivity: Their hierarchical organization in the personality characterization of drug-dependent patients from a decision tree learning perspective', Comprehensive psychiatry, vol. 55, no. 5, pp. 1227-1233.
[83]. Van Leeuwen M 2014, Interactive data exploration using pattern mining, in Interactive Knowledge Discovery and Data Mining in Biomedical Informatics, Springer, pp. 169-182.
[84]. Vilisov V 2017, 'Risk Proneness Estimation Method Developed in Relation to the Decision Taker that Controls the Robotic System', arXiv preprint arXiv:1703.06161
[85]. West IV TK & Gumbert C 2017, 'Multifidelity, Multidisciplinary Design Under Uncertainty with Non-Intrusive Polynomial Chaos',
[86]. Xu L, Hu Q, Hung E, Chen B, Tan X & Liao C 2015, 'Large margin clustering on uncertain data by considering probability distribution similarity', Neurocomputing, vol. 158, pp. 81-89.
[87]. Yan D, Zhao Z, Ng W & Liu S 2015, 'Probabilistic convex hull queries over uncertain data', IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 3, pp. 852-865.
[88]. Yang J, Ma J, Berryman M & Perez P 2014, 'A structure optimization algorithm of neural networks for large-scale data sets', Fuzzy Systems (FUZZ-IEEE), 2014 IEEE International Conference on, pp. 956-961.
[89]. Yun U, Lee G & Ryu KH 2014, 'Mining maximal frequent patterns by considering weight conditions over data streams', Knowledge-Based Systems, vol. 55, pp. 49-65.
[90]. Zhan L, Zhang Y, Zhang W & Lin X 2014, 'Identifying top k dominating objects over uncertain data', International Conference on Database Systems for Advanced Applications, pp. 388-405.
[91]. Zhang X, Gao L & Yu H 2016, 'Constraint based subspace clustering for high dimensional uncertain data', Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 271-282.
[92]. Zheng H, He J, Zhang Y & Shi Y 2017, 'A fuzzy decision tree approach based on data distribution construction', Proceedings of the Australasian Computer Science Week Multiconference, p. 5.
[93]. Zhou J, Pan Y, Chen CP, Wang D & Han S 2016, 'K-medoids method based on divergence for uncertain data clustering', Systems, Man, and Cybernetics (SMC), 2016 IEEE International Conference on, pp. 002671-002674.

[94]. Zhu J, Xu J, Zhang C & Gao Y 2017, 'Marine Fishing Ground Prediction Based on Bayesian Decision Tree Model', Proceedings of the 2017 International Conference on Management Engineering, Software Engineering and Service Sciences, pp. 316-320.