Technoarete Transactions on Intelligent Data Mining and Knowledge Discovery
Vol-2, Issue-3, August 2022
e-ISSN: 2583-195X

The Application of Decision Tree Method for Data Mining

Dr Komarasamy G [1], Dr. T. Nadana Ravishankar [2]

[1] VIT Bhopal University, India
[2] SRM Institute of Science and Technology, India
Corresponding Author E-Mail: [1] gkomarasamy@gmail.com

Abstract

The study showcases the impact of data mining considering the application of decision trees that helps to develop data from large number of datasets. Linear data set has been generated through the model of decision trees. Secondary data collection method has been selected in this study with inductive approach. Cross sectional research design has been used in this study to derived insight of the subject of the study. Themes are developed using the peer reviewed journals published after 2019 and secondary collected data has been interpreted in this study to meet the goal of the subject. Implications of data mining including decision trees and decision rules in different field have been discussed over here with positive perspective approach. Customization of data mining in creativity also has been focused here. Different types of decision trees methods have been depicted here with comparison vision. Advantages and disadvantages of decision trees also have been discussed over here to evaluate the actual impact of application of decision trees in data mining. Interpretation of results also has been depicted here to analyse the consequences of decision trees in data mining to conclude the study with informative justification..

Keywords

Customization, decision trees, data mining.

INTRODUCTION

Decision trees in data mining are the process to establish models in data mining. Controlling of linear data sets has been driven by decision trees effectively. There are two types of decision trees, one is categorical variable and other is continuous variable decision trees. Data mining is the process to sort data from large data sets considering the pattern and requirement for solving problems instantly considering the data analysis technique. Regression problems can be solved with the prediction ability of decision trees by categorization of the objects from insight of the data. The applications of decision trees have been considered to store hierarchical data, folder structure, organizational structure and XML/HTML data [1]. Decision trees in data mining are mostly used for operations research, decision analysis, identification of the strategy to reach the goal of the objective. Prioritizing the prior data prediction of the values of the target variables highlighting the decision rules is the major concern of decision trees in data mining.

Advantage and disadvantage of the decision trees in data mining have an influential impact on conducting business events. Interpretability, less data preparation, versatility and non-linearity are the most impactful features of decision trees in data mining. Beside this, over fitting, feature reduction, data re-sampling and optimization of decision trees can hamper the accumulation of the data. Interpretation and visualization of non-linear data patterns have been made easier by the implication of decision trees application in the data mining. The algorithm mostly used in the decision trees is called ID3. Main characteristics of ID3 is highlighting the entropy and information gained through data mining considering different attributions of the selected measures to construct a decision tree [2]. Composition of decision trees consist of decision nodes that reflect the choice, next one is chance nodes that define the probability and finally the end nodes that determine the outcomes.

MATERIALS AND METHODS

The study has focused on the aim of the subject that reflects the impact of decision trees application in data mining. Therefore, the secondary data collection method has been chosen for the study. Peer reviewed journals are used for the collection of the secondary data to make the study more authentic and trustworthy [3]. Thematic data analysis helps to express the observation skills of the writer based on the realistic perspectives of the real world. Generally secondary data collection has been interpreted with thematic data analysis procedure to meet the goal of the study. In thematic data analysis themes are developed relevant to the subject of the study along with prioritizing the peer reviewed journals. Time saving and budget friendly approach of secondary data collection procedure and thematic data analysis method has enhanced the quality of the study.

The study has also excluded the peer reviewed journals which have been published before 2019. Inductive research approach and cross-sectional research design has been prioritized in this study to conduct the secondary data collection procedure in the right way. Focusing on the realistic insights research execution process has been driven by the inductive research approach [4]. Decision trees and its implication in data mining to execute the business operations can be described through different themes considering the real observation of the writer retrieved from the real world. Flexibility and independence of writing have been prioritized in thematic interpretation of secondary collected data.

Limitation hazards can be avoided in the secondary data collection method that widens the scope for data collection along with broader the way to interpret the data with theme-based interpretation. Defining the concept of decision trees and importance of decision trees in data mining can be explained briefly in themes that help to justify the study considering the objectives of the study. Evaluation of the subject of the study can be done with logical informative justification through thematic analysis of secondary data. Reliability and validity of secondary data collection methods has increased the quality of the study that includes the authentication of the study in a cost-effective manner [5]. Considering all the pros and cons thematic data analysis has been helped to meet the goal of the study with justification.

RESULTS

Implication of data mining in decision trees and decision rules

Data mining helps to execute the business operations following the market trends to make a sustainable business in a global aspect. Retrieving large data sets has widened the scope for the organizations to select the strategy and planned an effective approach to grab the market segmentation. Decision trees help to select the linear data set and establish a model for the data mining [6]. Deviation and predicting outcomes considering the insight of the derived data highlighting the relationship with the large data sets helps in statistical data analysis, machine learning and computer science to implicate the actionable information in business operations. Commercial and industrial data mining applications help to make decisions in favour of the companies to increase the profit of the business. Following the decision rules and decision trees, data mining has been prioritizing the symbolic and instant solution for the business operational problems.

Decision rules in data mining is featured with the application of BPMN that automatically explores the correlations of the data which is uploading and mapping onto IBM process mining to derive the insights of data sets relevant to the business operations. Automatic detection capability of decision rules data mining has governed the entire business operations considering the linear data derivation. Input values are measured in the DRM that determines the categorization of the data into branches. Decision trees help to discreet and continue the variables of data sets following the DRM [7]. A subset of data is the significant attribution of the decision trees. Considering the establishment of the data model through decision trees, decision rules can be spitted into different categories such as conjunctive, disjunctive, elimination by aspects, lexicographic, compensatory. Innovations, marketing all types of business operations can be driven by the decision rules including the objectives, restaurants and variables.

Classification of problems can be predicted by the decision trees maintaining the decision rules. Decision trees and decision rules both are simultaneously important for data mining to balance statistics and economics considering the strategy and planning of the business execution. Retrieving the knowledge from the classification of data sets with an inductive approach has been called the decision tree induction that helps to predict furthermore problems arriving in the way of business execution considering the real scenario of current market trends [8]. Following decision rules, decision trees application started to create diagrams including analysing unpredictable outcomes and trying to reach the most logical solution. After creating the diagram, decision trees added chance and decision nodes and expanded the chances till reaching the end points. Next, calculating tree values is another prior concern of decision tree to evaluate the final outcomes.

Data mining procedure can be simplified with following the decision trees application steps that ensure instant solution of problems to make sustainable business with leveraging competitive advantage. Competing hypotheses also has been prioritized by the decision rules mining that ensures the continuous flow of the business productivity [9]. Highlighting all consequences of decision rules and decision trees, application data mining in business performance of globalized organizations plays a vital role with the implication of advanced technology.

Customization of data mining considering the creativity

Innovation is the key factor for increasing the business profit and reaching the goal of the business. Competitive advantage of the business can be leveraged by the implication of innovation in business platforms and data mining has widened the scope for creativity that enhances the ability of innovation of globalized companies. Considering the customization facility of the data mining procedure, innovation has been prioritized by the organizations. Collecting data sets from large sets of data has been customized by the data mining procedure including decision trees and decision rules has helped to derive the insights of data insights to analyse the market trends regarding creativity [10]. More creativity increases the probability of business sustainability in a competitive business market. Personal learning models also have been influenced by the customization of data mining techniques. Importance extracts of datasets can be derived from the large data sets considering the customization of data mining.

Filtration of different data mining techniques also can be defined as the customization of data mining. This characteristic of data mining helps the different fields including educational, commercial, professional perspectives to utilize the opportunity of optimization of data mining to retrieve knowledge from data sets. E-commerce platforms have highly preferred the customization of the data mining that helps to explore the innovation capability of the business [11]. Companies' strengths and weaknesses also can be evaluated with the implication of customization of data mining. Decision trees enhance the quality of customization of data mining by categorizing the data sets according to the perspectives of the data. Creative moves in different professional fields have been inspired by the optimization of data sets to generate ideas and planned a strategic path to implicate the creative move in operations.

Decision trees play a major role in optimization of data mining to make the availability and accessibility of retrieving the data in favour of different professions. Big data customization helps in cloud computing, reducing mapping errors and creating programming frameworks that accelerates the creativity ability of the organizations. Integrating search engines in ecommerce business has been considered the optimization of data mining which introduces the creative approach of online business with personalization features [12]. In the business field customer-oriented information gathering through data mining and customization of data mining to shorten the data from large amounts of data have increased the ability of organizations to create innovative ideas to meet the customer's needs.

Customer satisfaction is the prior concern of globalized business organisations; therefore, customers relevant data mining has a positive impact on the business operations and customizations of the data mining has added an extra advantage to the organization profile. Entrepreneurs are mostly benefited with the application of optimized data mining in the business platforms considering the innovation capability of the start-up business [13]. Customized decision trees have helped the entrepreneurs to strategize plans for innovation. On the other hand, optimized features of data mining techniques in the educational programmes have helped to accelerate the development of personal competence. Digitized innovation programmes are mostly influenced by the application of customized data mining procedures.

Analysis of different decision tree algorithms

Decision trees algorithms can be categorized in different segments such as ID3, classification and regression trees (CART), Chi-Square and reduction in variation. Most popular decision trees algorithm is the CART which has helped to predict modelling problems considering the regression and classification of data mining [14]. Machine learning is highly influenced by the implication of classification and regression decision trees. The capability of CART mostly impacted the both tasks of classification and regression in data mining. Highly complex data can be explored with the application of CART and instantly reveal the important data to reach the patterns automatically to solve symbolic problems. CART analysis of data helps to classify the differentiation of the systems due to the natural uncertainties. Gini's impurity index has been followed by the CART decision algorithm that helps to identify the future problems arising in business operations to avoid the risk factors associated with it.

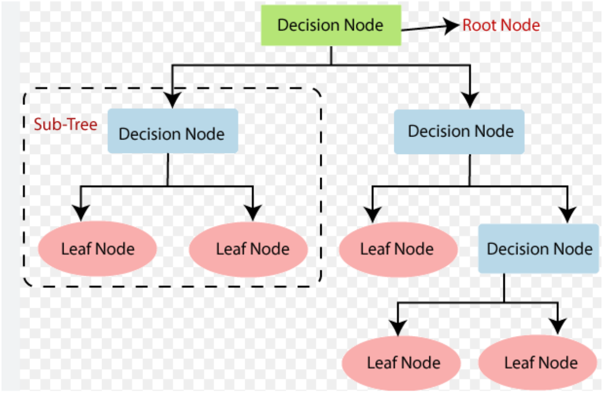

Figure 1: Decision tree algorithm in machine learning

Classification algorithms help to identify the distinct values and regression algorithms determine the continuous values. Most effective approach of CART decision tree algorithms is that it works on both data mining procedure tasks of classification and regression algorithms. CART is following the structure of binary tree to execute the decision tree mining considering shortening of data from a large data set. Limitation of CART model is that it cannot implicate any changes in the dataset as the small change can hamper the tree structure unstable that reflects on the variance in outcomes [15]. Natural language processing domain and machine learning have been prioritized the ID3 algorithm that helps to generate decision trees from significant dataset. Following a few steps ID3 algorithm is reaching out the data mining process completion, such as calculating the entropy first and then calculating the entropy of categorical values and gaining information for future considering each attribute.

Finally, feature with maximum information gained and repeatedly repeat the procedure to get the expected structure of the tree. Advantages of the ID3 algorithm are to predict understandable rules considering the training data. In comparison ID3 is a short tree that consumes short times for data mining. Classifiers accuracy has signified the CART algorithms rather than ID3 for data mining procedure [16]. Statistical test for identification of difference between categorical variables and random samples including observation for selection of the well-fitting results can be defined as the Chi-Square decision tree algorithm. The result of Chi Square decision tree algorithm meets the criteria of random, raw, mutually exclusive, large enough samples, independent variables-based data sets. Chi square data mining procedure is also helping in testing the hypothesis of the business projection that helps to predict the risk factors with a broader perspective approach.

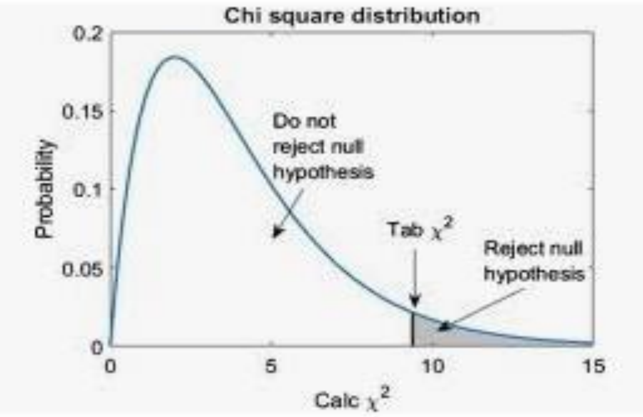

Figure 2: Hypothesis test of Chi Square algorithm

Most appreciable principle of chi-square algorithms is the comparison between the observed values and expected values in data mining. Following the principle data mining procedure helps to develop the structural decision tree to accelerate the operational execution of the business. Compacting the data set from a large number of data collections can be referred to as the data reduction decision tree in data mining. Records and different dimensions are considered as the data sources for the data reduction method [17]. Dimension based data reduction procedure has followed the principle component analysis, linear discriminant analysis, factor analysis, truncated singular value decomposition methods to collect the relevant data for business data mining. These all kinds of data mining algorithms considering the decision trees algorithms have been used in the different field of professions, especially in the commercial belt to add extra competitive advantage to the profile of organizations in the global aspect.

Advantage and disadvantage of data mining including decision tree

The aim and purpose of decision trees in data mining has been focused on collecting linear data sets from large data numbers to predict symbolic problems. However, advantages and disadvantages both are occurring in the decision trees in data mining. Interpretability is one of the major advantages in data mining that enforces less data preparation. Versatility of data sets increases the accuracy of data classification and stables the structure of decision trees in data mining. Non-parametric characteristic of decision trees helps to enhance the quality of data mining within a short time along with in a cost-effective manner. Machine learning is mostly influenced by decision trees in data mining [18]. Easy accessibility and availability of decision trees in data mining helps to understand and interpret the data in a simple way. Unique features of decision trees of data mining to control both categorical and numerical have enhanced the accuracy of data sets.

Easier detection of the errors makes the operation more authentic and accurate and decrease the probability of interruption of the execution flow of the business. Failure of solo nodes will not affect other nodes of decision trees considering the identification or errors ability of decision trees in data mining [19]. Beside this, over fitting of the data sets create a mess in selection of accurate data to implicate in business platforms. Data reduction and re-sampling is another major issue for the data mining procedure including decision trees. Sometimes optimization of data mining can create instability in the structure of decision trees.

A small change can hamper the accuracy of data mining following the structure of decision trees. Maintaining and configuring the decision trees in data mining has created difficulties in the data implications in various field-based operations. Obstacles can be generated during installing a topology network of trees in data mining [20]. The length of segments in tree topology is limited, which limits the cabling use of segments. On the other hand, tree structure has the tendency to grow on the basis of sampling that increases the probability of over fitting data. Evaluation of results can generate inaccurate data mining outcomes due to the over fitting of data in decision trees.

Discussion

Interpretation of the data in this study has helped to evaluate the importance of data mining including decision trees and decision rules. Impact of decision trees in decision making has a positive perspective to solve problems in the execution process of the different professional fields. Generating linear data sets has been driven by decision trees in data mining. Innovation capability of globalized organizations has been increased by the implication of data mining including decision trees to reach the goal of the objective. Commercial and industrial fields are mostly benefited with the implication of data mining considering the decision trees. Decision rules also have been discussed over here to highlight the automatic configuration of the selection of data and mapping of the data following the market trends. Considering the economics, decision trees and decision rules both have a positive perspective and importance.

Classification and regression of data sets also have been focused in the data mining that has helped to meet the subject of the study. Customization of data mining also plays a vital role in the different fields of professions such as machine learning, computer science operations and educational learning procedures for the students. Accuracy and time saving approach of data mining customization also has increased the creativity ability of different globalized organizations. Reduction of computing, errorless mapping and enhanced quality of programming framework has been a creative initiative of data mining customization followed by decision trees. Big data customization also has impacted the data mining procedure that has been interpreted in the study. Entrepreneurs and existing sustainable business organizations have been benefited with the implication of customized data mining. Customer satisfaction also has been to some extent driven by data mining optimization.

Different types of methods of decision trees have been discussed over here considering the characteristics of the decision trees techniques. CART, ID3, chi-square algorithm along with reduction of data has been depicted here and compares the abilities for better implication of the data mining. Highlighting the records of data and dimension-based data retrieving information have been focused by the reduction of data decision trees that has increased the ability of compacting a large number of datasets according to its needs. Considering all the facilities CART has been appreciated among all the methods of decision trees. Building up a model for data mining analysis is the main motive of decision trees. Observed values and expected values of data mining can be evaluated with the implication of the chi-square method of data mining. Classifiers of data mining are more efficient in the CART method rather than the ID3 method.

Advantages and disadvantages also have been highlighted and interpreted here to meet the aim of the study with proper informative justification. Linearity efficiency of the decision trees has highly impacted the data mining procedure. Most influential advantage of data mining is to understand and interpret the data in a short period of time maintaining the accuracy of the data set. On the other hand, small changes in the data set can hamper the decision tree's stability considering data mining. Short time detection of the errors has been executed by the data mining applications including decision trees. Optimization of decision trees often creates hazards that hamper the structural configuration of the decision trees. Over fitting of data sets also creates inaccuracy in the decision trees in data mining that delays the problem detection of the organizational operations. Evaluating the result of the study can be interpreted as data mining considering decision trees determines the competitive advantage for organizations in recent days.

CONCLUSION

The study has evaluated the impact of decision trees in data mining considering different fields of professions. Secondary data collection has been selected in this study considering the inductive research approach and cross-sectional research design. Thematic data analysis has been used in this study to interpret the data briefly. Decision trees concept has been briefly discussed in this study highlighting the data mining procedure. Principles and objectives of decision trees have been depicted in this study to signify the subject of the study with proper justification. Conceptualization of data mining and its importance in recent days has been shown in this study to establish a strong relationship between the data mining and decision trees. Different types of decision trees have been highlighted in this study to evaluate the different methods of decision trees. Customization of data mining techniques including decision trees has been depicted here to prioritize the application of it in different professional fields such as machine learning, computer science and educational learning.

ID3, CART, Chi-Square decision trees algorithms are discussed over here vastly considering the principles, characteristics and comparison of different decision trees. The study also has been focused on the advantages and disadvantages of different decision trees in data mining that enhance the quality of the study. Time saving and cost-effective approach of decision trees have determined the operational execution process of business and accelerates the way to reach the business goal quickly. Data mining helps to compact data from large amounts of data from where organizations have selected the proper data for the wellbeing of the companies including predicted symbolic risk factors to resolve the problems. Customer relevant data information retrieving is another effective approach of decision tree in data mining that helps to maintain the market trends and meet the customers demand to make a sustainable business in the global aspect. Considering all consequences, it can be concluded that data mining including decision trees has an influential impact on retrieving data information to solve problems instantly in operational function.

REFERENCE

[1] Bhagwat, Ganesh, et al. "Novel Static Multi-Layer Forest Approach and Its Applications." *Mathematics* 9.21 (2021): 2650.

[2] Sunday, Kissinger, et al. "Analyzing student performance in programming education using classification techniques." *International Journal of Emerging Technologies in Learning (iJET)* 15.2 (2020): 127-144.

[3] Weston, Sara J., et al. "Recommendations for increasing the transparency of analysis of preexisting data sets." *Advances in methods and practices in psychological science* 2.3 (2019): 214-227.

[4] Krueger, Anne Elisabeth, et al. "Guided user research methods for experience design—a new approach to focus groups and cultural probes." *Multimodal Technologies and Interaction* 4.3 (2020): 43.

[5] Baker, Courtney N., et al. "Validation of the Attitudes Related to Trauma-Informed Care Scale (ARTIC)." *Psychological Trauma: Theory, Research, Practice, and Policy* 13.5 (2021): 505.

[6] Verma, Anurag Kumar, Saurabh Pal, and Surjeet Kumar. "Classification of skin disease using ensemble data mining techniques." *Asian Pacific journal of cancer prevention: APJCP* 20.6 (2019): 1887.

[7] Wang, Ruohan, et al. "Dynamic prediction of mechanized shield tunneling performance." *Automation in Construction* 132 (2021): 103958.

[8] López, Cefe. "Artificial Intelligence and Advanced Materials." *Advanced Materials* (2022): 2208683.

[9] Siderska, Julia. "The adoption of robotic process automation technology to ensure business processes during the COVID-19 pandemic." *Sustainability* 13.14 (2021): 8020.

[10] Triguero, Isaac, et al. "Transforming big data into smart data: An insight on the use of the k-nearest neighbors algorithm to obtain quality data." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9.2 (2019): e1289.

[11] Li, Shugang, et al. "Text Mining of User-Generated Content (UGC) for Business Applications in E-Commerce: A Systematic Review." *Mathematics* 10.19 (2022): 3554.

[12] Kumar, Biresh, et al. "E-Commerce Website Usability

Analysis Using the Association Rule Mining and Machine Learning Algorithm." *Mathematics* 11.1 (2022): 25.

[13] Lopez-Vega, Henry, and Nicolette Lakemond. "Tapping into emerging markets: EMNEs' strategies for innovation capability building." *Global Strategy Journal* 12.2 (2022): 394-417.

[14] Zhao, Long, Sanghyuk Lee, and Seon-Phil Jeong. "Decision tree application to classification problems with boosting algorithm." *Electronics* 10.16 (2021): 1903.

[15] Mozo, Alberto, et al. "Chlorophyll soft-sensor based on machine learning models for algal bloom predictions." *Scientific Reports* 12.1 (2022): 1-23.

[16] Reddy, G. Sekhar, and Suneetha Chittineni. "Entropy based C4. 5-SHO algorithm with information gain optimization in data mining." *PeerJ Computer Science* 7 (2021): e424.

[17] Salo, Fadi, Ali Bou Nassif, and Aleksander Essex. "Dimensionality reduction with IG-PCA and ensemble classifier for network intrusion detection." *Computer Networks* 148 (2019): 164-175.

[18] Aydin, Nezir, and Gökhan Yurdakul. "Assessing countries' performances against COVID-19 via WSIDEA and machine learning algorithms." *Applied Soft Computing* 97 (2020): 106792.

[19] Brunello, Andrea, et al. "J48SS: A novel decision tree approach for the handling of sequential and time series data." *Computers* 8.1 (2019): 21.

[20] Yang, Aimin, et al. "Review on the application of machine learning algorithms in the sequence data mining of DNA." *Frontiers in Bioengineering and Biotechnology* 8 (2020): 1032.