A Review on Data Mining Algorithms Based on Decision Trees: ID3 and C4.5

Sheenam, Aanshi Bhardwaj

Abstract— Data mining is a process of detection of valuable data (information) from massive data. It aids in exploring various patterns and rules from the given data. It is helpful for several purposes in private and public sectors. Many Industries use Data Mining for extract the valuable information from the large database to increase research, reduce price and enhance sales i.e. banking, medicine, insurance and retailing. Techniques of data mining are Association Rules, Classification, Clustering, Decision Trees. Classification is a process of classifying the data based on training set and class labels. It is a supervised learning technique. Decision Tree constructs a tree like structure that anticipates the target variable value. Each internal node of the tree represents the input variables. These variables are linked to child node based on the value of those variables. Last node of the tree is the leaf node which contains the value of the result of target variable based on the input values. The most commonly practiced decision tree algorithms are ID3 and C4.5.The intent of this study is to scrutinize these decision tree algorithms. At first we present concept of Data Mining, Classification and Decision Tree. Then we present ID3 and C4.5 algorithms and we will make comparison of these two algorithms.

Index Terms—Classifiacation, Data Mining, Decision Tree, Traffic-accident.

I. INTRODUCTION

Data mining process is a recursive process which normally comprises of following sub processes

i. **Problem Definition**- to formulate the problem.
ii. **Data Exploration**- to diagnose the data.
iii. **Data Preparation**- to sample and transform the data.
iv. **Modeling**- to create the model for the data.
v. **Evaluation**- to test and evaluate the data.
vi. **Deployment**- to prepare the custom reports and also consists of various external applications.

Classical data mining techniques comprise classification of different users, discovery associations between different item for consumption or customer actions and clustering of users.

Classification is a process of organizing or grouping of data for its efficient and effective use. Classification follows two step processes.

- First step, build a model from the training data having values of the class label already known. It is just like constructing a classifier.
- Second step, check the preciseness of the training data with the help of new data. If it gives satisfactory results then this model can be used to classify the class labels which are unknown. it is just like using classifier for classification.

Decision Trees

Decision tree is one of the techniques of classification. It is basically a tree like structure. Internal nodes of the tree consist of splits and splitting attributes. Each node represents test on an attribute and the edges flowing in and out are contains the results of the test. Last node of the tree is called as the leaf node which represents the class label. With the help of training set, decision tree is constructed. This decision tree can be further used to classify the tupples having unknown class label. Based on some predefined attributes, source dataset of the decision tree is divided into subset. It is repeated on every subset i.e. each subset is further divided into small subset. This process is termed as recursive partitioning. It is completed when the same value of the target variable is there at subset of node, or when there is no beneficial result after splitting.

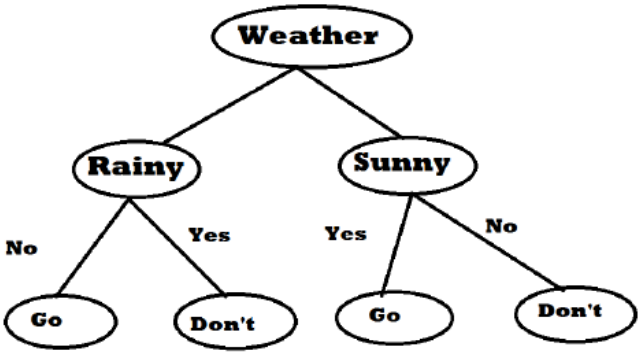

Fig 1: To show whether to go on a trip or not based on weather.

II. LITERATURE SURVEY

Data mining for safety transportation by means of using Internet survey, [1] For smart city application, this research proposes a vehicle infrastructure cooperative function as an illustration which would be assimilate into the safety system of vehicles. Support safety function due to flexibility in the position of the driver can be considered as one of the main functions which can combine road traffic safety with road infrastructure in an interactive way. Accordingly, through the analysis of data of experiences on traffic incidents, this research fond the main cause traffic accidents. There were two sources of data collection- direct interviews and internet. Analysis revealed that just before the accidents there were two major factors of a driver's unconscious states- haste, distraction. Eyes and head moements, and heart rate were also taken into possession as physiological information. AdaBoost and Error-Correcting Output Coding (ECOC) was set up using pattern recognition for detection of driver's cognitive

Revised Version Manuscript Received on May 06, 2016.

Student, Sheenam, Department of Computer Science, Lovely Professional University, Phagwara, India.

Asst. Prof. Aanshi Bhardwaj, Department of Computer Science, Lovely Professional University, Phagwara, India.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

distraction. For smart city applications, this study proposed a conceptual combination of potential intelligent safety function and Intelligent Transportation System service(ITS).

A data mining framework to analyze road accident data, [2] To pinpoint the leading part for the traffic mishaps is the fundamental target of accident data analysis. Nevertheless, divergent data leads it to a difficult task. Mostly preferred process for this is to segment the data. This framework proposed in this paper has used K-modes technique of clustering as a former segmentation step. The data set taken is of Dehradun(India) of round about 11,574 road accidents. The data set is from 2009 to 2014. After this, the next step is to apply association rule for mining the data set. This discover varied situations related to the accident incidents for K-modes algorithm of clustering and (EDS)Entire Data Set. Conditions of accidents are revealed with the rules defined for that within that specific cluster. The detections of both are then compared. The outcomes of this proclaim that this amalgamation results in vital information which is kept invisible if there is no data segmentation in the former step. In addition to this, EDS accidents (Entire Data Set) conducted a trend analysis. This helped in visualizing various clusters for various trends. The assertive approach is displayed with EDS. This methodology of trend analysis is also performed on hourly and monthly basis. Segmentation of accidental data is pivotal as it is shown by Trend analysis.

Review of Decision Tree Data Mining Algorithms: ID3 and C4.5, [3] This paper presents a review of the data mining concept and its two algorithms ID3 and C4.5 and their comparison details. To extract meaningful data from a large amount random shuffled data is done with the help of data mining. It basically uncovers meaning hidden patterns form the series of gigantic data. Data mining techniques include classification, clustering, association rules. In this paper classification techniques have been explored much in detail. Decision tree is a type classification method which defines the value of class variable based on input values. Decision tree algorithms include ID3, Reptree, Decision stump and C4.5. The key objective of this paper is to give a review of ID3 and C4.5 algorithms. The results reveal that C4.5 has better execution results then ID3. Also the error rate is less in C4.5.

Data mining model-decision tree for detecting emotions color, [4] The objective in search the human emotion is Affective computing. It poses a hint of the behavior instance. Hence, should be comprised at the time a machine learning system in the model that intends to replicate or predict human responses. C4.5, the decision tree model with emotion mode of Thayer's and color theory regarding an emotion detecting system are combined and manipulated. In this research, data set of 320 is taken and is divided into four emotion groups for training and building of decision tree. The result reveals the effectiveness of C4.5 to classify the emotion with the feedback color of humans. The affect of emotions was because of several factors. So, it leads to various patterns of colors. For further research emotions can be estimated by different means like facial expressions, speech modulation etc or to use this detection model in real life situation to estimate emotions.

Comparison of Classification Techniques for predicting the performance of Students Academic Environment, [5] This research is about predicting the performance of the student by comparing different classification techniques. In addition to this, it also assists in analyzing the slow learner or student performing poor in the end term are aided to perform better to get the goal at the end of the semester. This can be managed on several attributes. The Research focuses on analyzing the expertise on skill based on performance in academics with scope of knowledge by Prediction/ classification techniques. Besides this, various performance based algorithms have also been compared like C4.5, Naïve Bayesian classifier algorithm,Multi Label K-Nearest Neighbor algorithm, AODE for getting the accurate classification and decision tree algorithm for evaluation of performance of the students in the tool WEKA. Among them, the most accurate outcomes are from Multi Label K-Nearest Neighbor algorithm. For any future research students skills and their cognitive level can be found by using some methods like perceptual method or observation method. It can also be analyzed by the student feedback during classes. Neighbor algorithm can also be implemented on the data set.

Study on the reduction effect of traffic accident by using analysis of Internet survey, [6] There is a great declination in traffic accidents in Japan for the past twelve years. Safety measures in the inactive and preventive areas are attempted to escalate. With respect to passive, airbag system, seat belt and impact of accident on vehicle resulted in a vast reduction consequence. Preventive safety measures prove more fruitful if considered with this. Lately to reduce the traffic accident, driver's psychosomatic state adaptive driving support safety system was introduced. Collection and analysis of traffic incident data is anticipated to judge the effect of reduction of the traffic accident of psychosomatic adaptive safety function. To achieve this goal, this study initiated online (internet) survey by distributing the questionnaires to various people. The results revealed haste and distraction are the two fundamental subconscious state of driver. Initially the eye was kept on cause of severe accidents caused by distraction of a focused driver. The accuracy was measured with the help of pattern recognition. Both ASV (Advanced Safety Vehicle) and Intelligent Transportation Systems (ITS) were considered to analyze the traffic accidents due to distraction of a driver.

Large Imbalance Data Classification Based on MapReduce for Traffic Accident Prediction, [7] Traffic issues are the most viral issues in the today's world. Traffic Accidents can be predicted by classification analysis in which a learning model is built based on the data given. But it has some problems like hardware resource requirement are vast because the dataset is of large size data set. So there will be a hindrance in predicating precisely about the traffic accidents. Besides this, the data is of two types-one related to traffic accidents and other not suitable for traffic accidents. This paper is about settling all these issues. In this Hadoop framework is advised. This basically is used for efficient processing and analyzing large amount of data. It also supports sampling to refine the problem of data imbalance. Primarily, the data is preprcoessed and then analyzed to get data as a learning model. Correction of imbalanced data is achieved by sampling. For better and efficient prediction of the corrected data, this data is distributed among multiple

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

groups and on those groups classification is applied. This classification analysis is done by Hadoop framework.

Evolution of Traffic Flow Analysis under Accidents on Highways Using Temporal Data Mining, [8] In this paper Temporal Data Mining is suggested for analysis of traffic flow under accidents. It depicts the progress of flow in traffic when there are accidents on highways. It used the concept of Cell Transmission Model for the construction of Time series model to give a picture of flow of traffic by using ternary numbers. Euclidean Distance has some drawbacks due to which it, in the domain of time, does not consider the linear drift. To defeat this Discrete Fourier Transformation was instigated. This assisted in directing the time series to frequency domain from time domain. It reduced the calculation cost as well. An algebraic experimentation concluded the usefulness of the method proposed in this paper. The novel approach mentioned here has superior impended approach for the traffic flow and the influence of highway crashes. To check it's feasibility of the proposed technique, it is used on the real data of Harbin in the case study.

Mining traffic data for road incidents detection, [9] The approach proposed in this paper has a more organized feature extraction method which is based on the constantly changing features of data corresponding to those vehicles that are included in road traffic incidents. In this work it has been observed that how dynamic characters of data which is to be measured can be used for obtaining features that gives much measurable improvement of the road traffic incident detection rate with the help of a Support Vector Machines classification approach. All this is performed on the tool weka. The latter classification approach is considered as one among the most accurate solutions that is widely applicable to deal with specific road traffic incident detection problems. In this paper experiments are conducted for the appropriate evaluation which involved comparison of all known methods, for estimating the effect of the feature selection technique proposed on the basis of accuracy of the detection rate. The results reveal that the proposed approach proves to be more precise and has high speed in traffic incident detection. But it does not affect much on the false alarm rate. There is also a description of how the approach mentioned in this paper can be applicable to generic context of ITS. It results in efficient urban management.

A Decision Tree approach for traffic accident analysis of Saskatchewan Highways, [10] The paper provides the enhanced design facility and program which recognize the maximum cause of collisions among traffic on highway based on the study on human population and how they change. Data is collected from Saskatchewan, Canada, comprising rural and urban highway's data of over last 20 years. It exhibits a hybrid data mining model of C4.5 and ID3 algorithms. This involved auditing of data taken form collision of traffic. The outcome of this developed model can easily classify the factors contributing in traffic collision with accuracy. Tool used in this paper is WEKA. The results generated with this proposed algorithm are verified specimen which are precisely classified then those which are not by the traditional algorithm in tool WEKA The main objective is to examine the accidental data and generate decision tree for this analysis. The algorithm defined is self-developed whose results are measured in WEKA. Three facets considered are age, gender and season.

Mining traffic accident features by evolutionary fuzzy rules, [11] This lead analysis and modeling of traffic accident a good approach to develop data models on the real data set collected from the real-world data. In this paper, traffic accidental data set from Ethiopia is taken and treated using artificial evolution and fuzzy systems. It performed mining on the symbolic representation of selected attributes of the traffic accident data set. In this research genetic programming is used to mine the fuzzy rules. This study did a comparison of fuzzy rules to get classifiers for various attributes included. As a future research, there is an opportunity to apply the defined methods on other traffic accident data set.

Analysis of spatial autocorrelation for traffic accident data based on spatial decision tree, [12] Conventional Scientific methods are not enough to discover the novel, hidden patterns and their interconnections of geographical datasets as the span, magnitude and coverage of the geographic datasets is rapidly growing and in the research perspective it has gained a lot of interest in the last some years. In this paper, the key task is to estimate the performance or the process of execution of the conventional and spatial data mining methods. Classification analysis is done using Decision Tree approach. Algorithm used is the Iterative Dichotomiser 3) ID3. It constructs the traditional and spatial Decision trees. Two datasets are used for this research-synthetic Spatial accident dataset and real accidental dataset. After experiments SDT(Spatial DT) was shown having more significance in decision making purpose.

Using Decision Trees to Extract Decision Rules from Police Reports on Road Accidents, [13] According to the World Health Organization (WHO) traffic mishaps are considered as the major public health problem across the globe. So the safety managers are always trying to discover/recognize the main points that lead to as consequence of road accidents. In this paper, Decision Trees (DTs) which is one of the data mining techniques have been used to identify these factors. Traffic accidental dataset, in this paper, are taken from the province of Granada which is in Spain It has been properly analyzed. Decision Trees lead to extraction of the rules. These rules can be useful in campaigns organized for road safety. The results of the proposed algorithm are better. Although the proposed algorithm has some falls as it takes more computation time from the conventional one and is also not optimized locally. It is computation time can be improved as a task of future research on this paper. Multi-criterion Pruning Approach can be extended for critical or complex situations. The proposed algorithm can be evaluated on the basis of performance on other database with medical dataset.

III. DECISION TREE ALGORITHMS

Data Mining uses Decision Tree learning methods most commonly. The goal is create a model to predict value of target variable based on input values. Training dataset creates tree and test dataset aids to test accuracy of the decision tree. First, an attribute to split data in an efficient way is selected as

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

root node for creation of tree. Splitting attribute is the one about which maximum information is there. Three steps for Decision tree algorithm:

1. Selection of an attribute as target class to split tupples in partitions from a given dataset S.
2. Elect a criterion for splitting to generate a partitioning of all tuples belong to a single class. Choose best split to create a node.
3. Repeat the above steps iteratively until the completion of the tree or any stopping criterion is fulfilled.

ID3 (Iterative Dichotomiser 3): J.R. Quinlan, 1986 came up with this ID3 algorithm. As a splitting criterion ID3 uses Information gain. Root node which is the topmost decision node is the best predictor. The attribute with largest Information Gain is selected as split attribute. This helps in building tree from training instances. The tree formed is utilized to classify test data. It grows until the information gain approaches to zero or all instances belong to single target [15].

ID3 works in three simple steps:

1. First, target attribute is selected and entropy of attributes is calculated.
2. Attribute with the maximum information gain is selected.
3. Create node containing that attribute. Perform these steps repeatedly to create new tree branches and then stop after checking the stop criterion.

ID3 works in a top-down approach. It makes decision by using two concepts [15]:

1. Entropy
2. Information Gain (as referred to as just gain).

With these two concepts, creation of nodes and splitting criterion can be determined.

Entropy

Entropy is degree of randomness of data. It calculates the homogeneous attributes among the data. Zero entropy indicates that the whole data is homogeneous and is entropy is one then data is completely uncertain.

Information Gain

Information gain is defined as change in entropy. It basically selects highest information gain attribute for splitting.

$$ET(X, S) = \sum_{j=1}^{k} \frac{|Sj|}{|S|}.ET(Sj)$$

$$IG(X, S) = E(S) - E(X, S)$$

C4.5

C4.5 algorithm is an extension to ID3.C4.5 can handle continuous input attribute. It follows three steps:

1. Categorical attribute follow the same splitting criterion as of ID3 algorithm. Binary splits are generated by continuous attributes.
2. Selection of the attribute with highest gain ratio.
3. Perform these steps iteratively to create new branches of tree and stop when the stop criterion is met.

C4.5 uses a selection criterion which is Gain ratio and it is less biased in comparison to Information gain for large number of values.

$$GR(X, S) = \frac{IG(X,S)}{SI(X,S)}$$

$$\mathbf{SI}(X, S) = -\sum_{j=1}^{k} \frac{|Sj|}{|S|} \log \frac{|Sj|}{|S|}$$

Advantages of C4.5 over ID3:

1. C4.5 can handle both discrete and numerical attributes
2. C4.5 is also capable of handling missing value attribute.
3. C4.5 can implement pre and post pruning concepts which can avoid over fitting of decision tree.

IV. PROBLEM FORMULATION

In this study, an enhanced algorithm, Adta(Advanced decision tree algorithm), is developed from the traditional C4.5 algorithm. The algorithm is investigates the following equations which are explained briefly: information gain, gini index, likelihood ratio chi-squared statistics, gain ratio, and distance measure.

- Information Gain (IG): Information gain is based on Claude Shannon's work on information theory. InfoGain of an attribute A is used to select the best splitting criterion for that attribute. The attribute with the highest InfoGain is selected to build the decision tree.

$$InfoGain(A) = Info(D) - InfoA(D) \quad \text{.....eqn.3.1}$$

- Gini Index (GI): The Gini index is a criterion based on an impurity that measures the divergence between the probability distributions of the target attributes values.

$$GiniIndex(D) = Gini(D) - \sum_{j=1}^{v} pj\ X\ Gini(Dj) \quad \text{.....eqn3.2}$$

- Gain Ratio (GR): Gain ratio is the process in which the decision tree is biased against considered attributes with a huge number of distinct values. So it resolves the downside of information gain.

$$GainRatio(A) = \frac{InfoGain(A)}{SplitInfoA(D)} \quad \text{.....eqn3.3}$$

- Distance Measure (DM): Distance measure performs normalization. Like Gain Ratio, it normalizes the impurity criterion (GI). But it suggests normalizing it in an alternative way.

V. CONCLUSION

In this Research paper, we presented classification technique decision tree. We presented decision tree algorithm ID3 and C4.5.We focused on key elements of construction of decision tree. We did comparison of ID3 AND C4.5 algorithms. It is concluded that C4.5 is more advantageous for mining a data set. Also there is one problem definition. It is an advanced version of C4.5 decision tree algorithm.

REFERENCES

1. M. Miyaji, "Data mining for safety transportation by means of using Internet survey.", in In Data Engineering Workshops (ICDEW), 2015 31st IEEE International Conference, 2015, pp. 119-123, 2015.
2. S. Kumar and D. Toshniwal, "A data mining framework to analyze road accident data. Journal of Big Data", 2015, pp. 1-18, 2015.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

3. D. Kaur, R. Bedi and D. Gupta, "Review of Decision Tree Data Mining Algorithms: ID3 and C4.5", in International Conference on Information Technology and Computer Science, 2015, pp. 5-8, 2015.
4. M. Lee and G. Chen, "Data mining model-decision tree for detecting emotions color", in In Ubi-Media Computing and Workshops (UMEDIA), 2014 7th International Conference. IEEE, 2014, pp. 226-230, 2014.
5. M. Mayilvaganan and D. Kalpanadevi, "Comparison of Classification Techniques for predicting the performance of Students Academic Environment", in Communication and Network Technologies (ICCNT), 2014 International Conference, IEEE, 2014, pp. 113-118, 2014.
6. M. Miyaji, "Study on the reduction effect of traffic accident by using analysis of Internet survey.", in In Internet of Things (WF-IoT), 2014 IEEE World Forum, 2014, pp. 325-330, 2014.
7. S. Park and Y. Ha, "Large Imbalance Data Classification Based on MapReduce for Traffic Accident Prediction.", in In 2014 Eighth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2014, pp. 45-49, 2014.
8. Shi, Z. Tao, Z. Xinming and W. Jian, "Evolution of Traffic Flow Analysis under Accidents on Highways Using Temporal Data Mining", in In2014 Fifth International Conference on Intelligent Systems Design and Engineering Applications (ISDEA), 2014, pp. 454-457, 2014.
9. E. Gakis, D. Kehagias and D. Tzovaras, "Mining traffic data for road incidents detection.", in In Intelligent Transportation Systems (ITSC), 2014 IEEE 17th International Conference. IEEE,, 2014, pp. 930-935, 2014.
10. X. Zhang and L. Fan, "A Decision Tree approach for traffic accident analysis of Saskatchewan Highways", in In Electrical and Computer Engineering (CCECE), 2013 26th Annual IEEE Canadian Conference, IEEE, 2013, pp. 1-4, 2013.
11. P. Kromer, T. Beshah, D. Ejigu, V. Snasel, J. Platos and A. Abraham, "Mining traffic accident features by evolutionary fuzzy rules", in In Computational Intelligence in Vehicles and Transportation Systems (CIVTS), 2013 IEEE Symposium, 2013, pp. 38-43, 2013.
12. Ghimire, S. Bhattacharjee and S. Ghosh, "Analysis of spatial autocorrelation for traffic accident data based on spatial decision tree", in In Computing for Geospatial Research and Application (COM. Geo), 2013 Fourth International Conference,IEEE, 2013, pp. 111-115,2013.
13. L. Griselda, D. Juan and A. Joaquín, "Using Decision Trees to Extract Decision Rules from Police Reports on Road Accidents", Procedia - Social and Behavioral Sciences, vol. 53, pp. 106-114, 2012.
14. Desai1Sharmishta, Dr. PatilS.T, "Efficient Regression Algorithms for Classification of Social Media Data", In Pervasive Computing (ICPC), 2015 International Conference, IEEE, pp. 1-5, 2015.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.