IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 2, February 2021

DOI 10.17148/IJARCCE.2021.10224

Classification Techniques in Data Mining

Nita J Goswami[1], Asst. Prof Ketan Patel[2]

Student of M.E, C.E Dept,Gujarat Technology University, Himatnagar, Gujarat, India[1]

Asst Prof. Grow More Faculty of Engineering, Himmatnagar, Gujarat, India[2]

Abstract: This paper discusses the data mining and various data mining techniques of classification. The paper also describes the data mining strategies and the limitation of the data mining. Various classification techniques covered in the paper are based on the decision tree. The decision tree based classification J48, CART and ID3 are discussed in the paper. The paper is useful to discuss compares the decision tree based classification techniques and to select the useful classification technique according to requirement.

Keywords: Classification, decision tree, J48, ID3, CART.

I. INTRODUCTION

Today’s manufacturing, engineering, business, and computing processes in public and private organizations around the world are generating massive amounts of data. This explosive growth of data has outpaced the ability to interpret and digest the data. Therefore, data mining techniques and tools for automated data analysis and knowledge discovery are needed. Today’s enterprise data warehouse (EDW) focuses on developing, extending, and incorporating knowledge discovery technology into tools and methods for data analysis, including aspects of data modeling, algorithms, and visualization Knowledge gained by uncovering relationships and structure in the data will enable better understanding of customers, suppliers, and internal as well as external processes. This helps process owners to identify problems, reduce defects, and improve cost, aiding continuous quality improvement. The proliferation of computer databases, online automatic data collection and increased storage capacity has, in some situations, led to explosive growth in the size and complexity of data warehouses. The gigantic sizes of these enterprise databases overwhelm standard analysis methods, making it difficult to find useful information, knowledge, and relationships in the data There aremany software tool. implementing data mining and knowledge discovery techniques that are designed to work efficiently over large amounts of data and carry out simple analyses to uncover relationships in the data. The users and analysts may then perform further investigations and data analysis to confirm or better understand uncovered relationships.

II TYPES OF DATA MINING SYSTEM

Data mining systems can be categorized according tovarious criteria the classification is as follows

Classification of Data Mining Systems According to the Type of Data Source Mined

In an organization a huge amount of data’s are available where we need to classify these data but these are available most of times in a similar fashion. We need to classify these data according to its type (maybe audio/video, text format etc)

Classification of Data Mining Systems According to the Data Model

There are so many number of data mining models (Relational data model, Object Model, Object Oriented data Model, Hierarchical data Model/W data model) are available and each and every model we are using the different data .According to these data model the data mining system classify the data in the model

Classification of Data Mining Systems According to the Kind of Knowledge Discovered

This classification based on the kind of knowledge discovered or data mining functionalities, such ascharacterization,discrimination,association,classification, clustering, etc. Some systems tend to be comprehensive systems offering several data mining functionalities together.

DOI 10.17148/IJARCCE.2021.10224

Classification of Data Mining Systems According toMining Techniques used

This classification is according to the data analysis approach used such as machine learning, neural networks, genetic algorithms, statistics, visualization, database oriented or data warehouse-oriented, etc.
The classification can also take into account the degree of user interaction involved in the data mining process such as query-driven systems, interactive exploratory systems, or autonomous systems. A comprehensive system would provide a wide variety of data mining techniques to fit different situations and options, and offer different degrees of user interaction.

II. DATA MINING STRATEGIES

Data mining strategies can be grouped as follows:

Classification

Here the given data instance has to be classified into one of the target classes which are already known or defined. One of the examples can be whether a customer has to be classified as a trustworthy customer or a defaulter in a credit card transaction data base, given his various demographic and previous purchase characteristics.

Estimation

Like classification, the purpose of an estimation modelis to determine a value for an unknown output attribute. However, unlike classification, the outputattribute for an estimation problem are numeric rather than categorical. An example can be "Estimate the salary of an individual who owns a sports car?

Prediction

It is not easy to differentiate prediction from classification or estimation. The only difference is thatrather than determining the current behavior, the predictive model predicts a future outcome. The outputattribute can be categorical or numeric. An examplecan be "Predict next week's closing price for the Dow Jones Industrial Average".

Association Rule Mining

Here interesting hidden rules called association rules in a large transactional data base is mined out. For e.g. the rule {milk, butter->biscuit} provides the information that whenever milk and butter are purchased together biscuit is also purchased, such that these items can be placed together for sales to increase the overall sales of each of the items.

Clustering

Clustering is a special type of classification in which the target classes are unknown. For e.g. given 100 customers they have to be classified based on certain similarity criteria and it is not preconceived which are those classes to which the customers should finally be grouped into.
The main application areas of data mining are in Business analytics, Bioinformatics, Web data analysis, text analysis, social science problems, biometric data analysis and many other domains where there is scope for hidden information retrieval. Some of the challenges in front of the data mining researchers are the handling of complex and voluminous data, distributed data mining, managing high dimensional data and model optimization problems .

III. LIMITATIONS OF DATA MINING

As data mining products can be very powerful tools, they are not self-sufficient applications. To be successful, data mining requires skilled technical and analytical specialists who can structure the analysis and interpret the output that is created. Consequently, the limitations of data mining are primarily data or personnel related, rather than technology-related. Even though data mining can help reveal patterns and relationships that does not tell the user the value or significance of these patterns. All these types of determinations must be made by the user. Like, the validity of the patterns discovered is

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 2, February 2021

DOI 10.17148/IJARCCE.2021.10224

dependent on how they compare to "real world" circumstances. Example, to assess the validity of a data mining application designed to identify potential terrorist suspects in a large pool of individuals; user may test the model using data that includes information about known terrorists. While possibly re-affirming a particular profile, that does not necessarily mean that the application will identify a suspect whose behavior significantly deviates from the original model[4][8]. Some other limitation of data mining is that while it can identify connections between behaviors and/or variables that do not necessarily identify a causal relationship. Example, an application may identify that a pattern of behavior, like as the propensity to purchase airline tickets just shortly before the flight is scheduled to depart is related to characteristics such as income, level of education, and Internet use. So, that does not necessarily indicate that the ticket purchasing behavior is caused by one or more of these variables. Actually, the individual's behavior could be affected by some additional variable(s) such as occupation (the need to make trips on short notice), family status (a sick relative needing care), or a hobby.

.

IV. CLASSIFICATION IN DATA MINING

Classification produces a function that maps a data item into one of several predefined classes, by inputting a training data set and building a model of the class attribute based on the rest of the attributes. Decision tree classification has an intuitive nature that matches the user's conceptual model without loss of accuracy. However no clear winner exists amongst decision tree classifiers when taking into account tree size, classification and generalization accuracy.

Decision Tree

A decision tree is a flow chart like structure where each node denotes a test on an attribute value, each branch represents an outcome of the test and tree leaves representclasses or class distribution. A decision tree is a predictive model most often used for classification. Decision trees partition the input space into cells where each cell belongs to one class. The partitioning is represented as a sequence of tests. Each interior node in the decision tree tests the value of some input variable,and the branches from the node are labeled with thepossible results of the test. The leaf nodes represent the cells and specify the class to return if that leaf node is reached. The classification of a specific input instance is thus performed by starting at the root node and, depending on the results of the tests, following the appropriate branches until a leaf node is reached. Decision tree is represented in figure 1.

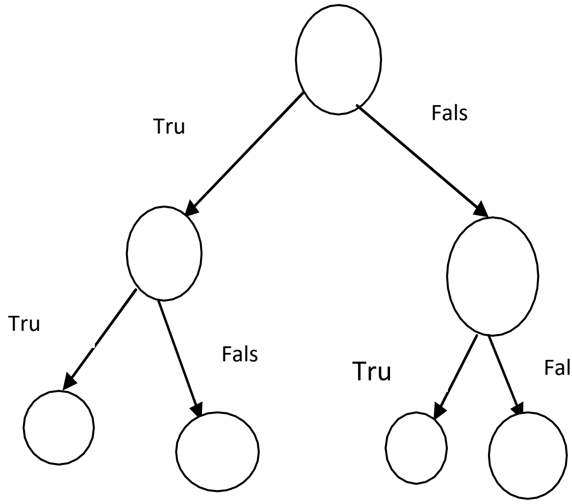

Figure1: Decision Tree

Decision tree is a predictive model that can be viewed as a tree where each branch of the tree is a classification question and leaves represent the partition of the data set with their classification.

V.I. J48

DOI 10.17148/IJARCCE.2021.10224

Decision trees represent a supervised approach to classification which is done recursively. A decision tree is a simple structure where non-terminal nodes represent tests on one or more attributes and terminal nodes reflect outcomes. By first selecting an attribute to be the root node, a branch is made for each possible value. As a result of these splits, one value is associated with each attribute. Recursion then occurs on each branch until all instances on a node have the same classification. A current and popular implementation of this approach is known as Quinlan's C4.5 model; C4.5 is well known as an "industrial strength algorithm" which has been modified and updated over the years[4][9]. The framework's version of C4.5 is called J48. J48 works by first choosing an attribute that best differentiates the output attribute values.

The Algorithm

Stage I: The leaf is labeled with a similar class if the instances belong to similar class.

Stage II: For each attribute, the potential data will be figured and the gain in the data will be taken from the test on the attribute.

Stage III: Finally the best attribute will be chosen depending upon the current selection parameter.

Limitations of J48 Algorithm

A few limitations of J48 are discussed below.

Empty Branches

Constructing tree with significant value is one of the important steps for rule generation by J48 algorithm. In our research, we have come out with many nodes with zero values or very close to that. But, these values don't contribute to create or help to create any class for classification task. Instead it makes the tree wider and still complicating. (Prerna Kapoor, 2015).

Insignificant Branches

Number of chosen distinct attributes produces the samenumber of potential division to build a decision tree. But the fact is, not all of them are significant for classification task. These least important branches not only decrease the usability of decision trees but also bring on the problem of over fitting. (Srishti Taneja, 2014)

Over Fitting

Over fitting happens when algorithm display gets information with exceptional attributes. This causes many fragmentations in the process distribution. Statistically unimportant nodes with least examples are known as fragmentations. Usually J48 algorithm builds trees and grows its branches 'just deep enough to perfectly classify the training examples'. This approach performs better with noise free data. But most of the time this strategy over fits the training examples with noisy data. At present there are two strategies which arewidely used to bypass this over fitting in decision tree learning. (SAGAR, 2015) Those are:

- If tree grows taller, stop it from growing before it reaches the maximum point of accurate classification of the training data.
- Let the tree to over-fit the training data then post-prune tree.

Yet, nothing of those is perfect solution of this problem. So we have proposed two tools to minimize the input space of data in this research. The first tool is Entropy of Information Theory and the second is Correlation Coefficient. In this experimentation, we have examined dengue medical data

Classification and Regression Tree (CART)

CART is one of the more popular methods of constructing the decision tree. It builds a binary decision tree by splitting the records at each node according to a function of a single attribute. The initial split produces two nodes, each of which we now attempt to split in the same manner as the root node. Once again we examine all the input fields to find the splitting candidate. If no split can be found that significantly decreases the diversity of a given node, then we label the node as a

DOI 10.17148/IJARCCE.2021.10224

leaf node. Eventually only the leaf nodes remain and we have grown a full decision tree. At the end of the tree- growing phase, every record of the training set has been assigned to some leaf of the full decision tree. Each leaf can now be assigned a class and error rate.

ID3

This algorithm is based on Hunts algorithm. The tree is constructed in two phases. The two phases are tree building and pruning. ID3 uses information gain measure to choose the splitting attribute. It only accepts categorical attributes in building a tree model. It does not give accurate result when there is noise. To remove the noise pre-processing technique has to be used.

To build decision tree, information gain is calculated for each and every attribute and select the attribute with the highest information gain to designate as a root node. Label the attribute as a root node and the possible values of the attribute are represented as arcs. Then all possible outcome instances are tested to check whether they are falling under the same class or not. If all the instances are falling under the same class, the node is represented with single class name, otherwise choose the splitting attribute to classify the instances. Continuous attributes can be handled using the ID3 algorithm by discrediting or directly, by considering the values to find the best split point by taking a threshold on the attribute values. ID3 does not support pruning.

VI CONCLUSION

The paper studied various decision tree based classification algorithm. The decision tree based classification algorithms are efficient for classifying the dataset. The paper describes three decision tree basedclassification algorithms i.e. ID3, CART and J48. The J48 algorithm seems to be better among three so in future the algorithm can be applied for the big data classification

REFERENCES

[1] S Patnaik, S. K., Sahoo,., & Swain, D. K. (2012). “Clustering of Categorical Data by Assigning Rank through Statistical Approach”. International Journal of Computer Applications, 43.

[2] Shrivastava, V. (2012). “A Study of Various ClusteringAlgorithms on Retail Sales Data”. International Journal, 1(2)

[3] Kanooni, A. (2004). “Survey of Existing Data Mining Techniques, Methods and Guidelines” Within the Context of Enterprise Data Warehouse (Doctoral dissertation, AthabascaUniversity).

[4] Kuang, L., Chen, L., Xie, Y.,& Wu, J. (2013, June). “Full Recognition of Massive Products Based on Property Set”.In Big Data (BigData Congress), 2013 IEEE International Congress on (pp. 294-301). IEEE.

[5] Padhy, N., Mishra, P., & Panigrahi, R. (2012). The Survey of Data Mining Applications And Feature Scope. International Journal of Computer Science, Engineering & Information Technology, 2(3).

[6] Fabricio Voz Ni Ka, Leonardo Vi Ana (2007) "Data mining classification".

[7] Han, J., Kamber, M. 2012. Data Mining: Concepts and Techniques, 3rd ed, 443-491.

[8] https://www.tutorialspoint.com/data_mining/dm_cluster _analysis.htm

[9] International Journal of Advanced Computer Scienceand Applications 7(7) August 2016

[10] Mrs.Sagunthaladevi.S's,Dr.bhupathirajuvenkataramaraju s scientific contributions Classification is a supervised learning technique in datamining