IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Special Issue-17, May 2019 E-ISSN: 2347-2693

Classification Techniques in Data Mining

[1*]M. Dukitha, [2]R. Sindhu

[1,2]Dept. of Master of Computer Applications, Er.Perumal Manimekalai College of Engineering, Hosur

Corresponding Author: dukitham@gmail.com, Tel. - 9486344552

Available online at: www.ijcseonline.org

Abstract— Data mining is a emerging field which has attracted a large number of information , industries due to huge volume of data managed in recent days. Classification techniques to improve business opportunity and to improve the quality of services provided. An classification is one of the most useful and important techniques. Classification techniques are useful to handle large amount of data. Various classification techniques covered in the paper is based on the decision tree. The regression tree based classification J48, CART and ID3.

Keywords: Classification, Decision tree, Regression, Naïve Bayes, J48, CART.

I. INTRODUCTION

Techniques and tools for automated data analysis and knowledge discovery are needed. There are many software tools for implementing data mining and knowledge discovery techniques. The classification techniques can be utilised to mine interesting sets of data from huge data bases and Data Mining refers to extracting or mining knowledge from large amount of data. Classification applied to information management, query processing decision making ,and many other applications.

II. LITERATURE REVIEW

Rathee A. and Mathur R. P. (2013) In this paper, the study of education database is done, which contain hidden knowledge for improving students performance. This paper gives the comparative study of decision tree algorithms like ID3, C4.5 and CART and as a result C4.5 is more accurate. The predictions obtained from the algorithms help the teacher to identify poor students and improve their performance.

Dharker S. and Rajavat A. (2012) By classification algorithms, the authors implement healthy diet recommendation system through web data mining. They compare two decision tree algorithms that are ID3 and c4.5 on the basis of accuracy and time performance analysis. Using these algorithms, they find the user access pattern. Then resulting outcome as ID3 algorithm is worked on each and every instance very exactly. C4.5 takes less time but will not work on each instance.

III. DATA MINING

Data mining means mining or digging deep into data which is in different forms to gain patterns, and to gain knowledge on that pattern. In the process of data mining, large data sets are first sorted, then patterns are known and relationships are reputable to achieve data breakdown and solve problems Classification:

In that describes and distinguishes data classes and concepts. Classification is the problem of identifying to which of a set of categories (sub populations), a new observation belongs to, on the basis of a training set of data containing observations and whose categories membership is known. Example:

Before starting any Project, we need to check it's feasibility. In this case, a classifier is required to predict class labels such as 'Safe' and 'Risky' for adopting the Project and to further approve it.

Two Step Process:
a).Learning Step (Training Phase**):**
Structure of Classification Model Different Algorithms is used to make a classifier by creation the model learn using the training set presented. Model has to be trained for guess of accurate results.

b).Classification Step:
Model used to predict class labels and testing the construct model on test data and hence approximate the correctness of the classification rules.

Types of Attributes:

a) Binary: binary data has only two values/states.

- Symmetric: Both values are equally important in all aspects.
- Asymmetric: When both the values may not be significant.

b).Nominal: When more than two outcomes are possible. It is in Alphabet form rather than being in numeral form. Example: One needs to choose some objects but of different colors. So, the color might be Yellow, Green, Black, Red. Different Colors: Red, Green, Black, Yellow

- Ordinal: Values that must have some meaningful order. Example: Suppose there are grade sheets of few students which might contain different grades as per their performance such as A, B, C, D Grades: A, B, C, D
- Continuous: May have infinite number of values, it is in float type Example: Measuring weight of few Students in a sequence or orderly manner i.e. 50, 51, 52, 53 Weight: 50, 51, 52, 53
- Discrete: Finite number of values. Example: Marks of a Student in few subjects: 65, 70, 75, 80, 90 Marks: 65, 70, 75, 80, 90 Classifiers can be categorized on two major types:
- Discriminative: It is a very basic classifier and determines just one class for each row of data. It tries to model just by depending on the practical data, depends deeply on quality of data rather than on distributions. Example: Logistic Regression Acceptance of a student at a University (Test and Grades need to be considered)
- Generative: It models the distribution of individual classes and tries to learn the model that generates the data behind the scenes by estimating assumptions and distributions of the model. Used to predict the unseen data.
  Example: Naive Bayes Classifier.

IV. Applications

- Marketing and Retailing
- Manufacturing
- Telecommunication Industry
- Intrusion Detection
- Education System
- Fraud Detection.

V. Classification

Classification is a data mining function that assigns items in a collection target categories or classes. The goal of classification is to exactly predict the objective class for each case in the data. For example: A classification model could be used to recognize loan applicants as low, medium, or high credit risks.

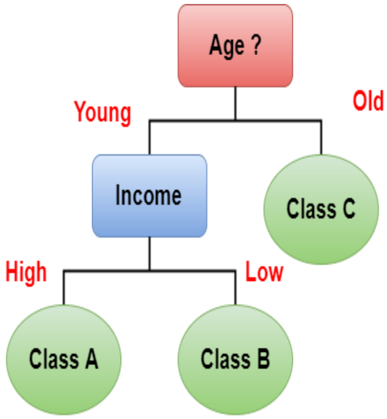

Fig. 1

VI. Classification Techniques

- Decision Trees
- Bayesian Classifiers
- Neural Networks
- Linear Regression

A. *DECISION TREE*

A decision tree is a structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch indicates the outcome of a test, and each leaf node contains a class label. The topmost node in the tree is the root node.

The following decision tree is for the concept buy_computer that indicates whether a customer at a company is likely to buy a computer or not. Each internal node represents a test on an attribute.

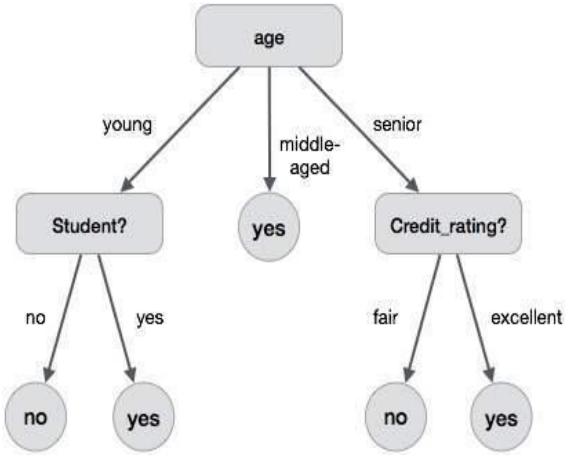

Fig 2 Decision tree

A decision tree consists of three types of nodes:

1. Decision nodes – typically represented by squares
2. Chance nodes – typically represented by circles
3. End nodes – typically represented by triangles

i.* Advantages Of Decision Trees*:

- Are simple to understand and interpret. People are able to know decision tree models after a brief details.
- Have value even with little hard data. Important insights can be generated based on experts relating a situation (its alternatives, probabilities, and costs) and their preferences for outcomes.
- Help determine worst, best and expected values for different scenarios.
- Use a white box model. If a given result is provided by a representation.
- Can be combined with other decision techniques.

ii. Disadvantages* Of Decision Trees*:

1. They are unstable, meaning that a small change in the data can lead to a large change in the structure of the optimal decision tree.
2. They are often relatively inaccurate. Many other predictors perform better with similar data. This can be remedied by replacing a single decision tree with a random forest of decision trees, but a random forest is not as easy to understand as a single decision tree.
3. For data including categorical variables with different number of levels, information gain in decision tree is biased in favor of those attributes with more levels.[7]
4. Calculations can get very complex, mainly if many values are undecided and/or if many outcomes are linked.

B. NAÏVE BAYES CLASSIFIER

- The Naive Bayes classification algorithm is a probabilistic classifier. It is based on probability models that include strong independence assumptions.
- The independence assumptions often do not have an impact on reality. Therefore they are considered as naive.
- A Naive Bayes model consists of a large cube that includes the following dimensions:
- ❖ Input field name
- ❖ Input field value for discrete fields, or input field value range for continuous fields.
- ❖ Continuous fields are divided into discrete bins by the Naive Bayes algorithm Target field value.

❖ **Baye's Theorem:**

Bayes' Theorem is named after Thomas Bayes. There are two types of probabilities −

- Posterior Probability [P(H/X)]
- Prior Probability [P(H)]

where X is data tuple and H is some hypothesis.

According to Bayes' Theorem,

P(H/X)= P(X/H)P(H) /P(X)

- Naive Bayes uses a related method to expect the probability of different class based on various attributes. This algorithm is mostly used in text classification and with problems having multiple classes.
- The Naïve Bayes Classifier technique is particularly suited when the dimensionality of the inputs is high. even though its simplicity, Naive Bayes can often outperform more difficult classification methods. The following example is a simple demonstration of applying the Naïve Bayes Classifier from Stat Soft.
- The objects can be classified as either GREEN or RED. Our task is to classify new cases as they appear (i.e., decide to which class label they belong, based on the currently exiting objects).

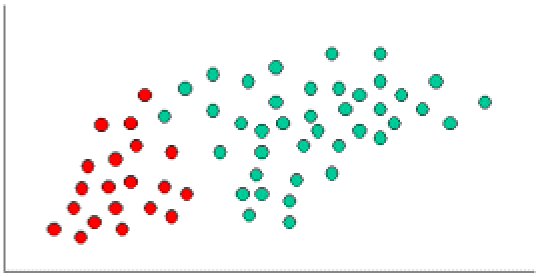

Fig. 3

C. NEURAL NETWORK

Neural networks represent a brain image for information processing. These models are purely inspired rather than an exact replica of how the brain actually functions. Neural networks have been shown to be very promising systems in many forecasting applications and business classification applications due to their ability to “learn” from the data, their nonparametric nature and their ability to simplify. Neural computing refers to a pattern detection methodology for machine learning. The consequential model from neural computing is often called an artificial neural network (ANN) or a neural network. Neural networks have been used in many business applications for pattern detection, forecasting, prediction, and classification. Neural network computing is a key factor of any data mining tool kit.

❖ *.Neural Network Method In Data Mining :*

Neural network method is used for classification, clustering, feature mining, prediction and pattern detection. It imitates the neurons structure of animals, bases on the M-P model and Hebb learning rule, so in essence it is a distributed matrix structure. Through training data mining, the neural network method increasingly calculates the weights the neural network connected.

The neural network model can be generally divided into the following three types:

(a)Feed-forward networks:

It regards the observation back-propagation model and the function network as representatives, and generally used in the areas such as prediction and pattern detection.

(b) Feedback network:

It regards discrete model and continuous model as representatives, and mostly used for associative memory and optimization calculation.

(c) Self-organization networks:

It regards adaptive resonance theory (ART) model and Kohonen model as representatives, and mostly used for cluster analysis.

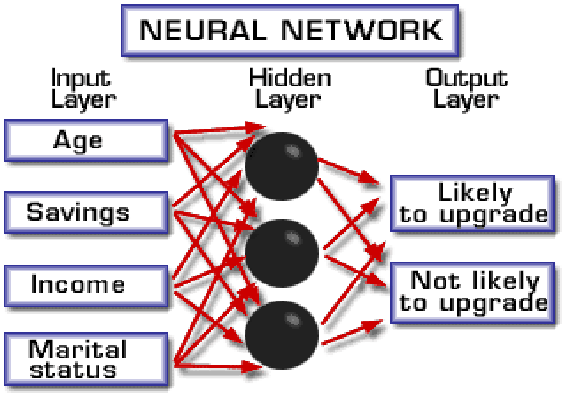

Fig. 4 Neural Network

D. Some Common Mistakes

Regression is a data mining method used to predict a range of numeric values given a particular dataset. For example, regression might be used to calculate the cost of a product or service, given other variables.

- In simple linear regression, we expect scores on one variable from the scores on a second variable. The variable we are predicting is called the *condition variable* and is referred to as Y. The variable we are basing our prediction is called the *interpreter variable* and is referred to as X. When there is only one interpreter variable, the prediction method is called simple *regression*. In simple linear regression the prediction of Y when plotted as a role of X forms a straight line.
- Linear regression consists of decision the best-fitting straight line through the points. The best-fitting line is called a regression line. The black diagonal line in the regression line and consists of the predict score on Y for each probable value of X. The perpendicular lines from the points to the regression line characterize the errors of prediction. As you can see, the red point is very near the regs small. By difference, the yellow point is much higher than the regression line and therefore its error of prediction is large.

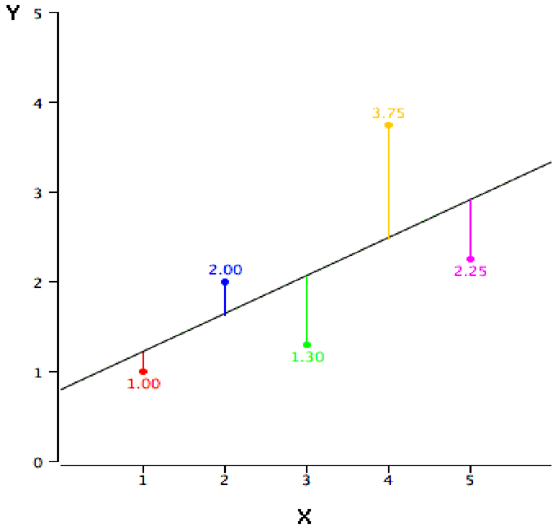

Fig 5: Linear Regression

VII. CONCLUSION

There are numerous proportional studies of the various classification techniques, but it has not been found that one single method is superior compared to others. Issues like accuracy, scalability, training time and many others supply to choosing the best technique to classify data for mining. Data mining is a wide area that integrates techniques from different fields such as machine learning, artificial intelligence, statistics and pattern recognition. Classification methods are naturally strong in modeling communications. Hence these classification methods show that how a data can be strong-minded and grouped when a new set of data is accessible.

REFERENCE

[1]Saranya Vani.M,Dr.S.Uma,Sherin.A,Saranya.K " Survey On Classification Techniques Used In Data Mining And Their Recent Advancements"(ISSN:2278-7798) IJSETR,vol.3,issue 9,september 2014.

[2]Mr.Sudhir M.Gorade,,Prof.Ankit Deo,Prof.Preetesh Purohit "A Study Of Some Data Mining Classification Techniques."(e-ISSN:2395-0056,p-ISSN:2395- 0072)IRJET,vol 4,issue 4,april 2017.

[3] Neha Mida And Dr.Vikram Singh "A Survey On Classification Techniques In Data Mining."(ISSN:2231-5268)IJCSMS,vol 16,issue 1,july 2015.

[4]C.Parimala,R.Porkodi "Clasiification Algorithm In Data Mining"(ISSN:2456-3307) ,IJSRCSEIT,2018,vol 3,issue 1.