IARJSET

ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 8, August 2021

DOI: 10.17148/IARJSET.2021.8874

SURVEY ON DIFFERENT DATA MINING ALGORITHM FOR PREDICTION

Valliammai V[1], Suruthi Selvi S[2], Akshaya Sibi S P[3], Nidharshna M[4], Lavanya U[5]

[1-5] UG - Computer Science and Engineering, Bannari Amman Institute of Technology, Erode

ABSTRACT: Data mining refers to the mining or extracting the useful information from a pool of data. Data mining algorithms are widely used to analyse in Education, Fraud Detection, Criminal Investigation and Bio Informatics. Data mining is a technique uniquely used for identifying unknown pattern and to convert the raw data into the unambiguous information and it is also defined as the finding the in depth data or the hidden information from the databases Huge algorithms has been proposed in the past for the data mining extraction. The various data mining techniques such as classification, clustering, association rule mining, decision tree and more on. The aim of this paper focuses on Different data mining algorithms that are used mostly for prediction and classification.

Keywords—Data Mining, clustering, decision tree, svm, classification

1.INTRODUCTION

The mining of gold from rocks is said to be an rock mining.so data mining should be defined as "knowledge mining from data," and it is also have a different meaning to data mining, such as, knowledge extraction, data/pattern analysis, data archaeology, Knowledge Discovery from Data, or KDD. Due to the evolution in technology there is a large amount of unprocessed data. It is a time dominating to view or process the needed information. In such circumstances there is a need to emerge a strategy which is useful to obtain the necessary data. Since, the volume of data is generated very fast and manual analysis, even if possible cannot keep pace. It is a tedious process. To overcome these pitfalls the concept of Data Mining is used. The techniques and algorithm of data mining will help the users to acquire the essential data. Data mining is the process of deriving hidden predictive Data from vast.

2. PREDICTION TECHNIQUE

Decision Tree
Naïve Bayes
Neural networks
Support Vector Machine
Random Forest algorithm
Logistic regression
K means clustering
Hierarchical clustering,
Agglomerative Clustering
Gaussian mixture model
Mean shift

3. DECISION TREE

There are various methods for classification, the classification algorithm of decision trees is crystalline, easy to understand and easy to convert into certain classification rules. Hence this algorithm is widely used. It depends on the background of "data platform for public petition", which mainly aim to study how data mining combined with other existing databases and extracting useful data from massive hidden in the data, provide better analysis for the decision makers. Currently, Classification is used frequently. The C4.5, ID3, CART decision tree algorithms are explained below-

Characteristics	ID3	CART	C4.5
Type of data	Categorical	Continuous	Categorical and continuous
splitting criteria	lead to multiway split	binary or multiway split	lead to multiway split

IARJSET
ISSN (O) 2393-8021, ISSN (P) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Vol. 8, Issue 8, August 2021
DOI: 10.17148/IARJSET.2021.8874

Boosting	Not supported	Supported	Not supported
Speed	Low	Average	Faster than ID3
Pruning	No pruning	Post pruning	Pre- pruning
Measure	Entropy	Gini	Entropy

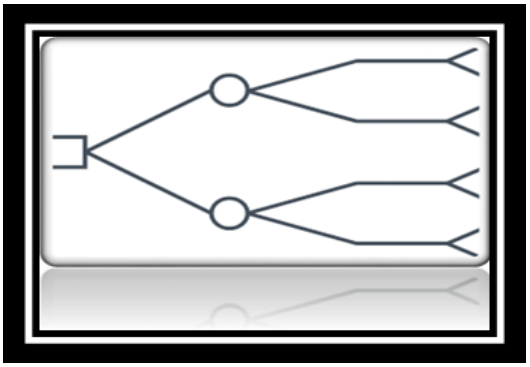

Fig 1. Impact of Decision Tree

4. NAIVE BAYES:-

Naive Bayes algorithm is based on Bayes theorem with naïve independence and strong assumptions. This learning method assumes that the features are independent of the given class. Application of the Naive Bayes algorithm for multiple real-life purposes. It is highly scalable due to the less training data. Continuous and discrete datasets can be handled and work on easily with missing values.

Text classification: Commonly used in probabilistic learning methods for text classification. This classifier is one of the most successful known algorithms when it comes to the classification of text documents.
Spam filtration: This is an example of text classification. Modern email services implement Bayesian spam filtration method.
Sentiment Analysis: Analyse the tone of tweets, reviews and comments—whether they are negative, positive or neutral.

5. NEURAL NETWORK:-

A neural network ensemble is a combined form of a finite number of neural networks, which are trained for a common classification task. Comparatively the ensemble is able to improve the generalization ability of the classifier more efficiently than the single neural network. The loss of data does not affect its working due to the input being accumulated in its own networks instead of a database. Mainly applied in Route detection.

4. Support Vector Machines:-

Support Vector Machines (SVMs) proposed by Vladimir Vapnik within the area of structural risk minimization and statistical learning theory, demonstrated to work successfully on various forecasting and classification problems. Support Vector Machines have the prospective to capture very large features, due to a principle which is based on the Structural Risk Minimization Theory (SRM) i.e., algorithm is based on guaranteed risk bounds of statistical learning theory.

5. Random forest algorithm:-

Random Forest keeps the benefits achieved by the Decision Trees but through the usage of bagging on samples, its by the voting scheme through which decision is made and a random subsets of variables, it more time achieves better results than Decision Trees. The Random Forest is efficiently appropriate for high dimensional data modeling because it can handle continuous, categorical and binary data and can handle missing values in the data. It produces good prediction output based on the most ranked output of the subset decision. In the Land sector it is used to identify the areas of similar lands. In the Marketing field it is used to identify the Marketing trends.

6. LOGISTIC REGRESSION:-

It is a very efficient method when it has linearly separable features in the datasets. This classification provides a solution to the general linear model when the dependent variable is nominal, either a two level(dichotomous),multilevel(polychotomous) or even ordered(polychotomous) response. The main application of this algorithm are Email classification by spam and not spam and used in the Online Credit card transactions. It is less prone to overfitting**.** It is not suitable for Nonlinear problems.

IARJSET

ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 8, August 2021

DOI: 10.17148/IARJSET.2021.8874

7. K MEANS CLUSTERING:-

It is easy to implement, good in capturing the structure of the data. The K-Means algorithm is a kind of cluster algorithm, and has advantages of briefness,certainty and efficiency. This algorithm is unsupervised and usually used in pattern recognition and data mining. Aiming at minimizing cluster performance index, and error criterion ,square-error are foundations of this algorithm. It is not suitable with clusters in the original data which have different size and density.The applications include public transport data analysis and clustering of IT alerts.

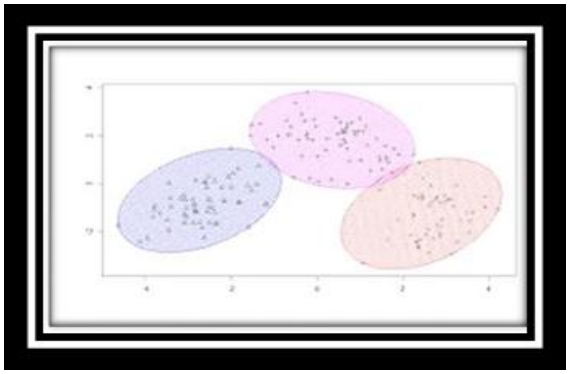

Fig 2. Impact of K means clustering

8. HIERARCHICAL CLUSTERING:-

It is very difficult to determine the correct number of clusters by the dendrogram for large datasets. The objective function is not directly minimized. Such hierarchical algorithms may be efficiently broken down into two groups of methods algorithm. The main application are social network data and US Senator Clustering through Twitter Charting Evolution through Phylogenetic Trees hierarchical algorithms may be efficiently broken down into two groups of methods algorithm. The main application are social network data and US Senator Clustering through Twitter Charting Evolution through Phylogenetic Trees.

9. AGGLOMERATIVE CLUSTERING:-

Various ranges of Agglomerative hierarchical clustering algorithms have been proposed at one time. The specificity of hierarchical methods is that they build a whole hierarchy of potential clusterings of the observations and can proceed in two ways. They can go bottom-up, i.e., start on the lowest hierarchy level and join clusters together towards the top; this approach is referred to as agglomerative clustering.

10. GAUSSIAN MIXTURE MODEL:-

This algorithm is very flexible and it is implemented in the Probabilistic
Method for acquiring a fuzzy classification. The main applications are modeling human height data and object tracking of multiple objects.

11. SPECTRAL CLUSTERING

It is elegant and fast for large datasets. It works very well when relations are approximately transitive. The application includes Pattern recognition and Speech processing. It is very Expensive for large datasets.

12. MEAN SHIFT

This algorithm doesn't have any Assumptions on the number of data clusters and the shape. The Output of the algorithm depends on the size of the window. The output obtained from the algorithm is independent of initializations.

Data Mining Technique	**Applications**	**Advantages**	**Limitations**
NAIVE BAYES ALGORITHM	Recommendation system. Sentiment Analysis. Medical data classification. Credit Scoring.	Suitable to work on large datasets and very easy to implement. It is a more efficient algorithm when it holds the independence assumption.	Zero frequency means the categorical value is not seen in training data then it assumes a zero probability to that model. Scarcity of the data.

IARJSET

ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 8, August 2021

DOI: 10.17148/IARJSET.2021.8874

RANDOM FOREST ALGORITHM	In the Banking sector to identify the loan risk. Medicine field diseases. Trends and risks of the disease can be identified.	Has high accuracy and flexibility, less variance. Works on easily with missing values. Regression and classification models can be implemented by using this algorithm.	In handling real time scenarios, random forest is slow. Due to the increase of decision tree it is ineffective. It consumes more time. Requires more resources for computation.

DECISION TREE ALGORITHM	Churn Analysis. Sentiment Analysis. Selecting a flight to travel. Handling late night cravings.	Easy to understand and interpret. Numerical and categorical data can be handled. Robust. Efficient algorithm for large datasets.	Complex calculations occur. Overfitting problems occur.
NEURAL NETWORK ALGORITHM	Financial Forecasting. Image processing. Language processing and translation.	It has a Good learning ability. It has a good speed. The output is not restricted to the input provided to them due to the good learning ability	Handles only numeric data. Need to translate each data into numeric form. Local optimates may occur.
SUPPORT VECTOR MACHINE ALGORITHM	Bioinformatics. Protein fold and remote homology detection. Face detection.	It Produces very accurate classifiers. It has Less overfitting and can handle noise data. It is highly Memory intensive.	It has the presence of discrete data. It has high algorithmic complexity.
LOGISTIC REGRESSION	Sentiment Analysis Object detection Disease prediction	It is very fast in classifying unknown records. It is very efficient method when it has linearly separable features in the datasets	Algorithms are very Sensitive to outliers. Linear Boundaries Construction takes place
K MEANS CLUSTERING	Identifying crime- prone areas. Fraud detection.	It works effectively for large datasets. Never required any prior distributional assumption	Very difficult to predict the K-Value. Not suitable With global well. Results in different final clusters. Due to the different initial partitions.
FuzzyC - Means Clustering Algorithm	Image segmentation. Computer forensics. Bioinformatics. Marketing.	Effective algorithm for overlapping. Data points may connect to more than one cluster center. Converges are the best feature. Effective algorithm for Unsupervised learning.	Specification of the number of clusters. It takes a long time. Computational time. Very Sensitivity to the initial guess speed, local minima,and noise.
HIERARCHICAL CLUSTERING	Tracking Viruses through Phylogenetic Trees.	Does not require an earlier specification on the number of clusters.	Never undo any previous steps. High time complexity due to the very long computation times.

IARJSET

ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 8, August 2021

DOI: 10.17148/IARJSET.2021.8874

GAUSSIAN MIXTURE MODEL	Speech recognition systems. Signal and information processing.	It gives highly applicable results for the real world datasets.	It results in high complexity due to the algorithms. Impotence to retrieve from database corruption.
MEAN SHIFT	Image Segmentation. computer vision. Video tracking.	Has better performance compared with k means clustering. It has only one parameter bandwidth which is the effective procedure.	Very slow when compared with k means clustering. Very Expensive for large features. Depends upon the parameter bandwidth.
SPECTRAL CLUSTERING	Speech separation. Hand drawings.	Can produce similarity /affinity matrix from the original dataset. It is an effective method for large datasets.	It is not suitable. For noisy and sparse data. It works effectively for convex and dense clusters.

13. RESULTS AND DISCUSSION

In this paper we used an analysis to discover the accuracy of the classification algorithm by using covid 19 india datasets from the kaggle and to identify the accuracy of the corona viruses. This analysis was carried out using google colab by applying 10 variables such as state ,cured, confirmed, deaths, confirmed Indian national, confirmed foreign national, age, gender, Date of observation, Time of observation Twelve classification and clustering algorithms were used Decision tree, naïve bayes, Random Forest algorithm, Neural Networks, Logistic regression. Support Vector Machine, K means clustering, Hierarchical clustering, Agglomerative Clustering, Mean Shift, Spectral Clustering, Gaussian

14. CONCLUSION

The performances of the algorithms were listed according to the advantages, limitations, definitions and application for the twelve classification and clustering algorithms used: Decision tree, naïve bayes, RandomForestalgorithm, Neural Networks, Logistic regression. Support Vector Machine, K Means Clustering, Hierarchical Clustering, Agglomerative Clustering, Meanshift, Spectral Clustering, Gaussian mixture model.

REFERENCES

[1] Yurong Zhong,Research Institute of Electronic Science and Technology, University oof Electronic Science Technology of China, Chengdu, China
[2] Himani Sharma1, Sunil Kumar21M.Tech Student, Department of computer Science, SRM University, Chennai, 2Assistant Professor, Department of computer Science, SRM University, Chennai, India
[3] Pouria Kaviani1, Mrs. Sunita Dhotre2 1M.Tech student, Department of Computer Engineering, Bharati Vidyapeeth University, College of Engineering, Pune
[4] Jehad Ali1, Rehanullah Khan2, Nasir Ahmad3, Imran
[5] Computer Systems Engineering, UET Peshawar, Pakistan 2, Sarhad University of Science and Information Technology, Peshawar, Pakistan 3 Computer Systems Engineering, UET Peshawar, Pakistan
[6] T. Haifley,Integrated Reliability Workshop Final Report, 2002, IEEE International
[7] Fionn Murtagh (1, 2) and Pedro Contreras (2) (1) Science Foundation Ireland, Wilton Place, Dublin 2, Ireland (2) Department of Computer Science, Royal Holloway, University of London.
[8] Youguo Li, Haiyan Wu Department of Computer ScienceXinyang Agriculture College Xinyang, Henan 464000, China.
[9] Nusa Erman, Ales Korosec, Jana Suklan, Faculty_of_Information_Studies_in_ china.