ISSN (e): 2250 – 3005 || Volume, 08 || Issue, 2|| February – 2018 ||
International Journal of Computational Engineering Research (IJCER)

Review on Data Mining Techniques

[1]Mrs.N.Sivanagamani Assoc.Prof
Dept. Of Computer Science and Engineering Geethanjali Institute Of Science & Technology
Approved by AICTE, New Delhi & Affiliated to JNTU, Anantapur)
Corresponding author: Mrs.N.Sivanagamani Assoc.Prof

ABSTRACT
Nowadays there is tremendous measure of information being gathered and put away in databases wherever over the globe .The propensity is to continue expanding a seemingly endless amount of time. It isn't intangible databases with Terabytes of information in undertakings and research agencies. Classification is a data mining technique used to predict group membership for data instances. In this paper, we present the basic classification techniques. Several major kinds of classification method including decision tree induction, Bayesian networks, k-nearest neighbor classifier, and support vector machines, genetic algorithm and fuzzy logic techniques. The goal of this review is to provide a complete review of different classification techniques in data mining.
Keywords: Bayesian, classification technique, data model, fuzzy logic, k-nearest neighbor classifier, Neural Networks, support vector machines,

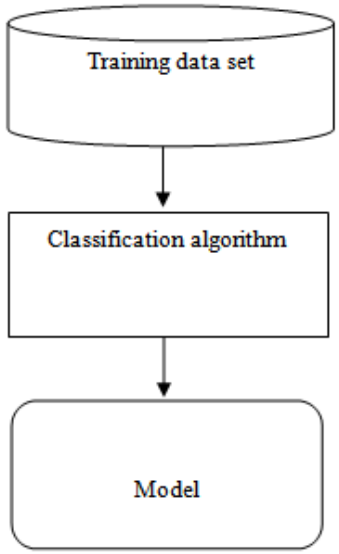
---

Date of Submission: 08-03-2018 Date of acceptance: 24-03-2018

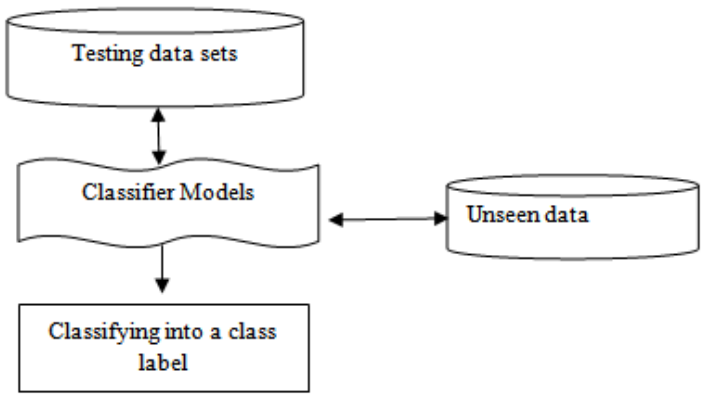
I. INTRODUCTION:

Introduction
Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. Classification used two steps in the first step a model is constructed based on some training data set, in seconds step the model is used to classify an unknown tuple into a class label.

1.1 Step 1 - Construction of a model



1.2 Step 2 - Model used for unknown tuple:



II. QUALITIES OF CLASSIFIERS:

Every last classifier has some quality which differential the classifier frame other. The properties are known as attributes of the classifiers. These qualities are

Correctness: - How a classifier orders tuple precisely depends on these attributes. To check exactness there are some numerical esteems in light of number of tuple characterize accurately and number of tuple order off-base.

Time: - How much time is required to build the model? This additionally incorporates an opportunity to use by the model to order at that point number of tuple (estimate time). In other word this refers to the computational costs.

Strength: - capacity to group a tuple accurately even tuple has a commotion. Noise can be wrong value or missing value.

Data Size: - Classifiers ought to be free frame the extent of the database. Model must to be multipurpose. The execution of the model isn't subject to the extent of the database.

Extendibility: - Some new element can be included at whatever point required. This element is hard to actualize.

III. DIFFERENT CLASSIFICATION MODEL:

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. There are several model techniques are used for classification some of them

1. Decision Tree
2. K-Nearest Neighbor
3. Support Vector Machines
4. Naive Bayesian Classifiers
5. Neural Networks

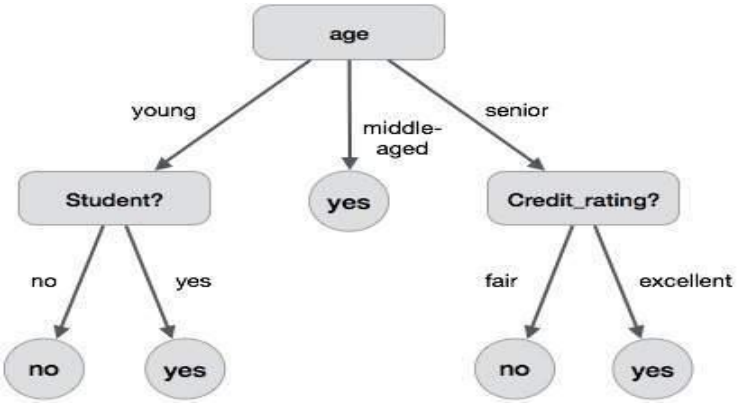
3.1Decision Tree:

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 denotes the result of a test, and each leaf node holds 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. Each leaf node represents a class.

The benefits of having a decision tree are as follows

1. It does not require any domain knowledge.
2. It is easy to comprehend.
3. The learning and classification steps of a decision tree are simple and fast.



Tree Pruning

Tree pruning is performed in order to remove anomalies in the training data due to noise or outliers. The pruned trees are smaller and less complex.

Tree Pruning Methods

There are two approaches to prune a tree

Pre-pruning − the tree is pruned by halting its construction early.

Post-pruning - This approach removes a sub-tree from a fully grown tree.

Cost Complexity
The cost complexity is measured by the following two parameters −
Number of leaves in the tree, and
Error rate of the tree.

3.2 K-Nearest Neighbor: This classifiers are based on learning by training samples. Each sample represents a point in an n-dimensional space. All training samples are stored in an n-dimensional pattern space. When given an unknown sample, a k-nearest neighbor classifier searches the pattern space for the k training samples that are closest to the unknown sample. "Closeness" is defined in terms of Euclidean distance, where the Euclidean distance, where the Euclidean distance between two points, X=(x1,x2,……,xn) and Y=(y1,y2,….,yn) is denoted by d(X, Y).

$$d(X,Y) = \sqrt{\sum_{i=1}^{n} (xi - yi)^2}$$

3.3 Support Vector Machines:support vector machines networks are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier . An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.
In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

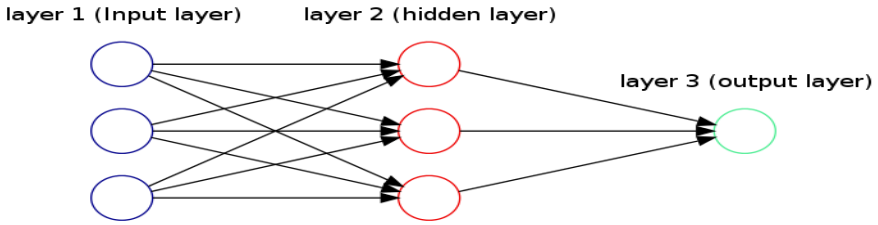
3.4 Naive Bayesian Classifiers:
Bayesian classifiers are statistical classifiers. They can predict class membership based on probabilities. The Naive Bayes Classifier technique is particularly suited when the dimensionality of the inputs is high. Naive Bayes can often outperform more sophisticated classification methods. Let D be a training set associated class labels. Each tuple is represented by an n-dimensional attributes, A1, A2,.., An. . Suppose that there are m classes, C1, C2,…, Cm. Given a tuple, X, the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X. That is, the naïve Bayesian classifier predicts that tuple x belongs to the class Ci if and only if P (Ci / X) > P (Cj/X) for 1<= j <= m, j ≠ i. Thus we maximize P(Ci / X). The class Ci for which P(Ci / X) is maximized is called the maximum posteriori hypothesis. By Bayes' theorem

$$P(C_i/X) = \frac{P\left(\frac{X}{Ci}\right)(P(Ci)}{P(X)}$$

P(X) is constant for all classes, only P (X/Ci) P (Ci) need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is, P(C1) = P(C2) =……=P(Cm), and we would therefore maximize P(X/Ci). Otherwise, we maximize P(X/Ci)P(Ci).

3.4.1 Applications of Naive Bayes Algorithms
Real time Prediction: Naive Bayes is an enthusiastic learning classifier and it is certain quick. In this way, it could be utilized for making forecasts in real period.
Multi class Prediction: This calculation is additionally notable for multi class forecast highlight. Here we can anticipate the likelihood of different classes of target variable.
Text classification/ Spam Filtering/ Sentiment Analysis: NaiveBayes classifiers generally utilized as a part of content characterization (because of better outcome in multi class issues and autonomy govern) have higher achievement rate when contrasted with different calculations. Accordingly, it is broadly utilized as a part of Spam sifting and Sentiment Analysis.**Recommendation System:** Naive Bayes Classifier and Collaborative Filtering together forms a Recommendation System that utilizations machine learning and information mining methods to channel inconspicuous data and foresee whether a client might want a given asset or not.
3.5 Neural Networks: An Artificial Neural Network, often just called a neural network, is a scientific model inspired by biological neural networks. A neural network consists of an interrelated group of artificial neurons, and it processes information using a connectionist approach to computation. In most cases a neural network is

an adaptive system that changes its structure during a learning phase. Neural networks are used to model complex relationships between inputs and outputs or to find patterns in data.
Neural networks are also similar to biological neural networks in that functions are performed collectively and in parallel by the units, rather than there being a clear definition of subtasks to which various units are assigned. The term "neural network" usually refers to models employed in statistics, intellectual psychology and artificial intelligence. Neural network models which match the central nervous system are part of theoretical neuroscience and computational neuroscience.



3.5.1Real-life applications:
Capacity estimation, or relapse investigation, including time arrangement forecast, wellness guess and displaying. Order, including example and grouping acknowledgment, oddity location and successive basic leadership. Information handling, including sifting, grouping, dazzle source partition.

IV. CONCLUSION

There are a few grouping methods in in data mining and each and every technique has its advantage and disadvantage. Decision tree classifiers, Bayesian classifiers, classification by back propagation, support vector machines, these techniques are eager learners they use training tuples to construct a simplification model.
Some of than are lazy learner like nearest-neighbor classifiers and case-based reasoning. These store training tuples in pattern space and wait until presented with a test tuple before performing generalization.
Data mining has significance in regards to finding the examples, determining, revelation of learning and so forth.In various business spaces. Data mining methods and calculations, for example, arrangement, bunching and so forth.Helps in finding the examples to settle on the future patterns in organizations to develop. Data mining has Wide application area nearly in each industry where the information is created that is the reason data mining is Thought about a standout amongst the most critical outskirts in database and data frameworks and a standout amongst the most promising interdisciplinary improvements in Information Technology.

REFERENCES

[1]. M. Akhil jabbar & Dr. Priti Chandrab "Heart Disease Prediction System using Associative Classification and Genetic Algorithm" International Conference on Emerging Trends in Electrical, Electronics and Communication Technologies-ICECIT, 2012.
[2]. M. Akhil Jabbar, B.L Deekshatulu & Priti Chandra "Classification of Heart Disease using Artificial Neural Network and Feature Subset Selection" Global Journal of Computer Science and Technology Neural & Artificial Intelligence Volume 13 Issue 3 Version 1.0 Year 2013 International Research Journal Publisher: Global Journals Inc. (USA)
[3]. S. Olalekan Akinola, O. Jephthar Oyabugbe Accuracies and Training Times of Data Mining Classification Algorithms: An Empirical Comparative Study" Journal of Software Engineering and Applications, 2015, 8, 470-477 Published Online September 2015 in SciRes. http://www.scirp.org/journal/jsea
[4]. Jaimini Majali, Rishikesh & Niranjan, Vinamra Phatak "Data Mining Techniques For Diagnosis And Prognosis Of Cancer" International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 3, March 2015
[5]. Nikhil N. Salvithal " Appraisal Management System using Data mining "International Journal of Computer Applications (0975 – 8887) Volume 135 – No.12, February 2016
[6]. Tanvi Sharma, Anand Sharma & Vibhakar Mansotra"Performance Analysis of Data Mining Classification Techniques on Public Health Care Data" International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 4, Issue 6, June 2016
[7]. B RosalineJetta "EFFICIENT CLASSIFICATION METHOD FOR LARGE DATASET BY ASSIGNING THE KEY VALUE IN CLUSTERING" International Journal of Computer Science and Mobile Computing A Monthly Journal of Computer
[8]. Science and Information Technology ISSN 2320–088X IJCSMC, Vol. 3, Issue. 1, January 2014, pg.319 – 324
[9]. DivaTamar and Sonali Agarwal " A survey on Data Mining approaches for Healthcare" international Journal of Bio-Science and Bio-Technology Vol.5, No.5 (2013), pp. 241-266 http://dx.doi.org/10.14257/ijbsbt.2013.5.5.25
[10]. V.Krishnaiah , Dr.G.Narsimha, Dr.N.Subhash Chandra "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (1) , 2013, 39 - 45
[11]. N S Nithyaand K Duraiswamy "Gain ratio based fuzzy weighted association rule mining classifier for medical diagnostic interface" Sadhana¯ Vol. 39, Part 1, February 2014, pp. 39–52. Indian Academy of Sciences
[12]. 1. Jiawei Han and Micheline Kamber (2006), Data Mining Concepts and Techniques, published by Morgan Kauffman, 2nd ed.

Mrs.N.Sivanagamani "Review on Data Mining Techniques." International Journal of Computational Engineering Research (IJCER), vol. 08, no. 02, 2018, pp. 38-41.