International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES)
Impact Factor: 5.22 (SJIF-2017), e-ISSN: 2455-2585
Volume 4, Issue 12, December-2018

AN EMPIRICAL STUDY OF DIFFERENT DATA MINING TECHNIQUES AND ALGORITHMS

Jismy Joseph[1], Dr.G. Kesavaraj[2]

PhD Research Scholar[1], Professor and head[2]
Department of Computer Science, Vivekanandha College of Arts and Science for Women (Autonomous), Elayampalayam, Thiruchengode, Tamil Nadu, India

Abstract:- Data mining is a process to find hidden information from a large data set. Data mining technologies are used nowadays in various sectors like banking, healthcare, education, marketing etc. One of the most important challenges in data mining is to choose the correct data mining technique and algorithm based on the type of problems tackled by businesses. A generalized approach can improve the accuracy and cost effectiveness of data mining process. The aim of this article is to give a comprehensive review of the most frequently considered techniques and algorithms for data mining.

Keywords - Data mining Techniques, Classification, prediction, Clustering, k-NN, Naïve Bayes classifier, Decision Tree, C4.5, classification.

I.INTRODUCTION

Nowadays millions of terabytes of data are there in the database and organizations need to analyse these data for performing a business task. Data mining techniques and algorithms are used for this analytical task, but selecting the best one for a particular situation is a big challenge because each algorithm produces a different result. Data mining means extraction of knowledge or hidden information from huge amount of data. It is also knows as knowledge discovery from Data (KDD). The knowledge recovery process consists data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge representation [1]. Data mining can be applicable to different data repositories like data warehouse, transactional database, data stream, object-relational database, text database, time series database, multimedia database etc.

II.DATA MINING TECHNIQUES

There are many data mining techniques, but most commonly used techniques are classification, clustering, perdition, association, decision tree, sequential pattern and regression analysis [2].

1. Classification

Classification is the most widely used technique in Data mining. The main goal of the classification process is to create a model that describes and distinguishes data classes. Two steps are used in classification process. In the first step, the classifier algorithm builds a classifier from a training data set made up of database tuples and their associated class labels. After that the classifier is used to classify each item. Different forms are there to present this model. They are decision tree classification rule based classification, neural network, mathematical formulae etc. The decision tree is a tree structure format, its internal node represents the test on attributes and external node represents the result of the test. Classification rule consists of a set of IF-THEN rule for classification. IF part consists of the condition and THEN part consists of the conclusion. Neural networks are mainly used when the relationship between the inputs and outputs are complex.

2. Clustering

Clustering is used to identify similar objects and create a cluster for these similar objects. In clustering the class labels are not known in advance. The main objective of clustering is to maximize the similarities between the objects within the cluster and minimize the similarities between the objects in different cluster. Different methods are available for clustering. They are partitioning, hierarchical, density-based, grid based and model based methods. In partitioning method 'n' data tuples are classified into 'k' groups. Each group should have at least one data tuples and each tuples must belong to exactly one group. A hierarchical method, decomposes the data into hierarchical manner. This method uses either agglomerative (bottom-up) or divisive (top-down) approach for hierarchical decomposition. The density method uses the notion of density to classify the group. In grid method the object space is quantized into finite number of cells that form a grid structure. Constraint based method uses an application oriented or user defined constraints for performing cluster.

3. Prediction

Prediction is a supervised learning technique used to predict missing or unavailable data values. Regression analysis is the commonly used techniques for numeric prediction.

4. Association

In some situation, we need to identify the relationships between different data in a data set. In such a scenario, association is the best technique for finding hidden pattern .This data mining technique is used to find the frequent item set and create strong rules from the frequent data set. There are three types of association rule. They are multilevel, multidimensional and quantitative association rule. Multilevel rule involves data at different levels while multidimensional uses more than one dimensional data. Quantitative association rule uses attribute that contains numeric data.

III. DATA MINING ALGORITHMS

Data mining algorithms contain a set of calculations, which are used to create a model from large set of data [3]. The algorithms first analyse the input data and then generate a model. Large number of data mining algorithms are available today which are used in field of health care, financial sector, sales, engineering corporate business etc. This paper focuses mainly on commonly used algorithms such as:

- K-NN
- Decision Tree
- Random Forest
- Naïve Bayesian Classification

1. KNN

Knn is a simple algorithm for classification. It is used for both classification and regression but mainly used for classification. In this method each tuple represents a point in an n-dimensional array. This algorithm compares the similarities between the training tuples and given test tuples.

When there is an unknown tuple, Knn algorithms compares this tuples with training tuples and finds 'k' training tuples that are closest to unknown tuple. These 'k' nearest tuples are called k nearest neighbours of unknown tuples. Euclidean distance is used to find the distance between these tuples. If two tuples are X1(x11, x12……..x1n) and X2(x21, x22……..x2n) then the distance between these tuples are calculated by using the following formula

$$\text{Dist}(X1, X2) = \sqrt{(\sum_{i-1}^{n}(x1i - x2i)2} \quad [1]$$

If distance between these tuples are less then there is a similarities between these tuples. Figure-1 is an example for K-NN classification.

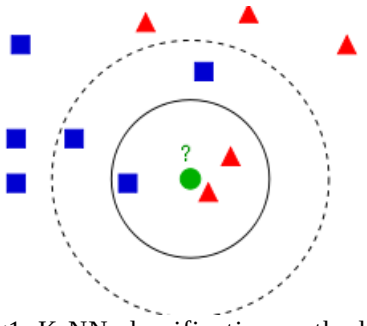

Fig1. K-NN classification method.

In the above figure, K-NN needs to classify a new data point with green circle into "Blue" or "Red" class. The KNN algorithm first calculates its nearest points. In this example the green circle has 3 neighbour points, two red and one Blue (K=3). The classifier then assign the new object to class which has higher neighbour points. From the figure above the new data point will be classified as "Red" because most of the nearest points are belongs to "Red" class.

Advantages of Knn Algorithm

- Simple technique that is easily implemented and building model is also easy.
- No cost for learning process.
- It is well suited for Multi-modal classes and Records with multiple class labels.

Disadvantages of Knn Algorithm

- It is a lazy learner because it simply uses the training data itself for classification and nothing is learnt from training data set.
- High memory requirement. It Stores all of the training data.
- When dataset is large it is expensive to find 'k' nearest neighbours.
- Accuracy may be reduced by the presence of irrelevant and noisy data.
- The performance depends on the number of dimensions.

2. Decision Tree

ID3, C4.5 and CART are different decision tree algorithms that uses a greedy/nonbacktracking method. In these algorithms, the decision tree is constructed in top-down recursive divide and conquer manner. The decision tree starts with a training data set called data partition and their associated class labels. The other inputs are a set of candidate attribute and an attribute selection method. The attribute selection method specifies the procedure to split the data tuples into individual classes. Initially all data tuples are in the same class and the node of the tree is considered as leaf node. After that training data set is divided into smaller subsets with respect to the attribute selection procedure and a tree is constructed. The partition stops when all tuples belong to the same class or there are no remaining attributes on which the tuples may be further partitioned.

Figure 2 shows how decision tree classifier works. The decision tree classifiers contains a set of test questions and conditions in a tree structure. The root node and all internal nodes contains test questions and the terminal node contains the class labels.

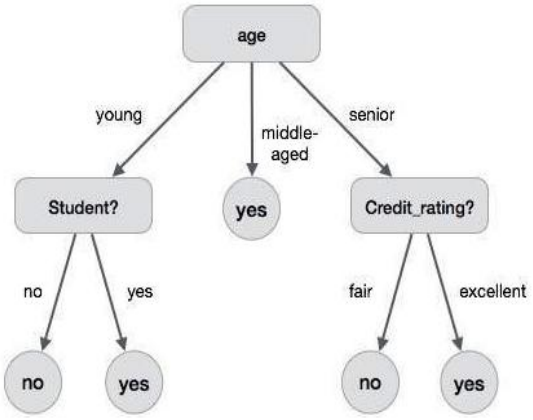

Fig2.Decision tree classification method.

Advantages of Decision tree

- For data preparation, decision tree needs little effort from users.
- Variable screening or feature selection is performed implicitly.
- Nonlinear relationships between parameters don't have any effect on the performance of tree.
- It can handle the dataset that may have errors and missing values.
- Decision tree is capable of handling nominal and numerical data.

Disadvantages decision tree

- If many complex interactions between the attributes are present, then the "divide and conquer" method used by decision tree for splitting will not perform well.
- Decision trees can be used in conjunction with other project management tools. For example, the project schedule can be evaluated using decision tree method. [4]
- The decision algorithms like ID3 and C4.5 requires that the target attribute will have only discrete values.
- Another disadvantage of decision tree is its greedy characteristic. This is its over-sensitivity to the training set, irrelevant attributes and to noise data [5].

3. Random Forest

Random forest is a machine learning and very flexible algorithm. It is a supervised machine learning algorithm used for classification and regression. These classifiers handle the missing values and can model the categorical values**.** It creates many decision trees and merges them together to form an accurate prediction. In this method the parameters are used to increase the predictive power and speed of the model. The basic building block of the random forest is decision tree. The random forest starts with 'k' randomly selected features from 'n' total features, then the splitting procedure is used to find the root node of the tree. After finding the root node, the splitting procedure is again applied to find the daughters' node. This procedure is repeated until 'm' number of nodes are created.

Advantages Random Forest

- Random forest is suitable for both classification and regression.
- Random forest is capable for handling missing values and categorical values.
- Random forest classifier doesn't overfit the model when it contains more number of decision trees.
- Random forest can identify the most important feature out of the available features from the training dataset.
- Capable for handling large data set.
- For finding interaction between the variables, it offers an experimental method.

Disadvantages Random Forest

- Random forest is not easily interpretable [6].
- Random forests have been observed to overfit some datasets with noisy classification [7].

4. Naïve Bayesian Classification

Naïve Bayes is a powerful and straightforward algorithm for classification .This approach can work on data sets that has millions of records [8].

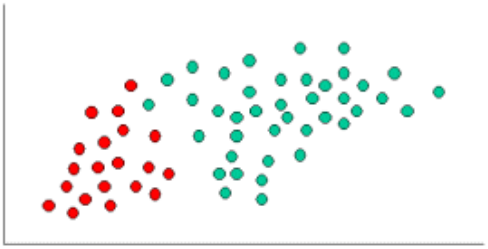

Fig.3. Naïve Bayesian Classification method.

In the above figure objects are classified into two groups, RED and GREEN. When a new object arrives, the classifier needs to decide to which class label they belongs to.

It is a supervised learning method based on conditional probability and also using independent assumption. Navie Bayes classifier is a probabilistic classifier, it uses probability distribution to classify the given input over a set of classes. Naïve Bayes classifier uses the following equation.

$$P(C \mid F_1.....F_n) = \frac{P(F_1.....F_n \mid C) * P(C)}{P(F_1.....F_n)}$$

[1]

The probability of an event can be calculated by using the conditional probability. The following formula is used for calculating the conditional probability.

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

[1]

In the above equation P (A|B) is the probability that the hypothesis A holds observed data tuples B. P (B|A) is the posterior probability of B conditioned on A. P (B) is the prior probability of B and P (A) is the priori probability of A.

This classifier assumes that the features are independent. This algorithm is used for

- Real time Prediction
- Text classification/ Spam Filtering
- Recommendation System

Advantages Naïve Bayesian Classification

- Naïve Bayesian algorithm is easy to implement.
- Its efficiency and classification performance are high.
- Compared to other sophisticated algorithms the amount of data required is this is very less.
- In most of the prediction and classification cases it produces accurate result.

Disadvantages Naïve Bayesian Classification

- Computing the probabilities by the traditional method of frequency count is not possible when an attribute is continuous.
- To implement it, one needs to find many conditional probabilities.

IV.CONCLUSION

The performance of an algorithm is based on the types of the problem. In some areas like credit risk analysis random forest may perform better but in some others like healthcare or education, other algorithms may perform better. So algorithm performance changes from domain to domain. In this paper the algorithm K-NN, decision tree, random forest and Naive Bayes are discussed. K-NN is a simple technique and it is easy to implement but when dataset is large it is expensive. Accuracy may be reduced by the presence of irrelevant and noisy data. Whereas the decision tree can handle the dataset that may have errors and missing values.

5. REFERENCES

1. Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, 2003.
2. Alton, L. (2017, December 22). Retrieved from www.datasciencecentral.com: https://www.datasciencecentral.com/profiles/blogs/the-7-most-important-data-mining-techniques

3. Microsoft. Data Mining Algorithms .Analysi Services-Data Mining. 2016.
4. Rafael Olivas, „Decision Trees – A primer for Decision-making Professionals", 2007.
5. Oded Maimon, Lior Rokach, „Data Mining and Knowledge Discovery Handbook", Second Edition, Springer Science+Business Media, p. 149-174.
6. Bahnsen, D. A. (2018). *machine-learning-algorithms-introduction-random-forests*. Retrieved from www.dataversity.net: http://www.dataversity.net/machine-learning-algorithms-introduction-random-forests/
7. Segal, Mark R. (April 14 2004). *Machine Learning Benchmarks and Random Forest Regression*. Center for Bioinformatics & Molecular Biostatistics.
8. Shrey Bavisi, Jash Mehta, Lynette Lopes. (2014). A Comparative Study of Different Data Mining Algorithms. *International Journal of Current Engineering and Technology* , 3248-3252.