A Study on Advantages of Data Mining Classification Techniques

0. Yamini,
Reasearch Scholar
Dept. of Computer Science
S.V.University
Tirupati, Andhra Pradesh

Prof. S. Ramakrishna
Dept. of Computer Science
S.V.University
Tirupati, Andhra Pradesh

***Abstract*— The data mining has a basic principle for analyzing the data from different angles and categorize it and finally to condense it. In today's world data mining have increasingly become very interesting and popular in terms of all application. Data and information have become major virtue for most of the organizations or institutions. The success of any organization confide in imposingly on the orbit to which the data acquired from business operations is utilized. Classification is an important task in knowledge discovery in databases (KDD) process. This paper provides different classification techniques analogous as Decision tree Induction, Bayesian Classification, Neural networks, Support Vector Machines.**

Keywords—Data Mining, Classification, Decision tree induction,Neural networks.

I. INTRODUCTION

There are many different methods used to perform the data mining task. These techniques not only required specific type of data structure but also betoken certain type of algorithm approach.

Data mining is considered to be an emerging technology that has made rioting change in the information world. The term data mining (often called as knowledge discovery) refers to the process of rehash data from different perspectives and summarizing it into useful information by means of a number of analytical tools and techniques, which in a proper may be useful to increase the performance of a system.

Data mining is an active area of research and the research is going-on to bring statically analysis and AI techniques to gather to address the issues [1]. Data Mining can also be defined as a multidisciplinary field which combines artificial Intelligence statistics and database technology. The relationship amid data mining, data base and AI & machine learning is shown in figure 1.

Many dodge like Supermarkets, hotels, factories have stored large amounts of data over years of operations, and data mining can concentrate much admired knowledge from this data. Then these business people can make greater profits by induce more customers and by convalescent sales. This is possible with engineering and medical fields [2].

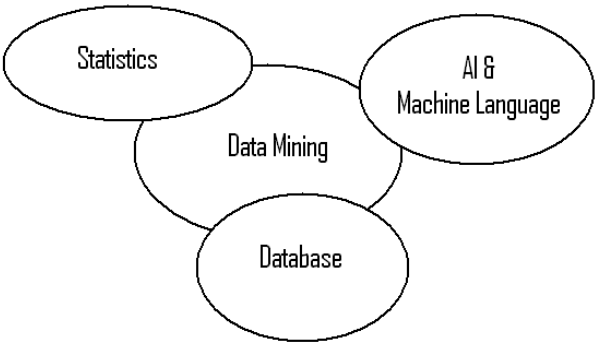

Figure 1. Relations ship in dispersion Data mining, AI with database

Commonly data mining contains several algorithms and techniques for picking out entrusting patterns from large data sets. Data mining techniques are restricted into two leagues: supervised learning and unsupervised learning. In supervised learning, a model is built antecedent to the analysis. We then exploit the algorithm to the data in order to estimate the framework of the model. Classification, Decision Tree, Bayesian Classification, Neural Networks, Association Rule Mining etc. are habitual examples of supervised learning. In unsupervised learning, we do not beget a model or hypothesis prior to the analysis [3]. We just utilize the algorithm precisely to the dataset and observe the results. Then a model can be erect on the ground work of the obtained results. Clustering is one of the exemplar of unsupervised learning. Various data mining tactics such as Classification, Decision Tree, Bayesian Classification, Neural Networks, Clustering, Association Rule Mining, Prediction, Time Series Analysis, Sequential Pattern and Genetic Algorithm and Nearest Neighbor have been used for knowledge discovery from large data sets [4]. Some of the common and useful data mining techniques have been discussed.

II. CLASSIFICATION

Classification is a well known data mining unsupervised learning technique, that which employs a set of pre-classified examples to mellow a model that can easily classify the population of records and some of the applications of data mining like Fraud detection and credit risk are particularly well suited to this type of analysis.

Classification approach frequently employs the decision tree or neural network-based classification algorithms. The data classification aciurgy that mainly tangle in the learning and in the classification, learning the training data is analyzed by the classification algorithms and in classification test data are used to estimate the accuracy of the classification rules. If the aciurgy is acceptable then the rules can be applied to the new data tuple. For a fraud detection application, this includes plenary records of both fraudulent and valid activities determined on a record-by-record. The classifier-training algorithm uses some of the pre-classified examples to determine the set of parameters that are required for proper intolerance. Then the algorithm conceals these parameters into a model called as a classifier [5].

Some of the well-known classification models are:

a) Classification by decision tree induction
b) Bayesian Classification
c) Neural Networks
d) Support Vector Machines (SVM)

A. Classification by decision tree induction

Decision tree is one of the most used data mining techniques because its model is easy to understand for all the users working on it. In this knack, the root of the decision tree is a simple question or otherwise called it as a posture that has multiple answers. Each answer then leads to a related set of questions or conditions that help us to persuade the data so that we can make the final decision based on it. [6]

For example, Let us see the following decision tree to ordain whether or not to play tennis:

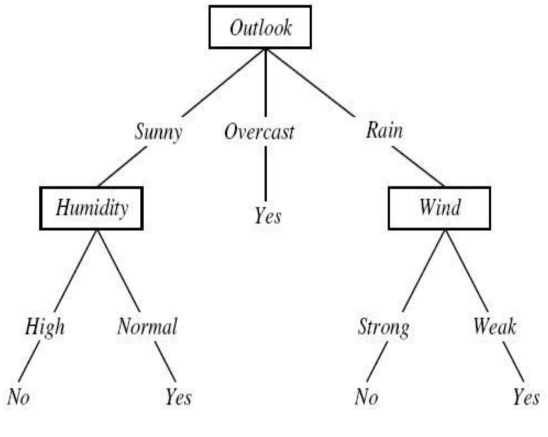

Fig. 1. decision tree

Starting at the root node, if the outlook is overcast then we can definitely play tennis. If it is rainy, we can only play tennis if the wind is week. If it is sunny then we can play tennis in case humidity is normal.

Parameters of Decision tree:

Data partition (Set of training tuples and their associated class labels), Attribute list, Attribute selection procedure (classify the attributes based on the associated classes).
Advantages:

- Decision trees don’t require realm knowledge.
- These are easy to understand.
- Handles High dimensional data.
- Classification and learning becomes simpler when decision trees are used.
- These are very accurate.

B. Bayesian Classification:

Bayesian classification is stationed on Bayes' Theorem. Bayesian classifiers are the powerful probabilistic representations that can predict class enrollment probabilities such as the probability that a given tuple belongs to a intrinsic class.

Bayes' Theorem is named after Thomas Bayes. These are classified into two types of probabilities −

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

Here X is a data tuple, where H is some hypothesis.
According to Bayes' Theorem,

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

Bayesian Belief Networks specify joint conditional probability distributions. They are otherwise known as Belief Networks. We can use a trained Bayesian Network for classification [7].Here we are having two constituents that define a Bayesian Belief Network −

- Directed acyclic graph
- A set of conditional probability tables

Directed Acyclic Graph Representation

The following diagram is a directed acyclic graph for six Boolean variables.

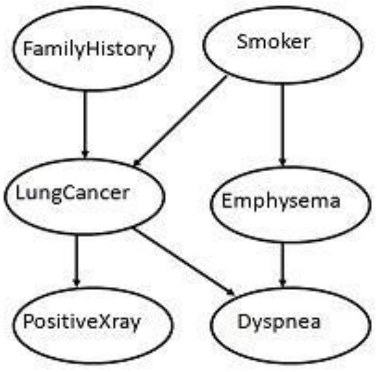

Fig. 2. Directed Acyclic Graph Representation

The arc in the picture allows representation of knowledge. For example lung cancer is affected by a person's family history of lung cancer, as well as in case or not the person is a smoker. It is worth noting that the variable Positive X-ray is independent of whether the patient has a family history of lung cancer or that the patient is a smoker, given that we know the patient has a disease as lung cancer

Advantages:

- Bayesian classifiers are efficient as efficient as decision trees and neural network classifiers.
- These are very accurate.
- Exhibit high speed.
- They exhibit class fortuitous independence (The attributes with in the class are independent of each other).
- They make the process of computation simple.

C. Neural Networks

One of the most in vouge NN algorithms is back propagation algorithm. In more practical terms neural networks are non-linear statistical data mold tools. They can be worn to model complex relationships between inputs and outputs or to find patterns in data [8]. Rojas [2005] claimed that Back Propagation algorithm could be broken down to four main steps. After exercising the weights of the network erratically, the back propagation algorithm is used to compute the quintessential corrections.

The algorithm can be degrading into the following four steps:

i) Feed-forward reckoning
ii) Back propagation to the output layer
iii) Back propagation to the hidden layer
iv) Weight updates

The algorithm is interrupted when the value of the howler function has become sufficiently small. This is very rough and basic formula for BP algorithm. There are some variations proposed by other scientist but Rojas definition seem to be quite accurate and easy to learn [9].

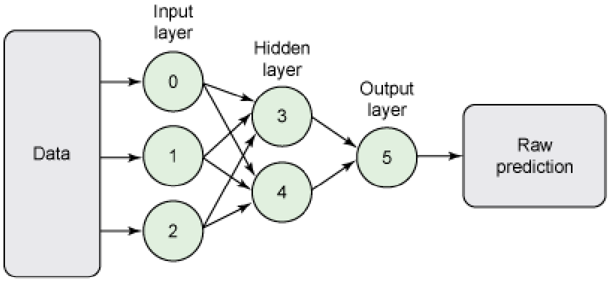

Fig. 3.A simple neural network model in which data is passed through a sequence of layers before a prediction is computed.

Back propagation algorithm has three steps:

1. Initializing the weights and biases.
2. Determine the input and output of each node.
3. Calculate the error at each node and adjust the weights and biases.

Advantages:

- High accuracy and tolerance of noisy data.
- Classifications of patterns.
- Requires knowledge about the relationship between tuples and classes.
- It is suitable for Continuous-valued data.
- It is very much suitable for real-world data.
- Uses parallelism to speed up calculation.
- Ease of Maintenance and can be implemented in parallel hardware

Disadvantages:

- It involves long training.
- Requires many parameters as topology or structure.
- Poor interpretability.

D. Support Vector Machine

SVM, a powerful machine Support Vector Machine (SVM) was first proposed by Vapnik and has since attracted a potency of interest in the machine learning research community [10] & [11]. Several recent studies have reported that the SVM (support vector machines) generally are proficient of delivering gassed-up in terms of classification accuracy than the other data classification algorithms. However, for some datasets, the exploit of SVM is very sensitive to how the cost parameter and kernel parameters are set. As a termination, the user habitually needs to conduct extensive cross validation in order to figure out the most favorable criterion ambience. This technique is generally raftered to as model selection. A special property of SVM is, SVM concurrently miniaturize the speculative classification error and maximize the geometric margin, So SVM called paramount Margin Classifiers. SVM is found on the Structural risk Minimization (SRM)

SVM can be used for both classification and prediction.

Advantages

- SVM uses maximum marginal hyper plane for classifying linearly separable data.
- Data can be clearly separated into portions.
- SVM extends itself in order to classify the linearly inseparable data.

III. CONCLUSION

Data mining is a process of extracting knowledge from massive data and makes use of different data mining techniques. Numbers of data mining techniques are discussed in this paper like Decision tree induction (DTI), Bayesian Classification, Neural Networks, Support Vector Machines. After my study on all the classification techniques it becomes more flexible to decide a technique for data mining. Mostly decision tree induction is most understandable when compared with the other techniques.

IV. FUTURE WORK

There are many future directions in data mining. As a part of future work, we supposed to do our research in different decision tree algorithms in data mining applications.

REFERENCES

[1] Dr. Pardeep Mittal, Sukhpreet Singh, Amritpal Singh, Priyanka, A Review of Data Mining Techniques with their Merits & Demerits, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 3, March 2014 ISSN: 2277 128X

[2] P. Meena Kumari et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 2458-2461

[3] Jiawei Han and Michelire Kamber, “Data Mining Concept and Technique”, Published by Morgan Kaufman, 2006.

[4] Monika Goyal and Rajan Vohra, “Application of Data Mining in Higher Education”, International Journal of Computer Science (IJCSI) Issues, Vol. 9, Issue-2, No.1, March 2012; pp-113-120.

[5] Tapas Ranjan Baitharu, Subhendu Kumar Pani, A Survey on Application of Machine Learning Algorithms on Data Mining, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-3, Issue-7, December 2013

[6] http://www.zentut.com/data-mining/data-mining-techniques

[7] http://www.tutorialspoint.com/data_mining/dm_bayesian_classification.html.

[8] Dr. Yashpal Singh, 2 Alok Singh Chauhan, Neural Networks in Data Mining, Journal Of Theoretical And Applied Information Technology, Journal of Theoretical and Applied Information Technology

[9] http://www.dataminingmasters.com/uploads/studentProjects/NeuralNetworks.pdf.

[10] Boser, B. E., I. Guyon, and V. Vapnik (1992). A training algorithm for optimal margin classifiers. In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, pages. 144 -152. ACM Press 1992.

[11] V. Vapnik. The Nature of Statistical Learning Theory. NY: Springer-Verlag. 1995.