A Survey on Data Mining Algorithms

D.Keerthika¹, G.Sangeetha²

PG Scholar, Department of CSE, Valliammai Engineering College, Kancheepuram, India¹
Assistant Professor, Department of CSE, Valliammai Engineering College, Kancheepuram, India²

Abstract: Data mining is multidisciplinary field of computer science. It is the process of recognizing patterns from massive data sets, which is big data. It contains the approaches of Machine Learning, database systems, artificial intelligence and statistics. A data mining algorithm is a set of problem solving and computation that generates a data mining miniature from data. Data mining contains large collection of algorithms. In this paper, some of the popular algorithms for data mining are closely examined with their advantages and disadvantages.

Keywords: Data mining, data mining algorithms, data sets.

I. INTRODUCTION

In today’s internet world, data present in internet is vast. Data mining is the process of evaluating the data in various contexts and summing up it into useful information—information that can be used to increase gain. It engages data pre-processing, complexity considerations, model inference, online updating and data management. Data mining is the scrutiny step of KDD. This process is normally illustrated with the following things:

- Selection
- Reprocessing
- Conversion
- Data Mining
- Evaluation [1].

Terms related to Data Mining

Data: Data are any text, numbers, facts and that can be processed by a computer. At present, organizations are acquiring broad and growing volumes of data in various formats and various databases [2].

Information: The relationships, associations, or patterns among all this data can render information. For example, analysis of trade point of sale transaction data can yield information on which products are trading and when [2].

Knowledge: Information can be changed into knowledge about classical patterns and ultimate trends. For example, summary information on retail jewellery sales can be examined in light of advertising efforts to render knowledge of customer purchasing behaviour. Thus, a manufacturer or vendor could determine which items are most susceptible to promotional efforts [2].

Data warehouse: Due to the development of data transmission, processing power, storage capabilities and data capture are permitting organizations in order to collaborate databases into data warehouse. It is the process of retrieval and centralized data management. Data warehouse serve as a unique vision of preserving essential repository of whole data of organization. Data Centralization is used to increase client access and evaluation. Evolution in data analysis software provides free access to data [2].

From fig.1, data are first selected from the data base. Then the target data is cleaned using pre-processing. Pre-processing is the method of changing the data into well-defined format. Next the transformed data is mined and converted into patterns. The last step in data mining is Evaluation, where the data is evaluated into knowledge. The Main steps in data mining are

1. Pre-processing
2. Transmutation
3. Data Mining
4. Assessment.

Fig.1 Data mining process

II. CLASSIFICATION OF DATA MINING ALGORITHMS

Fig.2 Classification of Data Mining Algorithms
Classification algorithm is the well-knownly used data mining algorithms, which contains a set of reclassified examples to develop a miniature that can categorize the records. It finds its applications in fraud detection and credit-risk [3]. This method generally contains support vector machine algorithm, neural networks. It involves classification and learning. The training data are analysed in learning. In classification, it checks the accuracy of test data. The rules can be applied to fresh tuples, if accuracy is adequate. This can be divided into

- Neural network
- Classification based on association
- Bayesian classification
- Support Vector Machine
- Decision Tree

Classification algorithm is an algorithm used to develop models that can categorize records. It finds its applications in fraud detection and credit-risk [3]. This method generally contains support vector machine algorithm, neural networks. It involves classification and learning. The training data are analyzed in learning. In classification, it checks the accuracy of test data. The rules can be applied to fresh tuples, if accuracy is adequate. This can be divided into:

- Neural network
- Classification based on association
- Bayesian classification
- Support Vector Machine
- Decision Tree

Advantages
1. It yields an entangled relationship between input and output.
2. It is widely useful in creation of prototype and clustering.
3. It is able to evaluate and construct data using inmost features beyond outermost guidance.

Disadvantages
1. It does not work well, when the input is increased.
2. It does not provide adequate performance for complicated problems.
3. It is strenuous to recognize the neural network model and how the basic data affects output guessing response.

Advantages
1. It provides the most accurate rules for classification.
2. It provides performance better than other algorithms.

Disadvantages
1. Selecting a rule simple will affect the accuracy.
2. For a large number of rules, training sets and long pattern rules, the efficiency will be low [7].

a. Classification based on Association

Classification is one of the types of data mining algorithms. Association is the invention of association relationships. Classification based on association is the grouping of Association rule mining and Classification rule mining. It is called as CBA technique [6]. It combines both the advantages of Classification and Association. In association, rules are found based on the specification of user-Constraints. In Classification rule mining, a few rules are found to develop an accurate model.

Advantages
1. It provides the most accurate rules for classification.
2. It provides performance better than other algorithms.

Disadvantages
1. Selecting a simple rule will affect the accuracy.
2. For a large number of rules, training sets and long pattern rules, the efficiency will be low [7].

b. Bayesian classification

Naive Bayes Classification is a straightforward probabilistic classifier depends on applying Bayes theorem with strong (naive) independence postulation. A more descriptive term for the elementary probability model would be “self-reliant characteristic miniature”. Hinge on the actual nature of the probability model, these classifiers can be trained very steadily in a setting of supervised learning. Although their naive design and apparently oversimplified assumptions, Bayes classifiers have worked utterly well in many complicated real-world situations. Abroad comparison with other methods of classification showed that Bayes classification is exceed by more trending approaches, such as boosted trees or random forests [8].

Advantages
1. Bayes classifier only requires cramp amount of data to evaluate the classification.
2. It uses a very innate technique. Bayes classifiers, contrast to neural networks, do not have many free

Fig.4 Human System and the neural network model

Fig.3 Types of Classification Algorithms
parameters that must be specified. So the design process is simplified.
3. As the classifier returns probabilities, it is very easy to apply these results to a large number of tasks than if an arbitrary scale was used[9].

Disadvantages
1. Ascertain independence of characteristics.
2. There is a possibility for false positives.

c. Support Vector Machine
SVM’s are set of supervised learning methods. It is used in classification and regression. They reside in the group of generalized linear classification. An important feature of SVM is, SVM concurrently reduce the error of empirical classification and increase the geometric margin. So it is called Maximum Margin Classifiers. SVM is situated on the Structural risk Minimization (SRM). SVM trace input vector to a higher dimensional space where a maximal separating hyper plane is developed. Two parallel hyper planes are developed on each side of the hyper plane that divide the data. The dividing hyper plane is the hyper plane that increase the distance amidst the two parallel hyper planes. A suspicion is made that the greater the margin or distance between these lateral hyper planes the finer the generalization error of the classifier [10].

Advantages
1. It works well in high dimensional spaces.
2. It works well in cases where number of dimensions is greater than number of samples.
3. It uses subset of training points in decision function.

Disadvantages
1. Biggest issue is speed and size in training and testing.
2. Even though they have good working, they are little bit slowly in testing phase.

d. Decision Tree
Decision tree is one of the popular data mining algorithms. In decision tree, each branch refers to question for classification. It is a prognostic model and objects are categorized by pursuing the path on the decision tree by considering the edges and the values associated with the object. It is used in prediction work, exploration analysis and data pre-processing [4].

The procedure in decision tree algorithms is very akin when they build trees. These algorithms look at all available different questions that could possibly divide the initial training dataset into segments that are roughly alike with reference to the distinguishing classes being prognoses. Few decision tree algorithms use heuristics in order to elite the questions. As example, CART (Classification and Regression Trees) elites the questions in a much authentic way as it tries full of them. After it tried, CART chooses the finest one, uses it to divide the data into more classified segment and then again ask all doable questions on each of this novel segment singly. The Decision tree for a two year child is shown in fig.5. In this figure each branch represents one decision and followed by its causes.

Advantages
1. It can handle mostly all variables even with mislaid values.
2. Decision tree is not affected by Co-linearity and Linear-separability issues.
3. The portrayal of tree is used to analyze the cause of inspected behavior.

Disadvantages
1. It is not used widely in diagnostic test.
2. It do not trespass peculiar requirements.
3. It does not bout the pursuance of linear regression.

III. CONCLUSION
In this paper the most important data mining algorithms are discussed from the research point of view. It has been well understood that each data mining algorithms has its own strength and drawbacks. Each algorithm is useful in some field. Most popular applications use data mining techniques for efficient performance. Data mining is one of the most popular research area. Hence it should be implemented at the company level to take finer decisions.

REFERENCES
[7]. Bing Liu, Wynne Hsu, Yiming Ma, "Integrating Classification and Association Rule Mining." KDD-98, 1998.


**BIOGRAPHIES**

**D. Keerthikawas** born in Pudukkottai in the year 1993. She received her Bachelor degree B.Tech Information Technology from Anna University of Technology, BIT Campus, Anna University, Chennai. She is now pursuing Master’s Degree M.E Computer Science and Engineering at Valliammai Engineering College, Anna University, Chennai.

**G. Sangeetha** was born in Pondicherry in the year 1977. She received her B.E degree in Computer Science and Engineering from University of Madras Chennai in the year 1999, and M.E Degree in Computer Science and Engineering from Sathyabama University, Chennai in 2005. She has fifteen years of teaching experience in various academic institutions. Currently she is working as Assistant Professor in the department of computer Science and Engineering at Valliammai engineering College, Chennai. She has authored a book “Computer Practice I” in the year 2004. She is a lifetime member of CSI (Computer Society of India) since 2005.