A DETAILED STUDY ON MACHINE LEARNING TECHNIQUES FOR DATA MINING

Sivaramakrishnan R Guruvayur
Research Scholar, Jain University
Bangalore, India
gr_shibu@hotmail.com

Dr. Suchithra R
Head, Department of MSc(IT) Bangalore, India
suchithra.suriya@gmail.com

ABSTRACT: **Data mining is the way of extracting the useful information, patterns from large volume of information by using various techniques. It is a powerful technology with great potential to help businesses to make full use of the available data for competitive advantages. This paper discusses various machine learning techniques and the detailed processes of Knowledge Discovery in Databases (KDD).This study also focus on various DM/ML approaches such as Classification, Clustering and Regression and discuss different types of each approach with its advantages and disadvantages.**

Keywords: Data mining, Machine Learning, Bayesian Network, Decision tree Induction, Support Vector Machine.

I. INTRODUCTION

Data mining and machine intelligence are currently a hot debated research area and are connected in database, artificial intelligence, and statistics and so on to find important information and the patterns in big data accessible to clients. Data mining is mainly about training unstructured information and extracting important data from them for end clients to help business choices. Data mining methods utilize scientific calculations and machine intelligence strategies. The prominence of such strategies in dissecting business issues has been upgraded by the arriving of huge information [1].

Data mining has turned out to be a standout amongst the most imperative tools for separating and handling data and to establish patterns to create helpful data for decision making. Of late, there has been a considerable measure of breakthrough in data gathering innovation, for example, standardized bar-code scanners in business spaces and sensors in logical and modern parts, which has prompted the era of enormous measures of data. This exceptional development in big data and databases has delivered a huge interest for new methods and tools that can change big data into helpful data. Statisticians, database researchers, and the MIS and Business people group began utilizing the term "data mining" at first to extract valuable data from big data. One of the procedures including data mining is known as Knowledge Discovery in Databases (KDD), which is utilized for finding helpful knowledge from data. KDD includes data preparation, determination, cleaning and legitimate understanding of consequences of the data mining procedure to guarantee helpful data is gathered from the data. Data mining contrasts from customary data investigation and statistical methodologies in that it utilizes logical systems from a few controls, for e.g. numerical investigation, pattern matching and areas of artificial intelligence, for example, machine learning, and neural systems and genetic algorithms [2],[3].Data mining, or knowledge discovery in databases, is the process of extracting knowledge from large databases.

In Data mining there are three types used to group objects into identified classes such as classification, regression and clustering [12] which is shown in Fig1.

Classification is used to separate the information into classes. A characterization of the classes can then be utilized to make expectations for new unclassified information. Classes can be generally binary partition, or can be difficult and multi-valued. There are two phases included in Classification. The first is learning process phase in which the analysis of training data is done then the creation of rules and patterns. The second phase used to tests the data and archives the accuracy of classification patterns [3].

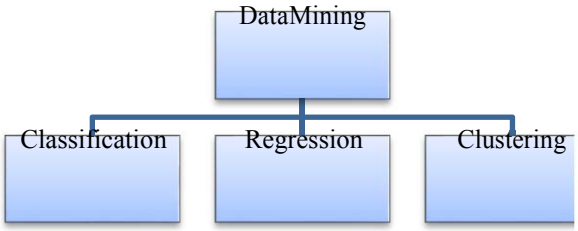

Fig1.Types of Data mining Approaches

Clustering: This approach comes under unsupervised learning because there are no predefined classes. The data may be grouped together as a cluster in this concept [4].

Regression: This is used to map data item into a really valuable prediction variable. Regression analysis can be utilized to show the connection between one or more free factors and dependent factors.

There is a significant overlap between Machine Language and Data Mining. These two terms are always confused because they regularly utilize similar strategies and hence overlap essentially. The pioneer of ML, Arthur Samuel, characterized ML as a "field of study that gives computers the ability to learn without being explicitly programmed." Machine Learning concentrates on prediction and Classification, in view of known properties already learned from the training information. Machine Learning calculations require an objective from the area (e.g., subordinate variable to predict). Data Mining concentrates on the revelation of known properties in the data. It needn't bother with a particular objective from the domain, yet concentrates on finding new and interesting knowledge.

A ML approach generally comprises of two stages: Training and testing. Regularly, the accompanying steps are performed:

Identify class attributes (elements) and classes from Training data.

Identify a subset of the attributes essential for classification.

Learn the model utilizing training data.

Use the trained model to group the unknown information

II. TYPES OF MACHINE LEARNING ALGORITHMS

Machine learning algorithms can be arranged in the following way:

Supervised Learning: These sorts of algorithms are the ones that are trained on illustrations called labeled cases where the inputs are furnished with the desired result already known.

Unsupervised Learning: Unsupervised machine learning is the machine learning task of inducing a function to depict concealed structure from "unlabeled" data. Since the unlabeled examples are given to the learner, there is no assessment of the correctness of the structure that is resulted by the applicable algorithm—which is considered as one way of recognizing unsupervised learning from other two learning method.

The conventional technique for transforming information into knowledge depended on manual examination and interpretation by a domain expert so as to discover valuable patterns in information for decision support.

In Fig[2] the overall process of KDD is shown.

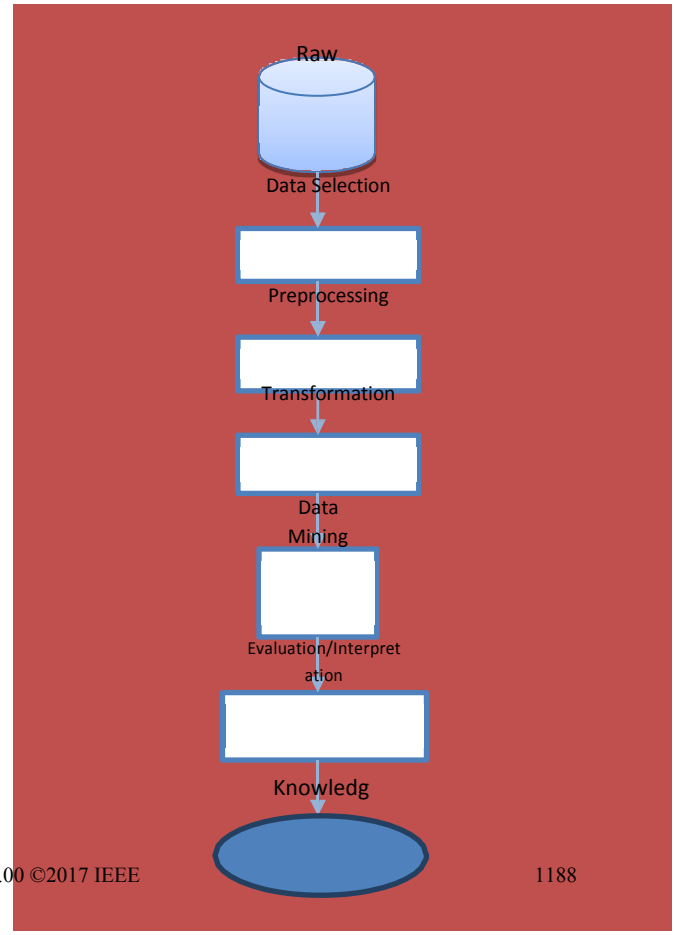

Fig2.Overall process of KDD

KDD process consists of the following steps[5]:

Understanding the application area: incorporates pertinent prior knowledge and objectives of the application. s

Extracting the target data set: incorporates choosing data set or concentrating on a subset of factors.

Data cleaning and preprocessing: incorporates fundamental operations, for example, noise removal and handling of missing information. Data from real-world sources are regularly erroneous, inadequate, and conflicting, maybe because of operation error or framework execution defects. Such low quality information should be cleaned before data mining.

Data integration: incorporates coordinating various, heterogeneous data sources.

Data reduction and projection: incorporates finding helpful elements to represent the data and utilizing dimensionality lessening or change strategies. 6) Choosing the function of data mining: incorporates choosing the motivation behind the model determined by the data mining algorithm

Choosing the data mining algorithm(s): incorporates choosing method(s) to be utilized for searching patterns in data, for example, settling on which model and parameters might be suitable.

Data mining: incorporates scanning for patterns of intrigue in a specific representational form or an arrangement of such indications.

Interpretation: It incorporates interpretation of the discovered data and additionally the possible perception of the extracted designs (patterns).

Using discovered knowledge: It incorporates consolidating this knowledge into the execution framework, taking actions with respect to knowledge.

Data mining involves model to discover patterns which consists of various components.

A. Classification

Classification is a supervised sort of machine learning in which there is arrangement of labeled information in advance. The classifier-training algorithm utilizes these pre-grouped cases to decide the set of parameters required for appropriate separation. The algorithm then encodes these attributes into a model named a classifier.

There are many classification methods available in data mining and the common techniques are as follows:

(a) Decision tree induction: Decision tree can be built from the class labeled tuples. It is like a tree like structure in which there are interior node, branch and leaf node. Interior node determines the test on trait, branch indicates the result of the test and leaf node speaks about the class label. Two stages that are learning and testing are straightforward and quick. The fundamental objective is to anticipate the result for continuous attribute be that as it may; decision tree is less fitting for assessing tasks. There might be mistakes in predicting the classes by utilizing decision tree approach. Pruning calculations are costly and building decision tree is additionally a costly errand as at each level there is division of node. There are several data mining algorithms such as C4.5, ID3, CART, J48, NB Tree, REP Tree etc.

(b) Bayesian Network (BN) is a graphical model for connections among an arrangement of different variable components. This graphical model structure S is a coordinated acyclic graph (DAG) and every one of the nodes in S are in coordinated correspondence with the components of an data set. The arcs represent the effects among the components while the lack of needed arcs in S encodes restrictive independence. Bayesian classifier has shown high exactness and speed when connected to vast databases [6] [7] Bayesian systems are utilized for displaying information Bioinformatics, engineering, medicines, Bio monitoring.

(c) Support vector machine (SVM) is a classification training algorithm. It prepares the classifier to predicate the class of the new sample. SVM is based on the machine learning algorithm designed by Vapnik in 1960's. It is additionally in view of the structure chance minimization rule to anticipate over fitting. This is a classifier in view of finding an isolating hyper plane in the component space between two classes in such a way that the separation between the hyper plane and the nearest data purposes of each class is augmented. The approach depends on a limited characterization chance [8] rather than on ideal classification.

B. Clustering

Clustering is a data mining procedure of grouping set of data items into various clusters or groups so that objects inside the bunch have high similarity, however are extremely dissimilar in alternate groups. Clustering algorithms are used to organize data, categorize data, for data compression and model construction, for detection of outliers etc.

Common Data clustering techniques are discussed.

(a) The k-means algorithm [9] is the most famous grouping way utilized these days in logical and mechanical applications. The name originates from representing to each of the k clusters C_j by the mean (or weighted normal) c_j of its focuses, the so-called centroid. While this portrayal does not work well with all attributes, it works well from a geometrical and statistical point of view for numerical qualities. The total of distance between components of a collection of points and its centroid communicated through a proper distance capacity is utilized as the target function.

(b) Hierarchical clustering consolidates data objects into clusters, those clusters into larger groups, etc, making a hierarchy. A tree which represents the command of groups is known as a Dendrogram. Singular data objects are the leaves of the tree, and the inside nodes are nonempty clusters. Sibling nodes divide the points secured by their common parent. This permits investigating information at various levels of granularity. Hierarchical clustering is strategies are ordered into agglomerative (bottom-up) and divisive (top-down) methodologies.

There are a various methodologies available for data clustering. In connectivity models (e.g., hierarchical clustering), information focuses are assembled by the distance between them. In centroid models (e.g., k-means), each cluster is mentioned by its mean vector. In distribution models (e.g., Expectation Maximization algorithm), the gatherings are thought to be submissive to a factual conveyance. Density models group the data points as dense furthermore, associated areas (e.g., Density-Based Spatial Clustering of Applications with Noise [DBSCAN]). Finally, graph models (e.g., clique) characterize each group as an arrangement of associated nodes (information focuses) where every node has an edge to at least one other node in the set[10].

C. Regression

There are two sorts of regression techniques, such as linear and non –linear [11].

(a) Linear regression: Linear regression is utilized where the connection amongst target and predicator can be represented in straight line. The advantage of using linear regression is it is easy to understand the hypothetical function.

$$\hat{y} = \theta_1 x + P2 + e$$

(b) Multivariate linear regression: The regression line can't be envisioned in two dimensional space.

$$\hat{y} = \theta_1 + \theta_2 x_2 + \theta_3 x_3 + \cdots + \theta_{n-1} x_{n-1} + \ell$$

(c) Non-Linear Regression: For this situation non-linear relationship can be there and this can't be mentioned to as straight line. This can be represented to as linear response by preprocessed the information.

III. COMPARISON OF DIFFERENT DATA MINING TECHNIQUES

TABLE I. COMPARISON OF DIFFERENT DATA MINING ALGORITHMS

Algorithm	***Findings***	***Drawbacks***
Decision Tree	It can deal with both consistent and discrete information. It gives quick result in classifying unknown records. It gives great comes about with small measure tree. Results does not influence with anomalies. It doesn't require preparation technique like normalization. It functions better with numeric information	It won't be able to predict the value of a continuous class attribute It gives error contained message when large number of classes used Unrelated attribute may leads to bad manner decision trees Even small changes made to the data can modify complete decision tree.
Naïve Bayesian	Compared to other classifier it gives less error rate. Easy to adapt It can handle continues data in a good manner When work with large database it gives high	Provides less accuracy since it concentrates more on independent features.

	accuracy and speed It can handle with discrete values	
Neural Networks	Used to classify the pattern on untrained data Works well with continuous values	Less interpretability It takes long training time.
K-Means	Simple and efficient algorithm It is relatively fast. It gives better result when distinct data is used	Does not work with noisy data and non-linear datasets.
Support vector machine	It can produce accurate and error free classification results even when input data are non-monotone and non-linearly separable. SVMs provide a good out-of-sample generalization, if the parameters are appropriately chosen. It can produce a unique solution, since the optimality problem is convex.	lack of transparency of results Scale dependent , iterative It works on slow training , nonlinear
Hierarchical Clustering	Embedded adaptability in regards to the level of granularity. x Well suited for issues including point linkages, e.g. scientific classification trees	Less interpretability with respect to cluster descriptors. Incorrect termination criterion Inability to make corrections once the splitting/merging decision is made.

IV. RELATED WORK

In [10] the authors have done a literature survey on machine learning and data mining methods for cyber analytics in support intrusion detection. They have discussed about various ML/DM methods. Described about well -known cyber data sets and the complexity of ML/DM. Various challenges for using ML/DM is addressed. In [8] authors have discussed about various classification techniques and have done a comparative analysis of different classification algorithms. The various classification techniques are decision tree, Support vector Machine, Nearest Neighbor etc. In [9] this paper for extending the K-means algorithm the authors have presented two algorithms with categorical domains and domains with mixed numeric and categorical values. Authors have used the well-known soybean disease and credit approval data sets for demonstrating the clustering performance of the two algorithms. In [5] the paper gives an overview of data mining and knowledge discovery database field, clearing up how data mining and knowledge discovery in databases are connected both to each other and to related fields, for example, machine learning, statistics, and databases. The work concentrates on specific real applications, particular data mining methods, challenges required in certifiable uses of knowledge discovery, and present and future research headings in the field.

V. CONCLUSION

This paper gives a survey on Machine learning techniques for data mining. Throughout the years data mining has delighted in enormous achievement, the application domains extended persistently yet the mining methods additionally kept up moving forward. Various issues have developed and solution have found by data mining scientists. In any case, there are ranges and issues that still require consideration for future upgrades in this innovation. More research on the most proficient method to manage the social issue of in some cases, unconscious and unsuspecting people's security require to be conducted. Data mining procedures should accordingly develop to coordinate with this challenge.

REFERENCE

[1] U Fayyad, G Piatetsky-Shapiro, P Smyth, “From Data Mining to Knowledge Discovery in Databases,” AI Magazine, vol.17, no.3, pp. 37-54, 1996.

[2]. P. R. Peacock, “Data mining in marketing: Part 1”, Marketing Management, pp. 9-18, 1998.

[3] Balagatabi, Z. N., & Balagatabi, H. N. (2013). Comparison of Decision Tree and SVM Methods in Classification of Researcher's Cognitive Styles in Academic Environment. Indian Journal of Automation and Artificial Intelligence, 1(1), 31- 43.

[4] A Survey of Clustering Data Mining Techniques P. Berkhin.

[5] U. Fayyad, G. P. Shapiro, and P. Smyth, "The KDD process for extracting useful knowledge from volumes of data," Commun. ACM, vol. 39, pp. 27–34, 1996.

[6]Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. Machine learning, 29(2-3), 131-163.

[7]. Jensen, F. V. (1996). An introduction to Bayesian networks (Vol. 210). London: UCL press.

[8] Machine Learning Techniques for Data Mining: A Survey, Seema Sharma1 , Jitendra Agrawal2 , Shikha Agarwal3 , Sanjeev Sharma.

[9] Huang Z (1998) Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. Acsys CRC, CSIRO

[10] A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection Anna L. Buczak, Member, IEEE, and Erhan Guven, Member, IEEE.

[11] Data Mining - Techniques, Methods and Algorithms: A Review on Tools and their Validity, Mansi Gera and Shivani Goel , International Journal of Computer Applications (0975 – 8887) Volume 113 – No. 18, March 2015

[12] J. Han and M. Kamber, "Data mining: concepts and techniques", Morgan-Kaufmann Academic Press, San Francisco, 2001.

[13] Decision Tree Induction: An Approach for Data Classification Using AVL-Tree ,Devi prasad Bhukya and S. Ramachandram, International Journal of Computer and Electrical Engineering, Vol. 2, No. 4, August, 2010 1793-8163.