Study of Data Mining Concepts

Ashish Barhate
Department of Computer Engineering
Saraswati College of Engineering, Kharghar

Sumit Gupta
Department of Computer Engineering
Saraswati College of Engineering, Kharghar

Shreyas Kinage
Department of Computer Engineering
Saraswati College of Engineering, Kharghar

Prof.Suhasini Parvatikar
Department of Computer Engineering
Saraswati College of Engineering, Kharghar

Abstract— presently, a very large amount of data stored in databases is increasing at a tremendous speed. This requires a need for new techniques and tools to aid humans in automatically and cleverly analyzing large data sets to acquire useful information. This growing need gives a view for a new research field called Knowledge development in Databases or Data Mining, which attract a attention from researchers in many different fields including database design, statistics, pattern recognition, database design, machine learning, and data visualization. Data mining is the process of discovering insightful, novel patterns and interesting, as well as descriptive, understandable and predictive models from large-scale data. In this paper we over viewed different tasks includes in Data mining. Data mining involves the tasks like anomaly detection, classification, regression, association rule learning, clustering and summarization.

Keywords—Data Mining, classification, clustering, association rules

I. INTRODUCTION

The last decade has experienced a uprising in information accessibility and exchange of it through internet. In the same strength more business as well as organizations began to collect data related to their own operations, while the database technologist have been seeking efficient mean of retrieving, storing and manipulating data, the machine learning community focused on techniques which used for developing, learning and acquiring knowledge from the data. Data Mining is the process of analyzing data from summarizing and different perspectives it into useful information. Data mining consists of transform, extract and load transaction data onto the data warehouse system, store and manage the data in a multidimensional database system, by using application software analyze the data, provide data access to business analysts and information technology professionals, present the data in a useful format, like a graph or table. Data mining involves the anomaly detection, association, classification, regression, rule learning, summarization and clustering.

II. DATA MINING

Data mining is the exploration and analysis of large data sets, in order to determine meaningful pattern and rules. The key idea is to find effective way to combine the computer's power to process the data with the human eye's ability to spot pattern. The objective of data mining is to design and work efficiently with large information sets. Data mining is the component of wider process called knowledge discovery from database. [4]. Data Mining is the process of analyzing data from different perspective and summarizing the results as useful information. It has been defined as "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" The definition of data mining is closely related to another commonly used term knowledge discovery [2]. Data mining is an integrated database, interdisciplinary, machine learning, artificial intelligence, statistics, etc. Many areas of hypothesis and technology in current era are databases, data mining, artificial intelligence and statistics is a study of three strong large technology pillars. Data mining is a multi-step process, requires accessing and preparing data for a mining the data, data mining algorithm, analyzing results and taking proper action. The data, which is accessed, can be stored in one or more prepared databases. In data mining the data can be mined by passing a variety of processes.

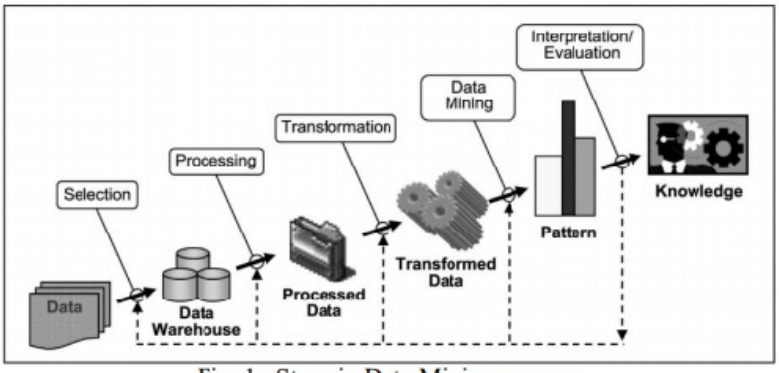

Fig. 1 : Steps in Data Mining process

In data mining the data is mined using two learning approaches that is supervised learning or unproven learning [5].

A. Supervised Learning:

In supervised learning (often also called directed data mining) the variables under examination can be split into two groups: explanatory variables and one dependent variables. The goal of the analysis is to specify a connection between the dependent variable and explanatory variables the as it is done in failure analysis. To precede with directed data mining techniques the value of the dependent variable must be known for a sufficiently large part of the data set.

B. Unsupervised Learning:

In unverified learning, all the variables are treated in same way; there is no difference between dependent and instructive variables. However, in contrast to the name undirected data mining, still there is some target to complete. This target might be as data reduction as general or more exact like clustering. The separating line between unverified learning and supervised learning is the same that distinguishes discriminate analysis from cluster analysis. Supervised learning require, target variable should be well defined and that a enough number of its values are given. Unsupervised teach typically either the target variable has only been recorded for too small a digit of cases or the target variable is unknown.

III. ISSUES IN DATA MINING

Data mining has evolved into an significant and active area of research because of the theoretical challenges and practical applications associated with the problem of discovering interesting and previously unidentified knowledge from real-world databases. The main challenges to the data mining and the equivalent concern in designing the algorithms are as follows:

a. Massive data sets and high dimensionality.

b. Over fitting and assess the statistical significance.

c. Understandability of patterns.

d. Non-standard incomplete data and data integration.

e. Mixed changing and redundant data.

IV. USED IN PREDICTIVE DATA MINING

A. C4.5

C4.5 is an algorithm used to produce a decision tree developed by Ross Quinlan [1]. It is an extension of Quinlan's prior ID3 algorithm. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a factual classifier. [6]This algorithm builds decision trees from a set of preparation data using the concept of information entropy. It is managing both continuous and discrete attributes, handling training data with missing attribute values and also handling attributes with differing costs. In building a decision tree we can deal with training sets that have records with unidentified attribute values by evaluating the gain, or the gain ratio, for an attribute by taking into consideration only the records where that attribute is defined. In using a decision tree, we can classify records that have unidentified attribute values by estimating the probability of the various possible results. [12]

a) Follows the algorithms employed in C4.5 using decision tree.

1 Decision trees

Given a set S of cases, C4.5 first grows an initial tree using the divide-and-conquer algorithm as follows:[7] If all the cases S belong to the same class or S is small, the tree is a leaf labeled with the most recurrent class in S. or else, choose a test based on a single attribute with two or more outcomes. Make this test the origin(root) of the tree with one division for each outcome of the test, partition S into consequent subsets *S*1, *S*2, . . . according to the outcome for each case, and apply the same process recursively to each subset. Use also information gain or gain ratio to rank the possible tests. Check the error. [10]

B. CLUSTERING ALGORITHM

Clustering is the task of handing over a set of objects into groups so that the objects in the same cluster are more similar to each other than to those in other clusters. Clustering is a key task of explorative data mining, and a common method for data analysis used in many fields including information recovery. Cluster study groups objects based on their similarity. The measure of similarity can be computed for a variety of types of data. [11] Clustering algorithms can be categorized into partitioning methods, hierarchical methods, density-based methods, grid-based methods, and model-based methods, k-means algorithm, graph based model etc.

a) K means algorithm

In data mining K-means clustering is a technique of cluster analysis which aim to partition n observations into k clusters in which each observation belongs to the cluster with the adjacent mean. The k-means algorithm (also referred as Lloyd's algorithm**)** is a easy iterative process to partition a Given dataset into a user specific number of clusters, k.[8]The algorithm operate on a set of dimensional

vectors, D = {**xi** | i = 1, . . . , N}, where **xi**

_d denotes the i th data point. The algorithm is initialized by selection k points in _d as the first k cluster representatives or "centroids". Techniques for selecting these initial seeds include sampling at random from the dataset, setting them As the solution of clustering a small subset of the data or perturbing the global mean of the data *k* Times. Then the algorithm iterates between two steps till convergence:[14]

Step 1: Data Assignment. Each data point is assigned to its *closest* centroid, with ties broken down arbitrarily. This results in a partitioning of the data.

Step 2: Relocation of "means". Every cluster representative is relocated to the center (mean) of every data points assigned to it. If the data points come up with a probability calculate (weights), then the rearrangement is to the expectations (weighted mean) of the data partitions. The algorithm converges when the assignments no longer change. Note that each iteration wants $N \times k$ comparisons, which determines the time difficulty of one iteration. The number of iterations compulsory for convergence varies and may depend on *N*, but as a first cut, this algorithm can be considered linear in the dataset size.

Advantages	**Limitations**
Relatively efficient and easy to implement.	Sensitive to initialization
Terminates at local optimum.	Limiting case of fixed data.
Apply even huge data sets	Difficult to compare by different numbers of clusters
The clusters are non-hierarchical and they do not overlap	Needs to specify the number of clusters in advance.
With a large number of variables, K-Means may be computationally faster than hierarchical clustering	Unable to handle noisy data or outliers.

Table 1: Advantages and Limitations of k-means algorithm

C. ASSOCIATION RULES ALGORITHM :

Association rule mining search for interesting relationships amongst items in a given set. Here the main rule interestingness is rule support and confidence which replicate the utility and certainty of discovered rules.

Association rule mining algorithm is a two step process where we should have to find all the frequent item sets and produce strong association rules from the frequent item sets.[9] If a rule concern association among the presence or absence of items, it is a Boolean association rule. Here the quantitative values for items are partitioned into interval. The algorithm can be shaped based on dimensions, based on level of abstractions involved in the rule set and also based on various extensions to association mining such as correspondence analysis.[13]

1. Multi dimensional Association Rules:

In multi dimensional databases, each unmistakable predicate in a control as a measurement. Affiliation decide that include at least two measurements or predicates every one of which happens just once in the govern can be alluded as multidimensional affiliation rules. Multi measurement affiliation rules with no rehashed predicates are called bury measurement affiliation manages and may with rehashed predicates which can contain numerous events of a few predicates are Called half breed measurement affiliation rules.

For example:

Age(X,50…..70)^ FAMILYHISTORY(X,DISEASE)=> DISEASEHIT(X,"TYPHOID").

Here the database attributes can be categorical or quantitative with no ordering among the values The basic definition of association rule states that, Let A={l1,l2,……..lm} be a set of items, and Let T, the transaction database, be a set of transaction, where each transaction t is a set of items and thus t is a subset of A. An association rule tells us about the association between two or more items. For example, If we are given a set of items where items can be referred as disease hit in an area and a large collection of patients who are subsets of some inhabitants in the area. . The task is to find relationship between the presences of disease hit within these group. In order for the rules to be useful there are two pieces of information that must be supplied as well as the actual rule: Support *is* how often does the rule apply? and confidence is how often is the rule is correct. [15]In fact association rule mining is a two-step process: Find all frequent item sets / disease hit - by definition, each of these item sets will occur at least as frequently as a predetermined minimum support count, and then generate strong association rules from the frequent item sets by definition, these rules must satisfy minimum support and minimum confidence. In this study predicting the chances of disease hit an area, by correlating the parameters or attributes such as climate, environmental condition, heredity, education with the inhabitants. And also finding how these parameters are associated with the chances of disease hit.

IV. CONCLUSION

In this exploration work, I have made an examination to make a correlation of the a portion of the current information digging calculation for high dimensional information bunches to evaluate expectation in information mining method. The principle methods incorporated into the overview are choice tree, bunching calculation, k-implies calculation and affiliation manage calculation. Considered every calculation with the assistance of high dimensional informational collection with UCI vault and discover the preferences and impediments of each. By contrasting the points of interest and detriments of every calculation, I am attempting to build up a cross breed calculation for multidimensional information examination. The productivity was figured based on time multifaceted nature, space many-sided quality, space prerequisites and so forth. The example utilized in this investigation incorporates UCI store. The proficiency of new calculation can be checked with constant information.

REFERENCES

[1] en.wikipedia.org/wiki/Data_mining
[2] Pavel Berkhin, A Survey of Clustering Data Mining Techniques, pp.25-71, 2002.
[3] R. Agarwal, T. Imielinski and A. Swamy "Mining association Rules between Set of Items in Large Database".In ACM SIGMO international conference on Management of Data .
[4] Han, J., Kamber, M., Data Mining Concepts and Techniques, Morgan Kaufmann Publisher, 2001
[5] K.Kameshwaran, K.Malarvizhi, Survey on Clustering Techniques in Data Mining, IJCSIT, Vol. 5, 2014, 2272-2276
[6] R. Agarwal, T. Imielinski and A. Swamy "Mining association Rules between Set of Items in Large Database".In ACM SIGMO international conference on Management of Data.
[7] "Classification Rules By Decision Tree for disease Prediction", Smitha.T, Dr. V. Sundaram, IJCA, vol-
[8] 43, No-8, April 2012
[9] "Knowledge Discovery from Real Time Database using Data Mining Technique", Smitha.T, Dr. V.Sundaram, IJSRP vol 2, issue 4, April 2012.
[10] "Another Look at Measures of Forecast Accuracy" Hyndman R and Koehler A(2005).
[11] BORG, I., and GROENEN, P. (1997): 'Modern multidimensional scaling: theory and application' (Springer-Verlag, New York, Berlin, Heidelberg, 1997).
[12] David Hand, Heikki Mannila, Padhraic Smyth," principles of Data Mining".

[13] Shekar B, Natarajan R 2004b A transaction-based neighbourhood-driven approach to quantifying interestingness of association rules. *Proc. Fourth IEEE Int. Conf. on Data Mining (ICDM 2004)*(Washington, DC: IEEE Comput. Soc. Press) pp 194–201
[14] El-taher, M. Evaluation of Data Mining Techniques, M.Sc thesis (partial-fulfillment), University of Khartoum, Sudan ,2009.
[15] Lee, S and Siau, K. A review of data mining techniques, Journal of Industrial Management & Data Systems, vol 101, no 1, 2001, pp.41-46.
[16] "A New Approach for Evaluation of Data Mining Techniques", Moawia Elfaki Yahia1, Murtada Elmukashfi El-taher2, IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 5, September