RESEARCHERID
THOMSON REUTERS

[Dokania * *et al.*, 7(5): May, 2018]
IC™ Value: 3.00

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

COMPARATIVE STUDY OF VARIOUS TECHNIQUES IN DATA MINING

Mr. Nilesh Kumar Dokania[*1] & Ms. Navneet Kaur[2]
[*1]Assistant Professor (Guru Nanak Institute of Management)
[2](Guru Nanak Institute of Management) (Research, MCA)

DOI: 10.5281/zenodo.1241440

ABSTRACT

Data mining (knowledge discovery from data) may be viewed as the extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns and models from observed data or a method used for analytical process designed to explore data. We know Data mining as knowledge discovery. Basically Extraction or "MINING" means knowledge from large amount of data. We use Data mining due to the explosive growth of data i.e. from terabytes to petabytes. We are drowning in data, but starving for knowledge! Alternative names of Data mining are: Data archeology, Data dredging, Information harvesting, Business intelligence, etc. Data mining techniques are used to find the hidden or new patterns to store the data. We know that data mining can use every sector like business, agriculture, marketing etc. There are many techniques for data mining like clustering, classification etc. There are various approaches and techniques of data mining which can be applied on data to build up a new environment to improve performance of existing data and help to create the new predictions on the data. [1].

Keywords: DM, KDD, data hiding, k-mean, Y-mean.

I. INTRODUCTION

Data Mining (DM) is the process of analyzing data from different perspectives and summarizing it into useful information. It is the efficient discovery of valuable, non-obvious information from a large collection of data. It is a knowledge discovery process helps us to understand the substance of the data in special unsuspected way.
Data mining consists of extract, transform, and load transaction data onto the data warehouse system, Store and manage the data in a multidimensional database system, Provide data access to business analysts and information technology professionals, Analyze the data by application software, Present the data in a useful format, such as a graph or table. [2]

Here is the list of steps involved in Knowledge discovery process:

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

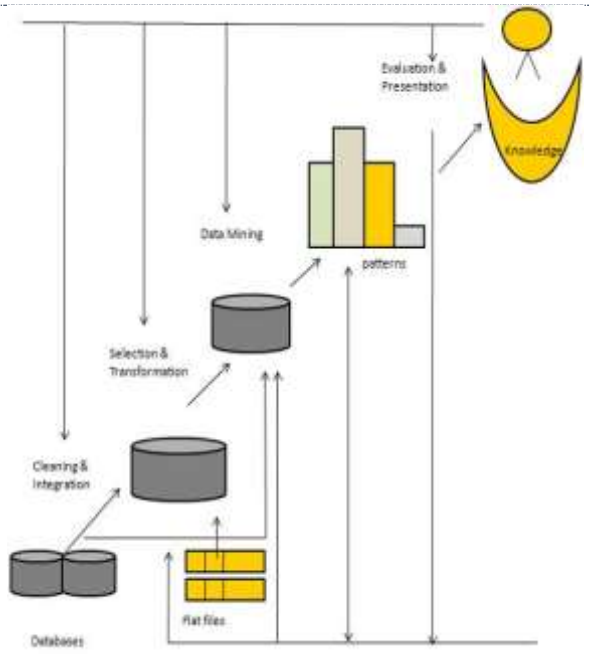

Fig 1: KDD Process

- Data cleaning
- Data Integration
- Data Selection
- Data Transformation
- Data Mining
- Pattern Evaluation
- Knowledge Presentation

II. LITERATURE REVIEW

This section summarizes various review and technical articles on data mining techniques. Several works have been carried out by many researchers. This section presents a brief summary on the basis of literature. In Sumit Garg & Arvind K. Sharma, the aim of their paper is how to use suitable data mining algorithms on educational dataset. This paper focuses on comparative analysis of various data mining techniques and algorithms.

In Aastha Joshi & Rajneet Kaur, Cluster analysis can be used as a standalone data mining tool to gain insight into the data distribution, or as a preprocessing step for other data mining algorithms operating on the detected clusters. Many clustering algorithms have been developed and are categorized from several aspects such as partitioning methods, hierarchical methods, density-based methods, and grid-based methods.

Further data set can be numeric or categorical. Inherent geometric properties of numeric data can be exploited to naturally define distance function between data points. Whereas categorical data can be derived from either quantitative or qualitative Data where observations are directly observed from counts.

In , Preeti Baser & Dr. Jatinder Kumar, attempt to do comparative analysis of various clustering techniques used for large dataset by comparing classifications of clustering techniques.

In , Mitchell D'silva & Deepa Vora, compare the Data mining techniques to enhance the Intrusion Detection. Different data mining techniques like classification, clustering, association rule mining are frequently to acquire information by the network data.

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

III. DATA MINING TECHNIQUES

Basic types of Data Mining techniques are as follows:-

- Predictive
- Descriptive

IV. TYPES OF PREDICTIVE

1. Classification
   a) Decision Tree
   b) Neutral Network
   c) Nearest neighbor Classification
2. Regression
3. Time Series analysis
4. Prediction

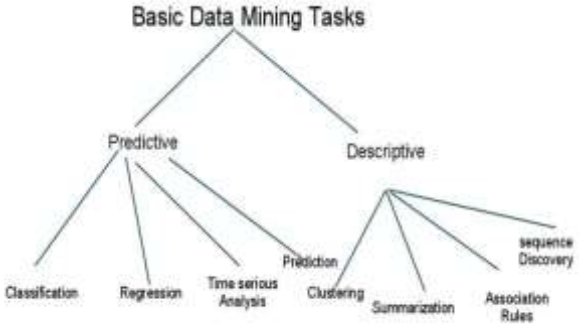

Fig 2: Data Mining Tasks

3.1 Classification: Classification is a data mining (machine learning) technique used to predict group membership for data instances. For example, you may wish to use classification to predict whether the weather on a particular day will be sunny, rainy or cloudy. Popular classification techniques include decision trees and neural networks.

3.2 Regression: Regression is a data mining function that predicts a number. Profit, sales, mortgage rates, house values, square footage, temperature, or distance could all be predicted using regression techniques. For example, a regression model could be used to predict the value of a house based on location, number of rooms, lot size, and other factors. [3]

Regression is a data mining (machine learning) technique used to fit an equation to a dataset. The simplest form of regression, linear regression, uses the formula of a straight line (y = mx + b) and determines the appropriate values for m and b to predict the value of y based upon a given value of x. Advanced techniques, such as multiple regression, allow the use of more than one input variable and allow for the fitting of more complex models, such as a quadratic equation.

3.3 Time Series Analysis: A time series represents a collection of values obtained from sequential measurements over time. Time-series data mining stems from the desire to reify our natural ability to visualize the shape of data. Humans rely on complex schemes in order to perform such tasks. We can actually avoid focusing on small fluctuations in order to derive a notion of shape and identify almost instantly similarities between patterns on various time scales.

3.4 Prediction: Predicting the identity of one thing based purely on the description of another, related thing

1. Not necessarily future events, just unknowns
2. Based on the relationship between a thing that you can know and a thing you need to predict
   Predictor => Predicted

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

3. When building a predictive model, you have data covering both
4. When using one, you have data describing the predictor and you want it to tell you the predicted value.

V. TYPES OF DESCRIPTIVE

1. Clustering
2. Summarization
3. Association Rules
4. Sequence Discovery

4.1Clustering: Clustering is used to store the data into groups according to their values, characteristics, similarities and dissimilates .In this approach same type data are store in same groups and these groups are known as clusters but data is heterogeneous between two clusters. Clusters can be apply on group of some schools to investigate the similarities and differences between these schools, students can be clustered together to study the differences in their behavior. [4]

From the three data mining techniques discussed above clustering is widely used for intrusion detection because of the following advantages over the other techniques:

1. Does not require the use of a labeled data set for training.
2. No manual classification of training data needs to be done.
3. Need not have to be aware of new types of intrusions in order for the system to be able to detect them.

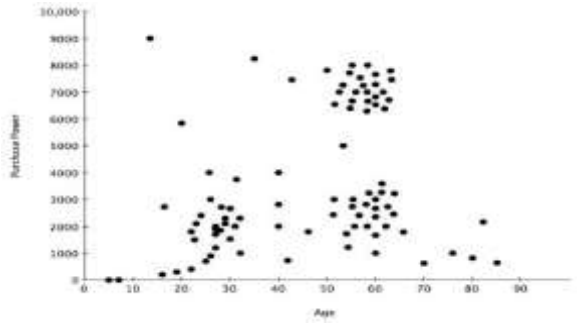

Fig 3: Clustering Technique

Some of the clustering techniques such as K-Means Clustering, Y-Means Clustering are discussed below.

4.2 Requirements of Clustering in Data Mining

1. Scalability **-** We need highly scalable clustering algorithms to deal with large databases.
2. Ability to deal with different kind of attributes **-** Algorithms should be capable to be applied on any kind of data such as interval based (numerical) data, categorical, binary data.
3. Discovery of clusters with attribute shape **-** The clustering algorithm should be capable of detect cluster of arbitrary shape. It should not be bounded to only distance measures that tend to find spherical cluster of small size.
4. High dimensionality **-** The clustering algorithm should not only be able to handle low- dimensional data but also the high dimensional space.
5. Ability to deal with noisy data **-** Databases contain noisy, missing or erroneous data. Some algorithms are sensitive to such data and may lead to poor quality clusters.
6. Interpretability **-** The clustering results should be interpretable, comprehensible and usable.

4.3 K-Means Clustering: K-Means algorithm is a hard partitioned clustering algorithm widely used due to its simplicity and speed. It uses Euclidean distance as the similarity measure. Hard clustering means that an item in a data set can belong to one and only one cluster at a time. It is a clustering analysis algorithm that groups items based on their feature values into K-disjoint clusters such that the items in the same cluster have similar attributes and those in different clusters have different attributes. [5]

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

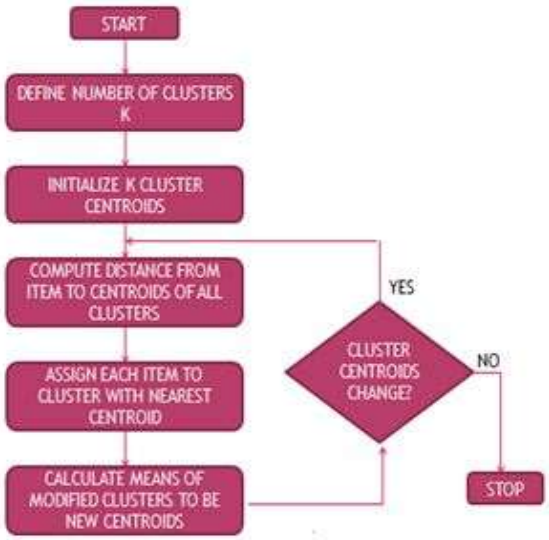

Figure 4: Flowchart of K-Means Clustering

4.4 Y-Means Clustering: Y-means is another clustering algorithm used for intrusion detection. This technique automatically partitions a data set into a reasonable number of clusters so as to classify the data items into normal and abnormal clusters. The main advantage of Y-Means clustering algorithm is that it overcomes the three shortcomings of K-means algorithm namely dependency on the initial centroids, dependency on the number of clusters and degeneracy. Y-means clustering eliminates the drawback of empty clusters. Y-means uses Euclidean distance to evaluate the similarity between two items in the data set. Y-means is an efficient clustering technique for intrusion detection since the network log data is randomly distributed and the value of K is difficult to obtain manually. [6]

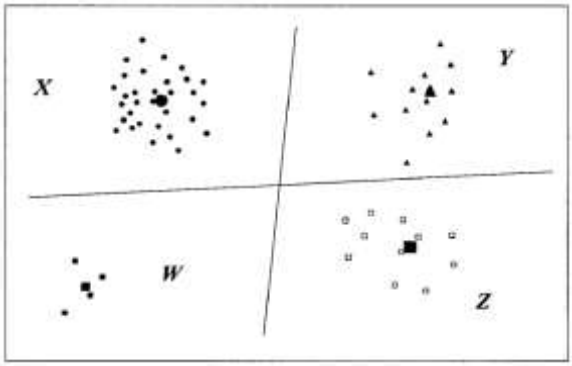

Figure 5: Y-Means Clustering

4.5 Comparison of K-Means and Y-Means Clustering

Criteria	**K-Means**	**Y-Means**
Input	Number of clusters K such that K<m Set of data items (x1,x2,..,xm)	Number of clusters K such that K<m Set of data items (x1,x2,..,xm)

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

Output	Set of K clusters where each cluster has similar	Set of non-empty clusters where each cluster has similar
	Items	Items
Membership value	Does not exist	Does not exist
Computation Time	Simple and straight-forward so requires less Time	Involves splitting and linking of clusters so requires more Time
Purity of cluster	Low	High
Empty cluster generation	May or may not generate	No
Efficiency	Works well for small data Sets	Works well for small as well as large data sets
Number of clusters an item belongs	One	One
Overall	Depends on the initial	Does not depend on the initial
Performance	number of clusters —K‖	number of clusters —K‖
Shape of cluster	Works well for compact and globular clusters	Works well for both globular and nonglobular clusters
Detection Rate	Highest	Higher

Figure 6: Comparison between K-means and Y-means Clustering

4.6 Summarization: Summarization is a key data mining concept which involves techniques for finding a compact description of a dataset. Simple summarization methods such as tabulating the mean and standard deviations are often applied for data analysis, data visualization and automated report generation. [7] Clustering [13, 23] is another data mining technique that is often used to summarize large datasets.

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

4.7 Association Rules: Association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using different measures of interestingness. Based on the concept of strong rules, Rakesh Agrawal et al introduced association rules for discovering regularities between products in large-scale transaction data recorded by point-of-sale (POS) systems in supermarkets. For example, the rule $\{\text{onions}, \text{potatoes}\} \Rightarrow \{\text{burger}\}$ found in the sales data of a supermarket would indicate that if a customer buys onions and potatoes together, he or she is likely to also buy hamburger meat. Such information can be used as the basis for decisions about marketing activities such as, e.g., promotional pricing or product placements. [8]

4.7 Sequence Discovery: Sequential Pattern mining is a topic of data mining concerned with finding statistically relevant patterns between data examples where the values are delivered in a sequence. It is usually presumed that the values are discrete, and thus time series mining is closely related, but usually considered a different activity. Sequential pattern mining is a special case of structured data mining.

There are several key traditional computational problems addressed within this field. These include building efficient databases and indexes for sequence information, extracting the frequently occurring patterns, comparing sequences for similarity, and recovering missing sequence members. In general, sequence mining problems can be classified as string mining which is typically based on string processing algorithms and item set mining which is typically based on association rule learning. [9]

Applications of Data Mining Technique

Application Area	Examples of Mining Functions	Mining Processes	Mining Techniques
Fraud Detection	Credit card frauds Internal audits Warehouse pilferage	Determination of variations from norms	Data Visualization Memory-based Reasoning
Risk Assessment	Credit card upgrades Mortgage Loans Customer Retention Credit Ratings	Detection and analysis of links	Decision Trees Memory-based Reasoning
Market Analysis	Market basket analysis Target marketing Cross selling Customer Relationship Marketing	Predictive Modeling Database segmentation	Cluster Detection Decision Trees Link Analysis Genetic Algorithms

Figure 7: Application of Data Mining

VI. CONCLUSIONS

This paper indicates the capabilities of data mining techniques provide effective improving tools for performance in organization. A comparative analysis of various data mining techniques is presented in this paper. Many data mining techniques can be implemented on data to predict their future performance.

Undoubtedly one of the toughest things to do when deciding to device a data mining technique is the determination of which technique to use and when. Most of the time the technique to be used are determined by trial and error. There are certain differences in the kinds of problems that are most favorable to each technique but the authenticity of real world data and the dynamic way in which markets, customers and data that represents them is formed which means that the data is continuously changing. These dynamics mean that is not always possible to build a "perfect" model on the historical data since whatever was known in the past cannot adequately predict the future. But sometimes the situation is very crucial for business person who is waiting for all information to come in before they make their decision. Since business, economy and even the world are

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

changing in unpredictable and even chaotic ways data mining techniques are not always predictable. So it's safe to select a robust model that may under-perform when compared to best data mining tools for analysis and execution at the earliest to take business decisions before it's too late. [10]

VII. REFERENCES

[1] http://research.ijcaonline.org/volume74/number5/pxc3889673.pdf
[2] http://www.ijarcsse.com/docs/papers/Volume_3/3_March2013/V3I3-0162.pdf
[3] http://www.ijcscn.com/Documents/Volumes/vol3issue5/ijcscn2013030504.pdf
[4] http://www.ijera.com/papers/Vol3_issue1/GP3112671275.pdf
[5] http:s//www.ibm.com/developerworks/library/ba-data-mining-techniques
[6] http://courses.cs.washington.edu/courses/csep521/07wi/prj/leonardo_fabricio.pdf
[7] http://www.saedsayad.com/decision_tree.htm
[8] http://www.statsoft.com/textbook/support-vector-machines
[9] http://www.jatit.org/volumes/research-papers/Vol5No1/1Vol5No6.pdf
[10] https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/regress.htm#DMCON052.

CITE AN ARTICLE

Dokania, N. K., Mr, & Kaur, N. (2018). COMPARATIVE STUDY OF VARIOUS TECHNIQUES IN DATA MINING. *INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY*,*7*(5), 202-209.