International Journal of Scientific Research and Engineering Development-– Volume 3 Issues 2 Mar- Apr 2020

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

A Study of Data Mining Techniques, Applications and Challenges

Mrs.N.Keerthikaa,
Assistant Professor,
Vivekanandha Educational Institutions,
Tiruchengode,Namakkal,
Tamilnadu,India

----------**********----------

Abstract:

Data mining is having spacious applications it is young and promising field for present generation .It has great deal of attention in the information industry and society. The wide availability of large amounts of data and the immediate need for such a data into useful information and knowledge. Organizing and analyzing such type of data is important. The concepts of data mining satisfy this need by providing tools to discover knowledge from data. Knowledge Discovery and Data Mining are interrelated terms they are used interchangeably. Data mining is an algorithmic technique for extracting information from collection of raw data. Data Mining is becoming a popular in various research areas due to its ability to detect hidden patterns or relationship among the objects in the data. These data mining used in different applications and various technique used in this data mining that things will be discussed in this paper.

Keywords **— mining, data discovery, knowledge discovery, classification.**

----------**********----------

I. INTRODUCTION

Data Mining is set of method it applies to large and complex databases. It eliminates the duplicates and discovers the hidden pattern. This method is computationally intensive. Using data mining tools, methodologies and theories for revealing pattern in data. Data mining is became an important are of research and product development. The process of retrieving high-level knowledge from low-level is called knowledge discovery database. Knowledge discovery in database contain different level of steps. The first step is selection of data which it is collected from various sources. The second step is pre processing the collected data. Third step is change the data into suitable format. Final step is applying the data mining techniques and extract the valuable information. This process is iterative process it consists of selecting data, pre-processing the data, conversion of data and interpretation or evaluation of data

II. PROCESS OF DATA MINING

It is a process of discovering interesting patterns and knowledge from huge amounts of data. The data sources include databases, data warehouses, web and other information resources .Data mining process includes the following steps.

1) Data Cleaning – this one remove the noise and inconsistent data.

2) Data Integration – It is used to combine multiple data sources.

3) Data Selection –It is used to retrieve the relevant data from the database for analysis task.

4) Data Transformation – It is used to transformed or consolidated data into particular appropriate form for mining by performing summary or aggregation operations.

5) Data Mining – Here the intelligent methods are applied in order to extract data patterns.
6) Pattern Evaluation – It is used to evaluate the data patterns.
7) Knowledge Presentation – Here the knowledge is represented

III. DATA MINING TECHNIQUES

Data mining is the process of extracting specific information from data and present related usable information .one of the most important task in data mining is to select the correct data mining technique. It is selected based on the type of your needs and type of problem of your faces. A Generalized approach is used to improve the accuracy and cost effectiveness of using data mining techniques. In this paper discuss some of the data mining techniques.

A. Association

It is one of technique in data mining .Sometime association techniques know as relation technique, because a pattern is discovered based on relationship between the sources. Example market basket analyses. This method is used to analyse customers mind set of buying, in this a set of products is frequently purchased by the customers. Market Retailers use this technique to analyse customer mind sets.

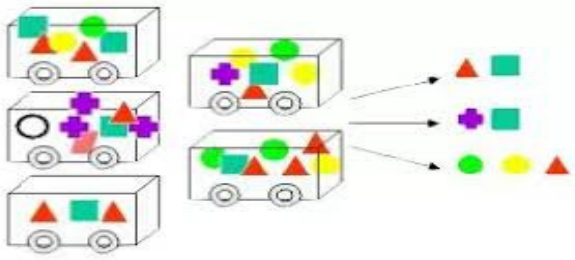


Figure.1.Associate rule in data mining

Our goal is to find all Rules Square -> Triangle specified *minimum* **support** and **confidence** constra ints, given a set of transactions, each of which is a set of items. For Example: Get a data set of all your past purchase from your Departmental Store and found a dependency rule minimizing with respect to the constraints between these items.

{square } —> {triangle }.

$$\text{SUPPORT} = \frac{\text{No.of transaction contains square \& triangle}}{\text{Total number of Transaction}}$$

$$\text{CONFIDENCE} = \frac{\text{No.of transaction contains square \&triangle}}{\text{No.of transaction contains square}}$$

Using this association rules analyzed vast and lots of value to different industries and verical business.Exapmle places are using association cross-selling and up-selling products,Network analysis areas,Physical organization of items,management and marketing,medical health care profession.

B. Classification

It can be performed on structured and unstructured data. It is the process to analyse a given data set and assign a class label to unknown class label. The goal of classification is to identify the class to which a new data will fall under. Data classification done in two steps. In first step is learning step, Training the data set where the model is created by applying a classification algorithm. Second step is the extracted model is test against predefined data set, to measure the trained model performance and accuracy. The classification task like as a supervised technique..

Classification Techniques

In this techniques used to find group of membership of data instantly. There are number if classification methods are available including decision tree induction, Bayesian network-Nearest

neighbor classifier, Case-Based reasoning genetic algorithm and fuzzy techniques [5]

i) Decision Tree induction

A decision tree is a structure it include a root node, branches, leaf nodes. Each internal node denotes attribute, each branch denote outcome of test, each leaf node holds class label. Decision tree are classify objects by sorting them based on their value.The objects are classified from the root node.

Classification and regression models are built by decision trees. It is used to create data models for decision making process. Models are built from training data set.This algorithm work with discrete and continuous variables. It separate data set into subsets based on most significant attribute in the data set. In the decision tree the data set is divided into homogeneous and non overlapping regions. It follows a top-down approach at the top-region all at single region it may be split into two or more branches this method is called greedy approach it only consider the current node between the worked. The decision tree algorithm .will work continuously until to reaching a minimum number of observations. After a decision tree s built many nodes represent outliers data tree pruning method is used to removes the unwanted data, it improves the accuracy of this model. Some of the decision tree algorithms include Hunt's lgorithm,ID3,CD4.5,CART.[1] Decision tree is more complex representation for some concept because of replication problem, Using an algorithm to implement complex features at nodes in order to avoid replication. The FICUS construction algorithm presented by Marko itch and Rosenstein (2002) which receives the standard input of supervised learning as a feature representation specification. This algorithm was designed to perform feature generation given any feature representation specification complying with its general purpose grammar. The well-know algorithm in the literature for building decision trees is the C4.5 (Quinlan, 1993). Quinlan's is earlier ID3 algorithm it was extended as C4.5.The latest study of decision trees and other learning algorithms has been done by (Tjen-Sien Lim et al. 2000). This study shows that C4.5 has a very good combination of error rate and speed compare to others. Ruggeri (2001) presented an analytic evaluation of C4.5 algorithm to know the run time behavior, which highlighted some efficiency improvements. Based on this evaluation, he implemented a more efficient version of the algorithm, called EC4.5. Finally he implemented decision trees as C4.5 with a performance gain of up to five times.

Finally one of the most useful characteristic of decision tree is comprehensibility. Easily understand a decision tree classifies belonging to a specific class. It constitutes a hierarchy of tests. Decision tree when dealing with discrete or categorical features it performs better.

ii) Bayesian Networks

It is one of the classification analysi. Bayes classifiers predict the probability of a given tuple if it is belong to a particular class or not. It is based on the Bayes theorem. bayes theorem is the basis of Bayesian statistics, It enables the user to update the probabilistic of unobserved events. Bayes theorem finds the prior or posterior density of parameters of a given data. It combines information about the parameters from prior density with the observed data [5]. He denotes both the case of discrete probability distributions of data and the more complicated case of continuous probability distributions. At the time discrete case, Bayes' theorem relates the conditional and marginal probabilities of events A and B, provided that the probability of B not equal zero:

$$P(x_i,\ldots,x_n) = \% P(x_i;pa_i)$$

It Contain two parts, a qualitative one based on a DAG for indicating the dependencies, and a quantitative one based on local probability distributions for specifying the probabilistic relationships. The DAG consists of nodes and directed links: Nodes represent variables of interest Even though Bayesian networks can handle continuous variables, exclusively discuss Bayesian networks with discrete nodes. Once fully specified,

a Bayesian network compactly represents the joint probability distribution (JPD) and, thus, can be used for computing the posterior probabilities of any subset of variables given evidence about any other subset.Using the relationships specified by our Bayesian network, representation of the joint probability distribution by taking advantage of conditional independence.

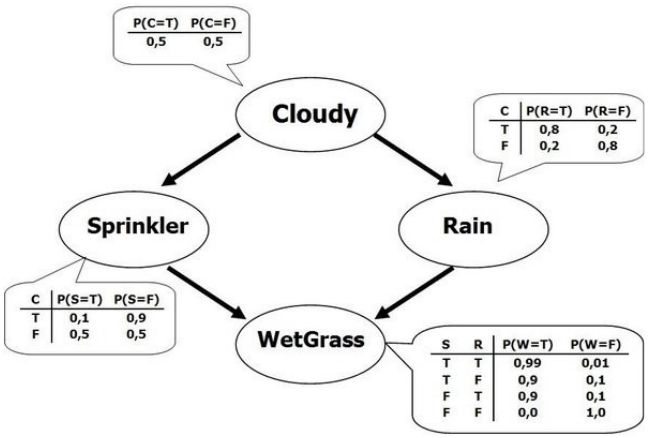
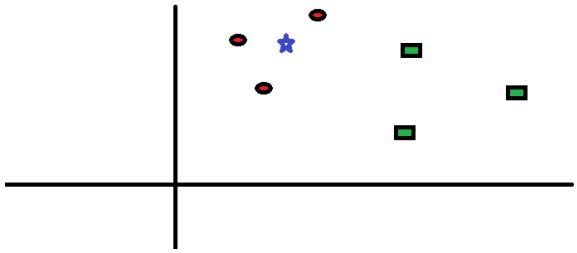


Figure.2.Bayesian Network in Data mining

It is a directed acyclic graph in which each edge corresponds to a conditional dependency, and each node corresponds to a unique random variable, if an edge (A, B) A and B are graph connecting random variables, it means that P (B|A) is a factor in the joint probability distribution, so we must know P (B|A) for all values of B and A in order to conduct inference. In the above example, since Rain has an edge going into Wet Grass, it means that P (Wet Grass | Rain) will be a factor, whose probability values are specified next to the Wet Grass node in a conditional probability table.

iii) K-Nearest Neighbor Classifiers

It is one of the data classification algorithm that attempts to determine what group a data point is in by looking at the data points around it. This algorithm, looking at one point on a grid, trying to determine if a point is in group A or B, looks at the states of the points that are near it. The range is determined arbitrarily, but the point is to take a sample of the data. If the most of the points are in group A, then it is likely that the data point in question will be A rather than B, and vice versa. For example following is a distribution of red circles and green squares.



.

Figure.3. Applying K-Nearest Neighbor Classifiers

You need to find out the class of the blue star. Blue star may be Red circle or Green squares and nothing else. Let's say K = 3. Now we will make a circle with Blue star as center just as big as to enclose only three data points on the plane. Refer to following diagram for more details. The three closest points to blue star is all red circle. We can say Confidently that the Blue Star should belong to the class Red circle.The choice is became very obvious as all three votes from the closest neighbor went to Red circle. The parameter K is play an important role in this algorithm. This method is used for all the training samples are stored in an n-dimensional pattern space. When given an unknown sample, a k-nearest neighbor classifier searches the pattern space for the k training samples that are closest to the unknown sample.We can implement a KNN model by following the below steps:

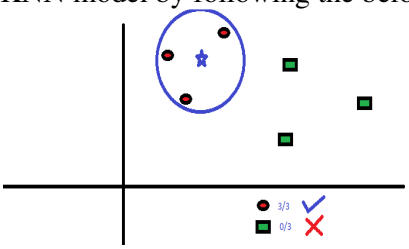


Figure.4 K-Nearest Neighbor Classifiers

1. Load the data.
2. Initialize the value of k.

3. To get the predicted class, iterate from 1 to total number of training data points.
   1. Distance is calculated between test data and each row of training data. Using Euclidean distance as our distance.metric since it's the most popular method. Chebyshev, cosine are the other metrics that can be used.
   2. Calculated distance sorted in ascending order based on distance values
   3. From the sorted array get top k rows.
   4. Finally get the most frequent class of these rows.
   5. Return the predicted class

C. Clustering

Clustering is the task of grouping or dividing objects based on their similarities and chacteristics. Clustering can be classified into two groups First one hard clustering, in this each data point either belongs to a cluster completely or not. Second one Soft Clustering. In soft clustering, instead of putting each data point into a separate cluster, a probability or likelihood of that data point to be in those clusters is assigned.A group of abstract objects into similar objects of one class .The main advantage clustering is ,it is adoptable to changes and helps ingle out useful features that different from different groups. Clustering provides Scalability, deal with different kinds of attributes, high dimensionality, ability to deal with noisy data, interpretability these are the points are tell why clustering is required in data mining.

Clustering methods

This methods can be classified into the following categories.

i. Partitioning method
ii. Hierarchical method
iii. Density-based method
iv. Grid Based method
v. Model based method
vi. Constraint based method

i. Partitioning method

The cluster analysis first partition the set of data into groups, based on the similarity and then assign the labels to the groups. Suppose if we are given a database of 'n' objects and the partitioning methods constructs 'k' partition of data. Each partition will represent cluster 'k <= n', but compulsorily each group contains at least one object, and also that object belongs any one of the group.

ii. Hierarchical method

This method creates a hierarchical decomposition of the given set of data objects. It has two approaches first one is bottom-up approach. Second one is Top-down approach. In the bottom-up approach each object forming a separate group. It keeps merging the objects or groups are close to one another. It doing this work until the termination condition holds. In Top-down approach we put all of the objects into same cluster. It is down until each object in one cluster or the termination condition holds.

iii. Density-based method

This method based on the notion of density, it continues growing the given cluster that is exceeding as long as the density of neighborhood threshold. Each data point in given cluster the radius of a given cluster has to contain at least number of points.

iv .Grid Based method

The objects together form a grid, the object pace is quantized into a finite number of cells that form a grid structure.

v. Model based method

This model is hypothesized for each cluster and it locates the clusters by clustering the density function. It reflects the spatial distribution of data points. This method provides way to determine the number of clusters based on statistics taking outlier or noise into account.

vii. Constraint based method

It is performed incorporation of a user or application oriented constraints. The constraints refer to the user expectations. It provides an interactive way to communication with the cluster process. It can be specified by user needs.

D. Regression

Regression technique adopted for predication. Regression analysis used to model the relationship.

between one or more independent variables. In data mining independent variables are attributes already known and response variables what we want to predict. Many real world problems are not simply predicted example sales volumes, stock prices, product failure. So we need more complex techniques to forecast future values. Regression used to predict a range of numeric values given a particular dataset. For example, regression used to predict the cost of a product or service, given other variables. It I used across multiple industries for business and market planning, financial forecasting, environmental modelling and analysis. There are two types of regression are available that are Linear regression, Multiple regression

i. Linear Regression

It is a simplest form of regression. It attempts to model the relationship between two variables by linear equation based on the observation of data. It also attempts to find the mathematical relationship between variables, if outcome is straight line it is linear model otherwise it is nonlinear model. Relationship between dependent variable is straight line and it has only one independent variable.

$$\boxed{Y=\alpha+\beta X}$$

Y is a linear function of X.

Y increase or decrease X also changed.

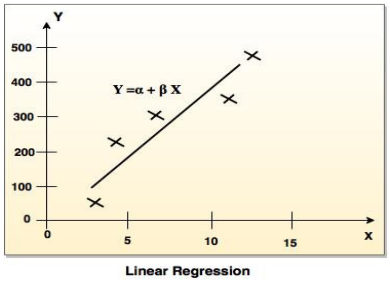


Figure.5.liner Regression

ii .Multiple Regression

Multiple Regression uses two or more independent variables to predict an outcome and a single continuous dependent variable. $\mathbf{Y=a_0 + a_1X_1 + a_2X_2 +\ldots\ldots\ldots.+ a_k X_k + e}$

Where "Y" - response variable,X_1 ,X_2,X_k – independent predictors'"-random ,A0,a1,a2,ak regression co-efficient.

IV. DATA MINING APPLICATIONS

S .NO	APPLICATIONS	USAGE
1	Communications	Data mining techniques are used in communication sector to predict customer activity to provide offer highly targeted and relevant campaigns
2	Education	It provides benefits to educators allow accessing the student data, predicting achievement levels and finding students or groups of students which need extra attention. For example, students who are weak in English subject
3	Manufacturing	Data mining used in base-level designing to obtain the relationships between product architecture, product portfolio, and data

		needs of the customers. It is used to forecast the product development period, cost, and expectations among the other tasks.
4	Banking	Data mining helps the finance sector to get an idea about market risks and manage regulatory compliance.
5	Retail	It helps retail shops and grocery stores identify and arrange most salable items in the most attentive positions. It helps store owners to comes up with the offer which encourages customers to increase their spending
5	Retail	It helps retail shops and grocery stores identify and arrange most saleable items in the most attentive positions. It helps Retailers to comes up with the offer which encourages customers to increase their spending
6	E-Commerce	Websites use Data Mining to offer cross-sells and up-sells through their websites.
7	Crime Investigation	Crime investigation agencies use Data Mining to deploy police workforce (where is a crime most likely to happen and when?), who to search at a border crossing etc.
8	Bioinformatics	Data Mining helps to mine biological data from large data sets gathered in biology and medicine.
9	Service Providers	Service providers like utility industries,are use Data Mining to predict the reasons when a customer leaves their company.
10	Insurance	Data mining helps insurance companies to sales their products into profitable and promote new offers to their new or existing customers.
11	Health Care	Data mining in healthcare has excellent potential to improve the health system. The data and analytics for better insights and to identify best practices that will enhance health care services and reduce costs. Analysts use data mining approaches such as Machine learning, Multi-dimensional database, Data visualization, Soft computing, and statistics.
12	Fraud Detection	Supervised methods consist of a collection of sample records, and these records are classified as fraudulent or non-fraudulent.

V.CHALLENGES OF IMPLEMENTING DATA MINING

Data mining is very powerful but it faces many challenges during its execution. Various challenges could be related to performance, data, methods and techniques.

A. Complex Data

Normally the data is heterogeneous, and it could be multimedia data, including audio and video, images, complex data, spatial data, time series, and so on. Handling these various types of data and extracting useful information is a tough task. Most of the time, new technologies, new tools, and methodologies would have to be refined to obtain specific information.

Figure.6. Challenges in Data Mining

B. Performance

The data mining system's performance relies primarily on the efficiency of algorithms and techniques used. If the designed algorithm and techniques are not up to the mark, then the efficiency of the data mining process will be affected adversely.

C. Data Visualization

In data mining, data visualization is a very important process because it is the primary method that shows the output to the user in a presentable way. The extracted data should convey the exact meaning of what it intends to express. But many times, representing the information to the end-user in a precise and easy way is difficult. The input data and the output information being complicated, very efficient, and successful.

D. Data Privacy and Security

Data mining usually leads to serious issues in terms of data security, governance, and privacy. For example, if a seller analyzes the details of the purchased items, then it reveals data about buying habits and preferences of the customers without their permission.

E. Incomplete and noisy data

When extracting useful data from large volumes of data, the data is heterogeneous, incomplete and noisy. Quantities usually inaccurate or unreliable**.**

F. Data Distribution

Actually data is usually stored on various platforms in a distributed computing environment. It may be in a database,personal systems, or even on the internet. Normally , it is a quite difficult task to make all the data to a centralized data repository mainly due to organizational and technical concerns.

VI.CONCLUSION

Data Mining helps to extract information from huge sets of information. It's the procedure of mining knowledge from data. Data mining is all about explaining the past and predicting the longer term for analysis. Data mining is the process it includes business understanding, data understanding, data preparation, modelling, and evolution. It provides various techniques association, classification, regression clustering to use. This mining is used in different industries such as insurance, education, banking, communications, Bioinformatics, manufacturing, retail, service providers, e-commerce. The noticeable drawback of data mining is that many analytics software is difficult to operate and requires advance training to work on.

REFERENCES

[1]. Thair Nu Phyu; Survey Of Classification Techniques In Data Mining. Phyuproceedings Of The International Multiconference Of Engineers And Computer Scientists 2009 Vol I Imecs 2009, March 18 - 20, 2009, Hong Kong

[2]. D.Usha Rani ; A Survey On Data Mining Tools And Techniques In Medical Field, International Journal Of Advanced Networking & Applications (Ijana) Volume: 08, Issue: 05 Pages: 51-54 (2017) Special Issue.

[3]. Rakhi Ray ; Advances In Data Mining: Healthcare Applications, International Research Journal Of Engineering And Technology (Irjet) E-Issn: 2395-0056 Volume: 05 Issue: 03 | Mar-2018 Www.Irjet.Net P-Issn: 2395-0072

[4]. M. Durairaj, V. Ranjani; Data Mining Applications In Healthcare Sector: A Study , International Journal Of Scientific & Technology Research Volume 2, Issue 10, October 2013 Issn 2277-8616.

[5].Https://Www.Bayesia.Com/Bayesian-Networks-Introduction

[6].Bouckaert, R. (2004), Naive Bayes Classifiers That Perform Well With Continuous Variables, Lecture Notes In Computer Science, Volume 3339, Pages 1089 – 1094.

[7].Cheng, J. & Greiner, R. (2001). Learning Bayesian Belief Network Classifiers: Algorithms And System, In Stroulia, E.& Matwin, S. (Ed.), *Ai 2001*, 141-151, Lnai 2056.

[8].D.Shobana, N.Uthra," The Data Mining Concepts And Techniques: A Survey" International Journal Of Trend In Research And Development, Volume 2(6),Issn 2394-9333.

[9]. Introduction To Data Mining And Knowledge Discovery, Third Edition Isbn: 1-892095-02-5, Two Crows Corporation, 10500 Falls Road, Potomac, Md 20854 (U.S.A.), 1999.

[10].Arun K Punjari, ―Data Mining Techniques ‖ , Universities (India) Press Private Limited, 2006.

[11]. Han, J. & M. Kamber, And Data Mining: Concepts And Techniques, San Francisco: Morgan Kaufma(2001).

[12].Https://Www.Analyticsvidhya.Com/Blog/2018/03/Introduction-K-Neighbours-Algorithm-Clustering.

