Data Mining using KNN

Sandeep Kaur[1]

[1]*Research scholar, Computer Science and Engineering department,*
GGS College of Modern Technology, Kharar Punjab Technical University, Jalandhar,India

Abstract- The recommendation framework assumes an essential part in web use mining. Data mining is a procedure in which vital or important qualities are extracted from database. Data warehouse is a sort of capacity of information where the information is kept to fetch it in close future when it is necessitated. Distinctive kinds of information can be productively stored in a Data Warehouse. This sort of information that will be stored is controlled by the kind of association which devours it. A large portion of the ventures keep record of every last sort of information while a few organizations just store that data which is valuable and significant for them. The information put away in a warehouse is useful in decision support network. Based on historical information the choice in regards to the future plans can be taken effortlessly or much adequately. This paper offers an overview to the work that has been done many authors.

Keywords- Data Mining, KNN, CART, CHAID.

I. INTRODUCTION

Extracting the important and meaningful information from the data base is called data mining. Data ware house stores the data so that it can be fetched whenever it is required [1,3]. Data Warehouse can store any kind of data particularly the type of data depends upon the kind of industry for which it is being used. Most of the industries keep record of each and every kind of data whereas some companies only store that information which is beneficial and meaningful for them [2, 5]. The data stored in a warehouse is helpful in decision support system. On the basis of historical data the decision regarding the future schemes can be taken easily or much effectively.

The example of a banking organization can effectively define the purpose of the data warehouse for an organization. For banking organization income is a very important source for describing the socio-economic situation [4,6,7]. The strategies for the benefits of the customer can only be created if and only if the bank has any record regarding the income of its customers. If the bank has a record regarding the income of the customers then they can offer the various schemes like concession on interest rates on various loans [8,9]. Bank can also alter the limit of credit card issued to the customer. Bank can know the income of the customer only in two ways such as amount deposited by the customer in his account on the daily basis and second way is to access the purchasing habits of the customer by using credit card issued to him [10].

There is another example which shows the failure of the institution of insurance products to access the data warehouse in a proper manner [12, 13]. By having the access to the information of transaction the institution can have information regarding whether the payments done by the customer to any other broker or not. In this way they can take decision regarding their prospectus [11, 15]. All of this can be achieved only if the data is stored in data warehouses and data mining is performed to extract valid data.

Basically data mining is used to examine the huge data set. The examined data belongs to a pattern. Large data set called as big data [14]. The big data pattern contains the method at the interaction of Artificial Intelligence and database systems or statistics etc. The main motive of the data mining is to mine those part of the data warehouse which meaningful and helpful in different decision making processes. Sometimes the stored data in warehouse is not meaningful or say important for every purpose. It is knowledge getting process. Data mining also performs online updating, complexity consideration, pre-processing, visualization, data management. It search the large amount of data, performs the pattern matching. Data mining also becomes a term which is repeatedly use at the place of large scale data processing or information processing and where the data processing includes the processes like collection of data, extraction of information, analysis of data warehouse etc. it is also substituted with the process decision making systems, AI, machine learning etc. the jargons used for data mining like data analysis on large scale database or data analytics or machine learning and AI is much appropriate terms.

II. DATA MINING

2.1 Issues in Data Mining

Data mining is the most common issue so it is a most common area of research. Theoretical issue is the one of the data mining issue. Related to practical implementation of the mining such as examination of interesting and unknown knowledge set from actual world data bases. There are some issues associated to the data mining with their algorithms and solutions:

1. Mixed changing and redundant data

2. Over decent and assessing the statistical significance.
3. Understanding the patterns.
4. Massive datasets and high dimensionality.

2.2 Tasks of Data Mining

It is the process which extracts the useful information from the large amount of data. There are some following tasks which perform by the data mining:

Classification
Explanation and visualization
Affinity grouping or association rules

- Prediction
- Clustering
- Estimation

Classification, Estimation and Prediction is performed for directed data mining. Directed Data Mining is a term which describes the process when the given data in database is used to create patterns or model that defines the single or multiple meaningful attributes as compared with other attributes. Association, Clustering and Description define the process of undirected data mining. Undirected data mining is the process in which the relationship between attributes is developed.

2.2.1 Classification

Classification is a process which is performed to evaluate the characteristics of given object and then these properties are allotted to the existing objects. Classification is done by using classes along with a training set of which includes reclassified objects or examples. The main aim of the classification is to classify the unclassified patterns or data. Examples of classification are as follows:

Classification of credit applicants on the basis of low, high or medium risk.
Classification of fruits and vegetables as dibble or hazardous.
To determine the telephone lines connected to internet.

2.2.2 Estimation

Estimation is the second process or task to perform after classification. In this the output is estimated on the basis of input pattern. The estimation is performed on the basis of given input parameters and these variables or parameters are unknown. Example of estimation process is as follows:

From the mother's qualification as an input variable estimating the number of children in a family.
On the basis of number of vehicles in a house estimating the total income of that house.

2.2.3 Prediction

It can be measured as estimation of process and classification of process. It is the process in which the outcome is predicted on the basis of some historical behavior or future values. The example of prediction process is as follows:

On the basis of buying behavior of the customer to predict whether he will leave in near future or not.
On the basis of user's behavior and activated plans predicting that which customer would like to have value added services on his connection.

2.2.4 Association Rules

Association rule is used to describe the relationship between set of objects. It defines that how various objects are linked or associated to each other with in a database. To understand the association rule let's consider an example, in a given set of transactions each and every transaction consist of items and X and Y are two data items. The relationship among them can be described as the database contained a X tends to contain Y.

2.2.5 Clustering

Clustering is a technique in which the data is divided into sub classes. These sub classes are known as clusters. These clusters contain the related data sets only. This leads to select the data easily when needed. Hence the data is divided into categorical clusters according to the nature of the data sets. There are many algorithms which are used to divide the data into various clusters. Some of the algorithms used for clustering are partitioning method, hierarchy method, grid based method of clustering. Clustering does not work upon variables that are predefined where as classification is based on predefined values. This is the main difference between the two.

III. DATA MINING TECHNIQUES

There are various techniques used for data mining. Some of them are explained as follows:
Methods for mining transactional or relational database.
Artificial Intelligence technique for data mining.
Decision Tree approach.
GA i.e. Genetic Algorithm.
Visualization.

3.1 Statistics

Statistics is considered to as an important part of data mining that is basically used for extracting knowledge and selecting data. For removing irrelevant data from meaningful data, statistics is used. It is mostly used for data cleaning as it helps in removing irrelevant data from data sets. By data cleaning, we mean process of removing irrelevant items and noise from data. Missing data can also be recovered with the help of statics. Various designing techniques like clustering and experimenting are used for data analysis.

3.2 Techniques for mining transactional/ relational database

In case of relational and transactional database, data mining based on links between items in a data set are performed. Pattern used in deriving association between data sets is where Ai for i {1,….., m} and Bj for j {1, ….., n} are value sets based on its attributes. Example sometimes customer buys chips or any other snacks along with beer. So data in a relational data is mined based on its relationship or association.

3.3 Artificial intelligence (AI) techniques

The concept of Artificial Intelligence (AI) is mostly used in data mining. By artificial intelligence we mean techniques like neural network, machine learning etc. many other techniques like knowledge acquisition, knowledge presentation and pattern recognition etc. also comes under Artificial Intelligence that are executed with the help of data mining. Classification is considered to be as most lacking point of data mining. By classification, it mean classifying data based on mutual exclusion [20]. The word mutual exclusion means how close are member of same groups to each other and how far they are from other group members. Example based on credit worthiness of a customer, its database is classified. To remove these classification's problems, neural network is used. Basic three steps are followed for neural network in data mining that are given below:
Network Construction and Training: Parameters like coding methods, number of attributes and number of classes etc. are used in this step for construction of neural network.
Network Pruning: In this step the links and data units that are repeated in data sets are removed from it without disturbing the error rate in that network. Or it can be said that this step helps in maintaining consistency.
Rule Extraction: This step helps in withdrawing the rules of classification.
Many other techniques like case based reasoning and intelligent agents and are also part of artificial intelligence. The patterns such formed are identified with the help of historical data. Patterns formed throughout the data are matched with the help of computer based program known as agent used by artificial intelligent agent.

3.4 Decision tree approach

Decision tree method works on the basis of tree structure. The decisions are presented in the form of tree structure. Decision tree structure is useful for decision support systems. Some mostly used decision trees are as follows:
CART
CHAID
The word CART refers to Classification and Regression Tree and the word CHAID refers to Chi Square Automatic Interaction Detection. Both of the techniques mentioned above are used for data classification. Both of the techniques are implied on unclassified data to convert it into the classified data after applying classification and predicting the outcome on the basis of classified data. These techniques need some data preparations but CART needs to have less data preparation as compare to CHAID.

3.5 Genetic algorithm

The basic motive for development of genetic algorithm is Darwin's theory of evolution. A population of rules is utilized to discover the solution of specific problems. The population rule is made at first and randomly.

3.6 Visualization

The word Visualization defines symbolic representation or a better understanding of any technique, problem or

system. Visualization plays an important role in data mining also. Data mining by using visualization can be achieved with the help of human brain or help. There is large number of examples available on the basis of visualization techniques which results in interactive pictorial representations. Some techniques used in it are:
Scatter Plot matrices,
Coplots projection.
Project matrix,
The techniques of Visualization used for geometric data sets are as follows:
Graph based techniques
Icon Based Techniques,
Dynamic Techniques,
Hierarchical Techniques,

IV. RELATED WORK

W. Lei, L. Chong et al., in this paper the creator had outlined the present database and computerized reasoning were hotly debated issues in the investigation of information mining innovation, this paper remains in the premise of idea and key innovation of distributed computing, Web information mining innovation of distributed computing techniques, the strategy to understand the mining innovation.

P.Bhargavi, in this paper the creator had anticipated a few ideal models utilized in information mining in the field of soil science database with a specific end goal to keep up a relationship. For the trial examination, soil information base has been extricated from the Department of soil sciences and agrarian science, Tirupati. The essential research was to assess whether the information mining methods were skilled in arranging the dirts or not. Additionally, the correlation has made between various order like Naïve Bayes and compelling instruments. There were a few advantages of the examination work had additionally recognized in this paper like Agriculture, Soil administration and natural.

Dhanashree S. medhekar, in this paper the creator had anticipated a classifier approach for the location of coronary illness. Besides, paper additionally clarified how Naïve Bayes calculation can be used with the end goal of arrangement. The average information taken for assessment has been sorted into five segments, for example, no, low, normal, high and high. On the off chance that an obscure example was recognized then framework will anticipate the class name of the specific example. A few capacities has performed named as preparing and testing i.e. Characterization and Prediction separately. Exactness of framework relies on the utilized calculation and database.

Zhang Zi-qiong, Ye Qiang et al., in this paper the creator had delineated that the audits were singular reports exhibiting sentiments or estimations. Despite what might be expected, non-audit reports frequently speak to genuine information unbiasedly. Estranging audits from non-surveys, or subjectivity order, was possibly huge for a few content technique applications, similar to information coercion and information recuperation. Additionally, it was a noteworthy method in assessment arrangement for online client audits. As a sort of type classification, the orders of emotional and target writings were differing from regular subject based orders. A few examinations had not been performed in this space and the vast majority of them were on English writings. Little work has been done on Chinese subjectivity classification. Therefore, the clarified instruments were used in English writings can't be used straight to Chinese in light of the differing highlights inside these several dialects. To offer the subjectivity arrangement on Chinese content based on a regulated machine learning worldview, Naive Bayes a system was offered in this paper. The reproduction results had been performed on two or three sorts of reports: film audits and motion picture plots written in Chinese. The recreation results had shown that the introduction of the anticipated system were similar to those of the customary English subjectivity order contemplates.

Parneet Kaur, Manpreet Singh et al., in this paper the creator had outlined the Educational Data Mining field focus on Prediction all the more regularly similar to deliver amend results for future reason. On the adjustments happened in educational programs designs keeping in mind the end goal to check an ordinary examination was mandatory of instructive databases. Keeping in mind the end goal to perceive the moderate students alongside understudies and demonstrating it through a prescient information mining model this examination had focused on it by applying classification based standards. Real World informational index from a secondary school was gotten and filtration of favored potential factors was readied using WEKA an Open Source Tool. The dataset of understudy scholastic records was confirmed and used on a few arrangement ideal models like Multilayer Perception, Naïve Bayes, SMO, J48 and REPTree using WEKA an Open source apparatus. Subsequently, measurements were created based on whole order standards and examination of entire five classifiers was likewise finished to explore the accuracy and to look through the best performing characterization worldview alongside whole. In this work, a learning stream show was likewise delineated alongside add up to five classifiers. This investigation had shown the essentialness of Prediction and Classification based information mining standards in the territory of training and furthermore spoke to different promising future lines.

Omer Faruk Arar, KürşatAyan, in this paper the creator had represented that the Naive Bayes was one of the significantly used ideal models in order issues because of the simplicity, productivity, and force. It was fitting for a few learning situations, similar to picture classification, extortion examination, web mining, and content order. Credulous Bayes was conceivable technique based on the suppositions that qualities were autonomous of each other and that their weights were just huge. In this manner, really, qualities may be interrelated. All things considered, such presumptions may cause an emotional decrease in introduction. In this work, through after preprocessing steps, a Feature Dependent Naive Bayes (FDNB) order system was anticipated. Attributes were included for assessment assets to create reliance inside each other. This system was used to the product imperfection examination issue and practices were done using extensively recognized NASA PROMISE informational collections. The accomplished outcomes delineated that this novel component was much fruitful relative to the standard Naive Bayes technique and that it had an aggressive introduction alongside other trademark weighting plans. The other objective of this work was to delineate that to be adaptable; a learning model ought to be made through using just preparing information, as generally deceptive results emerge from the use of the whole informational index.

Wa'el Hadi, Qasem A. Al-Radaideh, Samer Alhawari, in this paper the creator had anticipated a component that was named as cross breed AC calculation (HAC). The HAC had connected the intensity of the Naïve Bayes (NB) worldview keeping in mind the end goal to diminish the quantity of order controls and to produce different directions that present each trait esteem. Two or three tests were performed on an Arabic literary dataset and the standard Reuters-21578 datasets by applying six assorted ideal models, particularly J48, NB, grouping in light of affiliations (CBA), multi-class characterization in view of affiliation rules (MCAR), master multi-class order in view of affiliation rules (EMCAR), and quick acquainted arrangement worldview (FACA). The recreation yields of the practices exhibited that the HAC technique produced most extreme order rightness near to the MCAR, CBA, EMCAR, FACA, J48 and NB alongside increases of 3.95%, 6.58%, 3.48%, 1.18%, 5.37% and 8.05% individually. In addition, on Reuters-21578 datasets, the yields delineated that the HAC worldview had a magnificent and unfaltering introduction as far as arrangement precision and F measure.

Liang xiaoJiang, Shasha Wang, Chaoqun Li, Lungan Zhang, in this paper the creator had represented that the Bayesian framework could exhibit discretionary quality conditions, taking in an ideal Bayesian framework from high-dimensional content information was around impractical. The real reason was that taking in the ideal structure of a Bayesian framework from high-dimensional content information was massively time and space use. In this work, the creator had anticipated another model named structure broadened multinomial credulous Bayes (SEMNB). To learn SEMNB, a simple however effective learning worldview was anticipated in this paper with no structure looking. The recreation results on an extensive suite of benchmark content datasets outlined that SEMNB adequately outflanks MNB and was even extraordinarily improved near to other three best in class upgraded standards including TDM, DWMNB and CMNB.

B.V Baiyju, R.J.Remy, In this paper the creator delineated that the gigantic measure of information was anticipated in medicinal associations (doctor's facilities, restorative focuses) yet as this information was not appropriately used. There is an abundance of concealed information speak to in the datasets. This unused information can be modified into supportive information. A classifier plot for examination of coronary illness is anticipated in this work and showed that how Naive Bayes can be connected for order reason. In this instrument, the medicinal information is arranged into five classifications named as no, low, normal, high and high. Likewise, if obscure example comes then the system will explore the class name of that example. Thusly two or three fundamental capacities called as arrangement (preparing) and forecast (testing) would be introduced. Rightness of the system depends on the worldview and database used.

Vellingiri, J., S. Kaliraj, S. Satheeshkumar and T. Parthiban, In this paper the creator had offered an investigation of the programmed arrangement of web client route designs and anticipated another component to order client route examples and examining clients' future solicitations. The system was happened based on the consolidated mining of Web server logs and the substance of the recouped site pages. The literary substance of site pages was caught by blackmail of character N-grams that were converged among Web server log documents to begin client route profiles. The instrument was executed as an exploratory technique, and its introduction was figured based on a few employments: arrangement and expectation. The component got the arrangement rightness of about 70% and the examination accuracy around 65% that was roughly 20% more prominent near to the order accuracy through mining Web server logs alone. This system may be used to encourage enhanced web personalization and site association.

V. CONCLUSION

In this paper the author had surveyed that while web surfing useful data is derived from secondary data with the help of web usage mining. It is a technique which could guess user behavior during its interaction with the web. The aim of each domain is: to predict user’s behavior on the site, difference between actual and expected web site usage, adjustment of the Web site to the interests of its users. No exact distinction between the other categories and web usage mining. For the preparing the data of web usage mining, content data used as useful sources, that interacts with structured mining and web content mining. For the pattern discovery clustering is used, which associate to web content and structure is mining from usage mining.

VI. REFERENCES

[1] Wang Lei, Liu Chong, “Implementation and Application of Web Data Mining Based on Cloud Computing”, IEEE, International Conference on Intelligent Transportation, Big Data and Smart City, 2016.

[2] P.Bhargavi, “Applying Naïve Bayes data mining technique for classification of agriculture land soils”, IJCSNS, Vol. 9, No. 8, pp. 117-122, 2009.

[3] Dhanashree S. medhekar, “Heart Disease prediction System using Naïve Bayes”, International Journal of Enhanced Research in Science Technology & Engineering, Vol. 2, No. 3, pp. 1-5, 2013.

[4] Zhang Zi-qiong ; Ye Qiang ; Li Yi-jun, “Using Naïve Bayes Classifier to Distinguish Reviews from Non-review Documents in Chinese”, IEEE, International Conference on Management Science and Engineering, 2008.

[5] Parneet Kaur, Manpreet Singh, Gurpreet Singh Josanc, “Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector”, ELSEVIER, Procedia of Computer Science, Vol. 57, pp. 500-508, 2015.

[6] Ömer Faruk Arar, KürşatAyan, “A feature dependent Naive Bayes approach and its application to the software defect prediction problem”, ELSEVIER, Applied Soft Computing,Vol. 59, pp. 197-209, 2017.

[7] Wa'el Hadi, Qasem A. Al-Radaideh, Samer Alhawari, “Integrating associative rule-based classification with Naïve Bayes for text classification”, ELSEVIER, Applied Soft Computing,Vol. 69, pp. 344-356, 2017.

[8] Liang xiaoJiang, Shasha Wang, Chaoqun Li, Lungan Zhang, “Structure extended multinomial naive Bayes”, ELSEVIER, Information Science, Vol. 329, pp. 346-356, 2016.

[9] B.V Baiyju, R.J.Remy, “A Survey on Heart Disease Diagnosis and Prediction using Naive Bayes in Data Mining”, IJCET, International Journal of Current Engineering and Technology, Vol. 5, No. 2, pp. 1-5, 2015.

[10] Vellingiri, J., S. Kaliraj, S. Satheeshkumar and T. Parthiban, “A Novel Approach for User Navigation Pattern Discovery and Analysis for Web Usage Mining”, IJCSE, Vol. 5, No. 5, pp. 325-329, 2017.

[11] Jian Ming, Lingling Zhang, Jinhai Sun, Yi Zhang, “Analysis models of technical and economic data of mining enterprises based on big dataanalysis”, ICCCBDA, Pp 224-227, 2018.

[12] Vania Bogorny, Shashi Shekhar, “Spatial and Spatio-temporal Data Mining”, IICDM, Pp 1217-1217, 2010.

[13] Abbas Madraky, Zulaiha Ali Othman, Abdul Razak Hamdan, “Hair data model: A new data model for Spatio-Temporal data mining”, DMO, Pp 18-22, 2012.

[14] Hetal Thakkar, Barzan Mozafari, Carlo Zaniolo, “A Data Stream Mining System”, IICDMW, Pp 987-990, 2008.

[15] Jagat Sesh Challa, Poonam Goyal et al., “DD-Rtree: A dynamic distributed data structure for efficient data distribution among cluster nodes for spatial data mining algorithms”, IEEE, Pp 27-36, 2016.