A Study on Data Mining Technologies and Its Applications: A Survey Paper

Richa, Assistant Professor, Chandigarh University, Gharuan, India,

Abstract- This world is full of "Digital Data". On daily basis, abundance of data has been created by humans, weather it is transactional data, scientific data, environmental data, medical data, financial data, mathematical data, social networking data or educational data. with the increasing age of data on us, the risk of sinking in the flood of digital data is getting bigger. This digital data include unstructured data, structured data and semi structured data, which I am referring as "Combine Data". This research study gives the conceptual view of what data mining is and identifying the different technologies to understand the impact of data mining to convert the digital data into meaningful pattern for real world applications.

Keywords- *Data Mining, Data Mining models , Forecasting models, Applications.*

I. INTRODUCTION

Data mining is the "Power Tool" which is used to extract the useful and meaningful patterns from bulk of data to make the data for better usage. It also helps in analyzing the data by implementing various techniques and technologies ranging from artificial intelligence to machine learning, visualization methods to statistical analysis. Except the hypothetical concept, it is also refer as Knowledge Discovery in Databases known as KDD. With the help of knowledge discovery in databases, it creates the useful pattern to get the useful information. With the help of data mining many scientific fields has advanced their levels in the meaning of research, weather it is about medical field, environmental field or Scientific field. Now with the help of data mining analysis we can even detect the early signs of diseases like cancer. NASA also has been using data mining concept to make their missions more efficient and more safer.

In [1], the following definition given as: "Data mining is the process of exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules".

Data mining is an interdisciplinary subfield of computer science which involves computational process of large dataset to discover patterns. The goal of data analysis with data mining process is to extract information from a data set and transform it into an understandable structure format for further use. The methods used are at the juncture of artificial intelligence, machine learning, statistics, database systems and business intelligence. Data Mining is nothing but about solving problems by analyzing data already present in databases [2]. Intelligent mechanism are applied to extract the Data mining is also stated as essential process where intelligent methods are applied in order to extract the data patterns to help for better decision making in terms of financial, health and detection of cyber crime .

II. DATA MINING PROCESS

To draw the hidden and meaningful patterns for decision making, data mining has basically following major steps:

- Data cleaning and integration is done to remove the noise and inconsistency from raw data and convert data into target data.
- Selection, Extraction and Transformation is done to process the data for mining
- Different tools and techniques are applied by data mining technologies to extract and to evaluate the useful patterns.
- Then Knowledge representation is used to better understanding of mined knowledge or data.

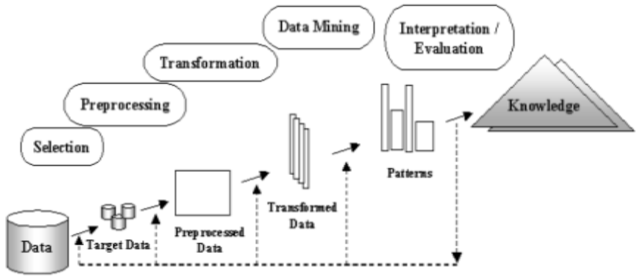


Fig.1:

Fig No. 1 is giving the pictorial representation of data mining process[3]. The main aim of data mining is to extract the information which can be utilized by the investors or traders to make their decision power stronger.

III. DATA MINING TECHNIQUES

Data mining is very viable as it draws upon at least one of these strategies[4]:

1. Tracking patterns- A standout among the most essential procedures in information mining is figuring out how to perceive designs in your informational collections. This is typically an acknowledgment of some deviation in your information occurring at standard interims, or a recurring pattern of a specific variable after some time. For instance, you may see that your offers of a specific item appear to spike just before the occasions, or notice that hotter climate drives more individuals to your site.

2. Classification-Classification is a more perplexing information mining method that powers you to gather different properties together into discernible classes, which you would then be able to use to reach assist determinations, or serve some values. For instance, in case you're assessing information on singular clients' budgetary foundations and buy accounts, you may have the capacity to characterize them as "low," "medium," or "high" credit dangers. You could then

utilize these arrangements to learn significantly more about those clients.

3. Association- Association is identified with following examples, however is more particular to conditionally connected factors. For this situation, you'll search for particular occasions or qualities that are exceptionally associated with another occasion or trait; for instance, you may see that when your clients purchase a particular thing, they additionally frequently purchase a second, related thing. This is normally what's utilized to populate "individuals additionally purchased" segments of online stores.

4. Outlier detection- Much of the time, just perceiving the overall example can't give you an unmistakable comprehension of your informational collection. You additionally should have the capacity to distinguish inconsistencies, or anomalies in your information. For instance, if your buyers are solely male, yet amid one interesting week in July, there's a colossal spike in female buyers, you'll need to examine the spike and see what drove it, so you can either repeat it or better comprehend your gathering of people all the while.

5. Clustering-Clustering is fundamentally the same as classification, yet includes gathering clusters of information together in light of their similarities or likenesses. For instance, you may bunch diverse social economics of your group of onlookers into changed bundles in light of how much extra cash they have, or how regularly they tend to shop at your store.

6. Regression-Regression, utilized basically as a type of arranging and displaying, is utilized to recognize the probability of a specific variable, given the nearness of different factors. For instance, you could utilize it to extend a specific cost, in view of different components like accessibility, customer request. All the more particularly, the main focus are of regression is to enable you to reveal the correct connection between (at least two) factors in a given informational collection.

7. Prediction- Forecasting is a standout among the most profitable information mining procedures, since it's utilized to extend the kinds of information you'll find later on. As a rule, simply perceiving and understanding verifiable patterns is sufficient to diagram a to some degree precise forecast of what will occur later on. For instance, you may audit buyers' records of loan repayment and past buys to anticipate whether they'll be a credit chance later on.

IV. DATA MINING APPLICATIONS

Today everything is related with internet which means internet is nothing but a cluster of information where any kind of data previously described as combined data is available in any form in any shape. As everything is related with data, data mining is there to convert combined data into useful information. Most prominent areas where data mining technologies are heavenly used are

- Financial Sector

Finance means production of huge data set. Financial term includes transactions on daily basis, stock market data, currency exchange rate, online trading etc. Data mining is becoming a strategically important area for many businesses including finance sector. It is a process of analyzing the data from various angles and summarizing it into valuable information. Everything in data Mining is about extracting the insightful information and produce such strategies for future actions that help investors for making their decision better .

A large portion of the organizations employ the information digging accomplice for them and some of them do it all alone. The determined facts help the fund organizations to search for shrouded patterns in a gathering of information and find relationship between them. Information Mining in Finance displays a thorough outline of major algorithmic ways to deal with prescient information mining, including factual, neural systems, ruled-based, choice tree, and fluffy rationale techniques, and after that looks at the appropriateness of these ways to deal with budgetary information mining.

- Health Sector

There is a relentless development in the measure of electronic wellbeing records or EHRs being gathered by healthcare offices. It has been the standard for attendants to take obligation in taking care of patient information input that was traditionally recorded in paper-based structures. Exactness is extremely imperative with regards to quiet mind and computerizing this enormous measure of information improves the quality of the entire framework. Be that as it may, how do social insurance suppliers filter through all the information proficiently? This is the place information mining has proven to be to a great degree viable. Information mining has been used to reveal designs from the extensive measure of put away information and after that used to manufacture prescient models. Since the mid 90s, this training has been utilized to help with misrepresentation discovery, credit scoring and upkeep scheduling yet it's at long last being used in human services programs around the nation. Enhancing the nature of patient consideration and decreasing healthcare costs are the perfect objectives of numerous projects. Information mining has helped these projects succeed.

- Cyber Crime Sector

Information mining applications are used in many saving money areas for customer division and profitability, FICO ratings and approval, foreseeing instalment default, promoting, distinguishing counterfeit exchanges, and so on. This paper introduces a general thought regarding the model of Data Mining strategies and differing digital violations in saving money applications. It additionally gives a comprehensive study of skilled and significant strategies on information digging for digital wrongdoing information investigation. The goal of digital wrongdoing information mining is to perceive designs in criminal conduct with a specific end goal to foresee wrongdoing envision criminal movement and prevent it.

V. METHODOLOGIES OF DATA MINING

A. K-Means

K-Means, Influenced Association Classifier and J48 Prediction tree for exploring the digital wrongdoing informational collections and deals with the available issues.

The K-Means calculation is being used for unsupervised learning group inside affected Association Classification. K-implies chooses the underlying centroids so the classifier can mine the record and figure forecasts of digital wrongdoings with J48 calculation. The aggregate learning of K- Means, Influenced Association Classifier and J48 Prediction tree tends positively to manage the cost of an upgraded, fused, and exact outcome over the digital wrongdoing expectation in the saving money parts Our law implementation associations require to be enough equipped to crush and keep the digital wrongdoing.

B. Neural Network

Neural Network or a counterfeit neural system is a natural framework that recognizes examples and makes forecasts. The best leaps forward in neural system lately are in their application to true issues like client reaction expectation, misrepresentation discovery and so on. Information mining procedures, for example, neural systems can show the connections that exist in information accumulations and can thusly be utilized for expanding business knowledge over an assortment of business applications. This great prescient demonstrating strategy makes extremely complex models that are extremely hard to comprehend by even specialists. Neural Networks are utilized in an assortment of utilizations. It is appeared in Fig No.2. Counterfeit neural system have turned into a great apparatus in assignments like example acknowledgment, choice issue or predication applications. It is one of the most up to date flags handling innovation. ANN is a versatile, non straight framework that figures out how to play out a capacity from information and that versatile stage is typically preparing stage where framework parameter is change amid activities. After the preparation is finished the parameter are settled. On the off chance that there are bunches of information and issue is ineffectively reasonable at that point utilizing ANN demonstrate is precise, the non direct qualities of ANN give it heaps of adaptability to accomplish input yield outline. Counterfeit Neural Networks, give client the capacities to choose the system.

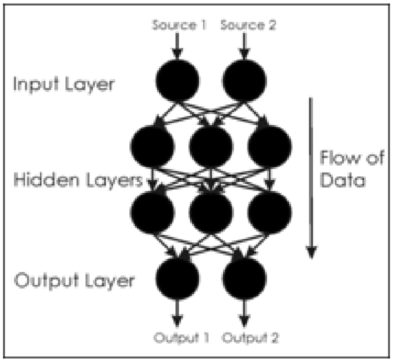


Fig.2:

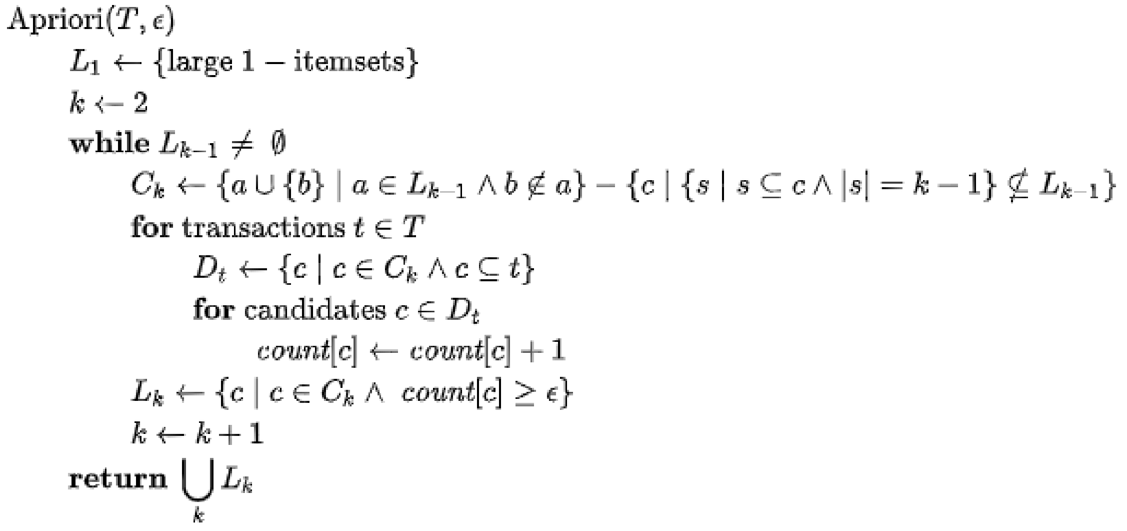
C. Decision Trees

A choice tree is a stream outline like structure where every hub indicates a test on a quality esteem, each branch speaks to a result of the test and tree leaves speak to classes or class circulation. A choice tree is a prescient model regularly utilized for arrangement. Choice trees parcel the info space into cells where every cell has a place with one class. The parceling is spoken to as a grouping of tests. Every inside hub in the choice tree tests the estimation of some info variable, and the branches from the hub are named with the conceivable aftereffects of the test. The leaf hubs speak to the cells and determine the class to return if that leaf hub is come to. The characterization of a particular info occasion is in this manner performed by beginning at the root hub and, contingent upon the aftereffects of the tests, following the proper branches until the point when a leaf hub is come to [5]. Choice tree is a prescient model that can be seen as a tree where each part of the tree is an arrangement question and leaves speak to the parcel of the informational collection with their characterization.

D. Apriori algorithm

Apriori calculation is an established calculation in information mining. It is utilized for mining continuous item_sets and pertinent affiliation rules. It is contrived to work on a database containing a ton of exchanges, for example, things brought by clients in a store. It is critical for viable Market Basket Analysis and it helps the clients in acquiring their things without any difficulty which builds the offers of the business sectors. It has additionally been utilized in the field of medicinal services for the recognition of antagonistic medication responses. It produces affiliation decides that demonstrates what all blends of drugs and patient qualities prompt ADRs.

The pseudo code[6] for the calculation is given beneath for an exchange database , and a help limit of . Regular set theoretic documentation is utilized, however take note of that is a multiset. is the competitor set for level . At each progression, the calculation is expected to create the applicant sets from the extensive thing sets of the former level, noticing the descending conclusion lemma. gets to a field of the information structure that speaks to applicant set , which is at first thought to be zero. Numerous points of interest are precluded underneath, more often than not the most critical piece of the execution is the information structure utilized for putting away the hopeful sets, and tallying their frequencies.



```
$\mathrm{Apriori}(T, \epsilon)$
    $L_1 \leftarrow \{\text{large } 1 - \text{itemsets}\}$
    $k \leftarrow 2$
    while $L_{k-1} \neq \emptyset$
        $C_k \leftarrow \{a \cup \{b\} \mid a \in L_{k-1} \wedge b \notin a\} - \{c \mid \{s \mid s \subseteq c \wedge |s| = k-1\} \nsubseteq L_{k-1}\}$
        for transactions $t \in T$
            $D_t \leftarrow \{c \mid c \in C_k \wedge c \subseteq t\}$
            for candidates $c \in D_t$
                $count[c] \leftarrow count[c] + 1$
        $L_k \leftarrow \{c \mid c \in C_k \wedge \; count[c] \geq \epsilon\}$
        $k \leftarrow k + 1$
    return $\bigcup_k L_k$
```

E. Genetic Algorithm

Genetic Algorithm endeavor to join thoughts of common assessment The general thought behind GAs is that we can fabricate a superior arrangement in the event that we by one means or another consolidate the "great" parts of different arrangements (schemata hypothesis), simply like nature does by consolidating the DNA of living creatures [7].

Genetic Algorithm is fundamentally utilized as a critical thinking procedure to furnish with an ideal arrangement. They are the most ideal approach to take care of the issue for which little is known. They will function admirably in any hunt space since they shape an extremely broad calculation. The main thing to be known is the thing that the specific circumstance is the place the arrangement performs extremely well, and a hereditary calculation will create a brilliant arrangement. Genetic calculations utilize the standards of choice and advancement to create a few answers for a given issue.

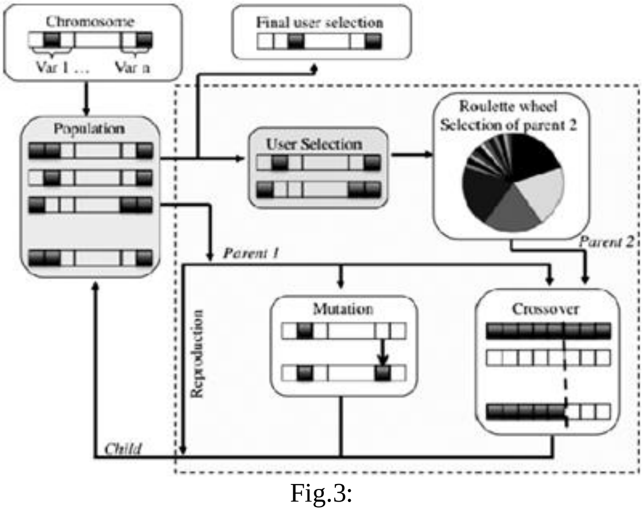


Fig.3:

Genetic calculations (GAs) [8] depend on an organic applications; it relies upon hypothesis of advancement. At the point when GAs are utilized for critical thinking, the arrangement has three unmistakable stages: • The arrangements of the issue are encoded into portrayals that help the essential variety and determination activities; these portrayals, are called chromosomes, are as basic as bit strings. • A wellness work makes a decision about which arrangements are the "best" living things, that is, most fitting for the arrangement of the specific issue. These people are supported in survival and proliferation, in this way offering ascend to age. Hybrid and change create another quality people by recombining highlights of their folks. In the long run an age of people will be deciphered back to the first issue space and the fit individual speaks to the arrangement.

VI. REFERENCES

[1]. Xingquan Zhu, Ian Davidson, “Knowledge Discovery and Data Mining: Challenges and Realities”, ISBN 978- 1-59904-252, Hershey, New York, 2007.

[2]. Joseph, Zernik, “Data Mining as a Civic Duty – Online Public Prisoners Registration Systems”, International Journal on Social Media: Monitoring, Measurement, Mining, vol. - 1, no.-1, pp. 84-96, September2010.

[3]. https://www.researchgate.net/figure/Data-mining-is-part-of-the-global-knowledge-discovery-in-database-KDD-Han-Kamber_fig13_268259025

[4]. https://www.datasciencecentral.com/profiles/blogs/the-7-most-important-data-mining-techniques

[5]. Lior Rokach and Oded Maimon,“Data Mining with Decision Trees: Theory and Applications(Series in Machine Perception and Artificial Intelligence)”, ISBN: 981-2771-719, World Scientific Publishing Company, , 2008.

[6]. https://en.wikipedia.org/wiki/Apriori_algorithm

[7]. AnkitaAgarwal,”Secret Key Encryption algorithm using genetic algorithm”, vol.-2, no.-4, ISSN: 2277 128X, IJARCSSE, pp. 57-61, April 2012.

[8]. Li Lin, Longbing Cao, Jiaqi Wang, Chengqi Zhang, “The Applications of Genetic Algorithms in Stock Market Data Mining Optimisation”, Proceedings of Fifth International Conference on Data Mining, Text Mining and their Business Applications,pp- 593-604,sept 2005.