International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 5 Issue: 5 678 – 682

A Survey On Data Mining Techniques and Applications

Sunil Kumar Patel
PG Scholar, Department of Computer Engineering and Applications, NITTTR, Bhopal, India
E-mail- sunil.skpatel3991@gmail.com

Kshipra Soni
PG Scholar, Department of Computer Engineering and Applications, NITTTR, Bhopal, India
E-mail: kshiprasony@gmail.com

Abstract— Data Mining refers to the analysis of experimental data sets to seek out relationships and to summarize the data in ways in which are each comprehensible and helpful. Compared with alternative DM techniques, Intelligent Systems (ISs) based mostly approaches that embody Artificial Neural Networks (ANNs), fuzzy pure mathematics, approximate reasoning, and derivative-free optimisation strategies similar to Genetic Algorithms (GAs), are tolerant of impreciseness, uncertainty, partial truth, and approximation. This paper reviews varieties of Data Mining techniques and applications.

Keywords- *Data mining, clustering, classification, MST.*

*****

I. INTRODUCTION

The wide-spread use of distribution information systems leads to the construction of large data collection in business, science and on the web. These data collections contain a wealth of information, which however needs to be discovered. Business will learn from their dealing data a lot of regarding the behavior of their customers and so will improve their business by exploiting this data. Science can obtain from observational data (e.g. satellite data) new insights on research questions. Web usage information can be analyzed and exploited to optimize information access [3]. Data mining provides methods that enable extracting from massive data collections unknown relationships among the data items that are helpful for higher cognitive process. Therefore data processing generates novel, unexpected interpretations of data [1] [2]. The development of data Technology has generated large amount of databases and large data in various areas. The analysis in databases and technology has given rise to association degree approach to store and manipulate this precious data for additional deciding. Data processing may be a process of extraction of useful information and patterns from immense information. It's additionally known as information discovers method, knowledge mining from data, information extraction or information/patter analysis.

II. KNOWLEDGE DISCOVERY PROCESS

Knowledge discovery could be a method that extracts implicit, probably helpful or antecedently unknown information from the data [11]. The knowledge discovery method is delineated as follows.

- Data comes from variety of sources is integrated into one data store known as target knowledge.
- Data then is pre-processed and remodeled into customary format.
- The data mining algorithms their information to the output in type of patterns or rules.

Then those patterns and rules area unit understood to new or helpful information. The ultimate goal of data discovery and data processing process is to seek out the patterns that square measure hidden among the large sets of information and interpret them to helpful knowledge and data. As delineated in method diagram higher than, data processing could be a part of information discovery process.

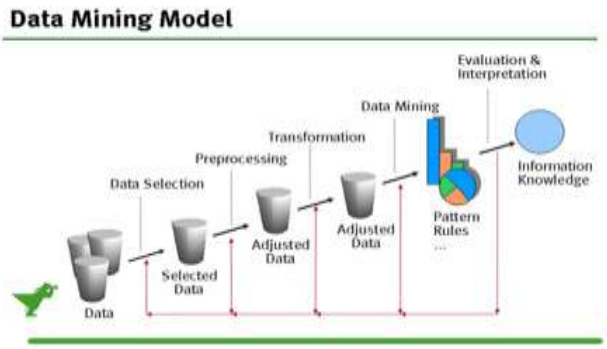

Fig 1: A Block Diagram of Data Mining Model

III. DATA MINING APPLICATIONS

1. *Financial Data Analysis:-*

The financial data in banking and money business is usually reliable and of top quality that facilitates systematic data analysis and data mining. [10]A number of the everyday cases area unit as follows –

- Design and construction of data warehouses for multidimensional data analysis and data processing.
- Loan payment prediction and shopper credit policy analysis.
- Classification and cluster of shopper for targeted selling.

- Detection of cash lavation and alternative money crimes

2. *Retail Industry:-*

Data Mining has its nice application in Retail Business as a result of it collects deal of information from on sales, shopper getting history, product transportation, consumption and services. It s natural that the amount of data collected can still expand apace attributable to the increasing ease, accessibility and recognition of the online. Data mining in retail trade helps in distinguishing client shopping for patterns and trends that cause improved quality of client service and smart client retention and satisfaction [10]. Here is that the list of samples of data mining with in the retail trade –

- Design and Construction of data warehouses supported the advantages of data mining.
- Multidimensional space analysis of sales, customers, products, time and region.
- Analysis of effectiveness of sales campaigns.
- Customer Retention.
- Product recommendation and cross-referencing of things.

3. *Telecommunication Industry:-*

Today the telecommunication trade is one among the foremost rising industries providing numerous services like fax, pager, mobile phone, chat messenger, images, e-mail, internet information transmission, etc. Because of the event of latest pc and communication technologies, the telecommunication trade is apace increasing. This can be the rational why data mining is become vital to assist and perceive the business [2]. Data mining in telecommunication trade helps in distinguishing the telecommunication patterns, catch fallacious activities, build higher use of resource, and improve quality of service. Here is that the list of examples that data processing improves telecommunication services –

- Multi-dimensional Analysis of telecommunication information.
- Fraudulent pattern analysis.
- Identification of bizarre patterns.
- Multidimensional association and serial patterns analysis.
- Mobile Telecommunication services.
- Use of visualization tools in telecommunication information analysis.

4. *Biological Data Analysis:-*

In recent times, we've seen an incredible growth within the field of biology reminiscent of genetics, proteomics, Genomics and medicine analysis. Biological data mining could be a vital a part of Bioinformatics. [10] Following square measure the aspects within which data mining contributes for biological data analysis –

- Semantic integration of heterogeneous, distributed genomic and proteomic databases.
- Alignment, indexing, similarity search and comparative analysis multiple ester sequences.
- Discovery of structural patterns and analysis of genetic networks and super molecule pathways.
- Association and path analysis.
- Visualization tools in genetic knowledge analysis.

5. *Other Scientific Applications:-*

The applications mentioned on top of tend to handle comparatively tiny and undiversified data sets that the applied math techniques are acceptable. Vast quality of data is collected from scientific domains resembling geosciences, astronomy, etc. An oversized quantity of data sets is being generated due to the quick numerical simulations in numerous fields resembling climate and scheme modeling, chemical engineering, fluid dynamics, etc [10]. Following are the applications of data mining within the field of Scientific Applications –

- Data Warehouses and data pre-processing.
- Graph-based mining.
- Visualization and domain specific Information

6. *Intrusion Detection:-*

Intrusion is a one type of action that threats the integrity, confidentiality and availability of the network resources. In the connected world the security is the one of most important issue. The large usage of net and convenience of the tools and tricks for intrusive and offensive network prompted intrusion detection to become an important part of network supervision. The data mining technology is applied in many areas for intrusion detection.–

- Development of data mining rule for intrusion detection.
- Association and correlation analysis, aggregation to assist choose and build discriminating attributes.
- Analysis of Stream information.
- Distributed data mining.
- Visualization and query tools.

IV. DATA MINING TECHNIQUES

There are square measure many data mining techniques are developed and utilized in data mining comes recently as well as association, classification, clustering, and prediction And successive patterns etc. square measure used for data discovery from. Data mining suggests that aggregation relevant data from unstructured knowledge. Therefore it's ready to facilitate attain specific objectives. The aim of an information mining effort is generally either to from a descriptive model or a prophetical model. A descriptive model presents, in prophetical type, the most characteristics of the information set. The aim of a prophetical model is to permit the

information jack to predict associate in nursing unknown (often future) worth of a selected variable; the target variable [7]. The goals of prophetical and descriptive model are often achieved employing a style of data mining techniques as shown in fig 1.

1) Classification:-

Classification supported categorical (i.e. discrete, unordered).This technique supported the supervised learning (i.e. desired output for a given input is known) .It square measure typically classifying the information supported the coaching set and values (class label). These goals square measure attain employing a call tree, neural network and classification rule (IF- Then).for example they apply the classification rule on the past record of the scholar United Nations agency left for university and measure them. Mistreatment these techniques we will simply establish the performance of the scholar. [9]

2) Regression:-

Regression is employed to map an information item to a true valued prediction variable [8]. In alternative word, regression is often custom-made for prediction. The regression techniques target worth square measure known. For instance, you'll predict the kid behavior supported case history.

3) Time Series Analysis:-

Time series analysis is that the method of mistreatment applied math techniques to model and justify a time-dependent series of information points. Statistic foretelling could be a methodology of employing a model to come up with prediction (forecasts) for future events supported known past events [9]. For instance exchange.

4) Clustering:-

Agglomeration could be an assortment of comparable knowledge object. Dissimilar object is another cluster. This method finds the similarities between knowledge per their characteristic [9]. This method supported the unsupervised learning (i.e. desired output for a given input isn't known). For instance, image process, pattern recognition, plans.

5) Summarization:-

Summarization is abstraction of information. It set of relevant task and provides an outline of information. For instance, long distance race are often summarized total minutes, seconds and height. Association Rule: Association rule is that the most well liked data processing techniques and penalized most frequent item set. Association strives to get patterns in knowledge that square measure primarily based upon relationships between things within the same dealing. Owing to its nature, association is typically spoken as "relation technique". This methodology of information mining is used analysis so as to spot a collection, or sets of merchandise that customers typically purchase at identical time [6].

6) Association:-

Association (or relation) is perhaps the higher known and most acquainted and uncomplicated data [9] processing technique. Here, you create an easy correlation between 2 or a lot of things, typically of identical kind to spot patterns.

7) Prediction:-

Proposes prediction methodology together with the opposite data processing techniques, involves analyzing trends, classification, pattern matching, and relation [9]. Prediction could be s wide topic and runs from predicting the failure of elements or machinery, to distinctive fraud and even the prediction of company profits. By analyzing past events or instances, you'll build a prediction concerning a happening.

8) Sequential patterns:-

DUAR and HART P [4] describe the assorted uses of successive patterns for distinctive trends, or regular occurrences of comparable events. For instance, with client knowledge you'll establish that customers obtain as selected assortment of merchandise along at completely different times of the year. During a hand basket application, you'll use this data to mechanically recommend that sure things are super imposed to a basket supported their frequency and past buying history [7].

V. SURVEY OF EXISTING RESEARCH

Ida Bagus Irawan Purnama, Neil Bergmann, Raja Jurdak, Kun Zhao, In this paper, authors first analyze the distribution of users by their predictability level and use their average distribution to define the related number of trips [1]. This is initially done to guide the definition of homogeneous user groups based on similar regularity patterns. Temporal and spatial analyses are then used to characterize the differences of the mobility patterns among the groups. To show the diversity of usage role in identifying the highly predictable users, entropy and predictability analysis is used with the classified users. Finally, the Markovian trait of the highly predictable users is tested by analyzing the similarity of real and conditional predictability.

Rokhmatul Insani and Hira Laksmiwati Soemitro, In this research, researchers attempts to try data mining applying in the field of telecommunication. They use data mining for customer segmentation and market basket analysis. For segmentation Authors compare 2 algorithms that are K-Means Clustering and Kohonen SOM Algorithm. And the result shows that K-Means Clustering is more suitable for this case [2]. This research take RFM model and K-Means Clustering,

then the result can be identified the profitable customers. After that, they use product packages of profitable customer to identify the relationship between the product packages. It is will be useful for company to give offering to their customer.

V´ıctor Mart´ınez, Fernando Berzal and Juan-Carlos Cubero, in this paper, authors have presented the NOESIS network data mining framework [3]. NOESIS is open source and lightweight. NOESIS algorithms are implemented using structured parallel programming patterns, which enable an Effective use of the available computing resources. The framework is built on top of a hardware abstraction layer that provides parallelization mechanisms and hides their underlying complexity. The NOESIS frame work is evolving and new data mining techniques are scheduled to be developed in the future, from overlapping community detection methods to quasi-local link scoring and prediction techniques, as well as additional graph layout techniques. Since the NOESIS graphical user interface is based on a model-driven application generator, creating ports of the application generator to other platforms, such as Android or the Web, will automatically enable the use of the NOESIS GUI in those platforms.

Michael Gowanlock, David M. Blair and Victor Pankrati us, uses three algorithms first is The Density-Based Spatial Clustering of Applications with Noise Algorithm (DBSCAN) and second is The Neighbor Search Algorithm, and third one is VARIANTDBSCAN Algorithm [4].These algorithms are combined shows great promise for helping to overcome these types of problems. In space weather applications it could contribute towards accelerated natural hazard warning systems. Authors approach allows for concurrent clustering of a dataset using multiple parameters. Authors maximize clustering throughput by mitigating the memory bottleneck through efficient indexing, reducing computation and memory pressure by reusing results of previously executed variants, and by optimizing the order of variant execution through scheduling heuristics. They also find that in the case Where low data reuse occurs between variants, the overhead of reusing data is not prohibitive in comparison to clustering the variant from scratch. The combination of techniques significantly out performs a sequential implementation across a wide range of application scenarios and input parameters on real-world datasets.

Xiao-jun Chen and Jia Key, proposes the concept of the time gap degree, designs an algorithm based on FPSPAN-Tree and FPSPAN-Growth [5]. It can merge the time data flow in the calculation of the single-dimensional FP-Growth algorithm in order to get a more reasonable method for mining frequent item sets in the time data flow. After the comprehensive analysis and comparison of the experimental results in terms of the running time of the system and the derivation rules, they draw a conclusion that FPSPAN-Growth leads to better Recall and Precision than other related methods while sharing similar time complexity.

Andreea Griparis and Daniela Faur, Mihai Datcu, this paper brings into focus a visualization based approach to mining the EO data [6]. This method aim stomp the existing data correlations in the multidimensional information space to the spatial correlations revealed by the 3D space. The assessment of the results considers a single and global quality criterion, involving the number of the intrusions and extrusion to reveal the performance of dimensionality reduction methods.

Ivan Kholod, Mikhail Kuprianov, Ilya Petukhov, suggested a approach in this paper makes that it possible to apply data mining algorithms for different architectures of IoT. Authors used the approach that decomposes a data mining algorithm into the functional blocks and maps them on actors. Using the actor's model as the execution environment allows us to use the data mining algorithms in the distributed architecture of IoT. Actors move part of computing to network layer of IoT and closer to source data. It increases performance of data analysis and decreases network traffic between the end devices and the cloud. They have maximum effect for big data [7].

Andrey N. Rukavitsyn, Mikhail S. Kupriyanov, Andrey V. Shorov and Ilya V. Petukhov, described a number of various methods of web page categorization have been implemented and modified in this article, particularly the methods based on "neighboring" web page classification [8]. A web page classification model was developed, which has an ability to define one of the 14 categories with precision micro-averaging of 96%. The main difficulties for Categorizations are web pages that contain no text. A person usually estimates the web page content on the basis of images; so possibly adding such attributes will help to improve the quality of the categorization. That is why one of the possible lines of the following research will be connected with new attributes based on images.

VI. CONCLUSION

This paper presents an outline of data mining and its techniques Units that are accustomed to extract fascinating and to develop vital relationship among variables keep in an exceedingly Brobdingnagian dataset. Data mining is required in several fields to extract the helpful data from themassive quantity of information. Great amount of information is maintained in each field to stay totally different records adore medical data, scientific information, academic information, dempgraphic information, financial data, marketing data etc. therefore, other ways are found to mathematically analyze the information, to summerize it, to find and characterize in it and to automaticaly flag anamolies.There are many data mining techniques are introduced by the researchers. These techniques are used to do classification and to try bunch and to search the interesting patterns.

REFERENCES

[1] Ida Bagus Irawan Purnama, Neil Bergmann, Raja Jurdak, Kun Zhao," Characterising and Predicting Urban Mobility Dynamics By Mining Bike Sharing System Data", 978-1-4673-7211-4/15 $31.00 © 2015 IEEE DOI 10.1109/UIC-ATC-ScalCom-CBDCom-IoP.2015.46

[2] Rokhmatul Insani, Hira Laksmiwati Soemitro, "Data Mining for Marketing in Telecommunication Industry", 2016 IEEE Region 10 Symposium (TENSYMP), Bali, Indonesia, 978-1-5090-0931-2/16/$31.00 ©2016 IEEE

[3] V´ıctor Mart´ınez, Fernando Berzal and Juan-Carlos Cubero, "The NOESIS Open Source Framework for Network Data Mining".

[4] Michael Gowanlock,David M. Blair, Victor Pankratius, "Exploiting Variant-Based Parallelism for Data Mining of Space Weather Phenomena", 2016 IEEE International Parallel and Distributed Processing Symposium, 1530-2075/16 $31.00 © 2016 IEEE DOI 10.1109/IPDPS.2016.10.

[5] Xiao-jun Chen, Jia Ke, "Fast Processing of Conversion Time Data Flow in Cloud Computing via Weighted FP-Tree Mining Algorithms, 978-1-4673-7211-4/15 $31.00 © 2015 IEEE DOI 10.1109/UIC-ATC-ScalCom-CBDCom-IoP.2015.87.

[6] Andreea Griparis and Daniela Faur, Mihai Datcu, "A Dimensionality Reduction Approach to Support Visual Data Mining: Co-Ranking-based Evolution",978-1-4673-8197-0/16/$31.00©2016 IEEE.

[7] Ivan Kholod, Mikhail Kuprianov, Ilya Petukhov, "Distributed Data Mining Based on Actors for Internet of Things", 5th Mediterranean Conference on Embedded Computing, MECO 2016, Bar, Montenegro.

[8] Andrey N. Rukavitsyn1, Mikhail S. Kupriyanov2, Ilya V. Petukhov,"Investigation of Website Classification Methods Based on Data Mining Techniques", 978-1-4673-8919-8/16/$31.00 ©2016 IEEE.

[9] Ranshul Chaudhary, Prabhdeep Singh, Rajiv Mahajan, "A SURVEY ON DATA MINING TECHNIQUES", International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 1, January 2014, ISSN (Print) : 2319-5940 ISSN (Online) : 2278-1021.

[10] Mrs. Bharati M. Ramageri, "DATA MINING TECHNIQUES AND APPLICATIONS", Bharati M. Ramageri / Indian Journal of Computer Science and Engineering Vol. 1 No. 4 301-305, ISSN: 0976-5166.

[11] Kalyani M Raval, "Data Mining Techniques", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 10, October 2012 ISSN: 2277 128X.