IJCSE International Journal of Computer Sciences and Engineering Open Access

Review Paper Vol.8, Issue.3, Mar 2020 E-ISSN: 2347-2693

Emerging Trends in Data Mining- An Algorithms used, Challenges and its Significance in Current Scenario

Gagan Gurung[1*], Rahul Shah[2], Dhiraj Prasad Jaiswal[3]

[1]Department of Information Technology, The ICFAI University Sikkim, Gangtok, India
[2]Department of Information Technology, The ICFAI University Sikkim, Gangtok, India
[3]Department of Information Technology, The ICFAI University Sikkim, Gangtok, India

DOI: https://doi.org/10.26438/ijcse/v8i3.7882 | Available online at: www.ijcseonline.org

Received:19/Feb/2020, Accepted: 07/Mar/2020, Published: 30/MAr/2020

Abstract— Data is defined as the collection of facts. When data are refined it becomes information. Data plays a vital role in obtaining any kind of information. These data are collected from different sources. It is difficult to handle these data sometimes. In data mining, all the data are collected from different sources as per their need. From telecommunication to retail industries, from finance to educational institutes all these data give vital information for any kind of execution that needs to be done for the future course of action. However, some challenges need to be taken care of while obtaining these data. Some data might be noisy, some might be scattered. Looking into this scenario we should be very careful while playing with these data. Every organization heavily depends on these data for their planning and execution. Data mining plays a vital role in achieving the target of the organization. It should be handled in a precise manner. Though the challenges are many their significance is equally high. This paper explains the brief ideas on what are the new trends in data mining and how it is helping to overcome our needs as per the current requirements.

Keywords - Data Mining, Semantic Web, Multirelational data mining, Bio-informatics, Algorithms, Challenges in data mining.

I. INTRODUCTION

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cut costs, or both. Since data are huge so it has to be mined in such a way that it produces an accurate result and it gives clear dimensions to our need [1]. Data mining is done to discover some knowledge in databases. The need for data mining is to identify interesting patterns and establish relationships to solve problems through data analysis. Various data mining software finds specific requirements as per the need [2]. For example, data mining software can help retail companies find customers with a common interest. Data mining can be applied in various fields like retail, telecom, finance and many more.

II. USES OF DATA MINING

1. The Fight against Terror

After the September 11 attack in the USA, many countries approved new laws in the fight against terrorism. These laws allow intelligence services to gather all information deemed necessary to prevent new attacks and to swiftly identify potential terrorists. Total Information Awareness program-The goal of this program was the creation of a huge central database that consolidates all the available information on the population. Data mining can be used to collect all the relevant data across the world to supervise the potential threats that may cause any of the attacks and identifying the potential measures to overcome it.

2. Bio-Informatics

Bioinformatics is the science concerning management, mining, and interpretation of biological sequences and structures. The Structural Genome Initiative has the aim of cataloging the structure-function information of proteins. Progress in technologies such as microarrays resulted at the beginning of the subdomains of genomics and proteomics. These fields study the genes, proteins and the circuit inside the cell that regulates the gene expression. Lots of data is being generated, data that must be mined if mankind ever wants to expose the mysteries of cells. Many accidents do occur where it is impossible to identify, so in these cases, data mining can help us to track the relevant data from the source.

3. Retail

Retailers were always on the front-line for the adoption of data mining. Since longtime association analysis is used to examine which products are frequently sold together and based on this analysis retailers can choose the optimal layout of the shelves to maximize [2]. By promoting self-scanning, the retailers can obtain worthwhile extra information: they can follow perfectly how their customers wander through the shop. This provides new possibilities for data mining. Retailers can find out the buying habits of a customer, the specific items that are purchased in large numbers and the items that are in more demand. Through data mining, the retailers can identify the potential area where they can supply and generate the new venture as per the requirements.

4. Semantic Web

Links that work today can suddenly stop working because someone changes the directory structure of the site that is being referred to. The mixture of content and lay-out tags into one document makes them unsuitable for machine processing. Based on the title of a web page and words that appear in the document, a search engine tries to grasp the meaning of a web page but this approach often fails. For example, a search for the exact birthday of Albert Einstein on the internet with the keywords 'birthday' and 'Einstein', returns several thousands of results but many of these results are unrelated to our inquiry. To fulfill this goal it is necessary to complement web documents with machine-readable meta-data and that is exactly what the second generation Internet, coined the Semantic Web or SemWeb, promises to realize.

5. Financial Data Analysis

Data mining can be used to track the fraud debit/credit card used. It also helps to identify the potential loyal users of the bank. It helps to identify the stock market pattern from past to present. It helps to identify the customers who are frequently using the cards in any payment methods. It helps us to identify the new areas where many schemes can be given to the customer based on the requirements. Financial data needs to be taken care of as it is one of the biggest parts of contribution to the national GDP of the country.

III. NEW TECHNIQUES IN DATA MINING

1. Multi Relational Data Mining

Most data mining algorithms are propositional; this means that they were devised to discover patterns in a single data table. However, larger databases generally contain several tables between which several relations have been defined. From this database, one wants to establish a decision tree so that the important customers can be identified swiftly. An example of such a rule is: IF (x is married to a person with income > 10800) THEN important customer (x) = YES.

2. Support Vector Machine

Classification and regression are probably the most widespread applications of data mining. For example, neural networks have proven to be excellent classifiers, but due to their complexity, it is tough to understand why certain classification decisions are made. Classifications by decision trees on the other are motivated by several rules that are represented by the tree. Recently, a new black-box technique has been proposed that shows even better performance: support vector machines. The basic idea behind SVMs is the following: the data is first being mapped into a high-dimensional space and afterward a linear classifier is constructed in this high- dimensional space.

IV. FIVE IMPORTANT FUTURE TRENDS IN DATA MINING

1. Multimedia Data Mining

It involves the extraction of data from different kinds of multimedia sources such as audio, text, hypertext, video, images, etc. and the data is converted into a numerical representation in different formats. This method can be used in clustering and classifications, performing similarity checks, and also to identify associations.

2. Ubiquitous Data Mining

This method involves the mining of data from mobile devices to get information about individuals. Despite of having several challenges in this type such as complexity, privacy, cost, etc. this method has a lot of opportunities to be enormous in various industries especially in studying human-computer interactions.

3. Distributed Data Mining

This type of data mining is gaining popularity as it involves the mining of a huge amount of information stored in different company locations or at different organizations. Highly sophisticated algorithms are used to extract data from different locations and provide proper insights and reports based upon them.

4. Spatial and Geographic Data Mining

This is a new trending type of data mining which includes extracting information from environmental, astronomical, and geographical data which also includes images taken from outer space. This type of data mining can reveal various aspects such as distance and topology which is mainly used in geographic information systems and other navigation applications.

5. Time Series and Sequence Data Mining

The primary application of this type of data mining is a study of cyclical and seasonal trends. This practice is also helpful in analyzing even random events that occur outside the normal series of events. This method is mainly being used by retail companies to access customer's buying patterns and their behaviors [3].

V. CHALLENGES IN DATA MINING

1. Noisy Data

The process of collecting data is from different areas. The data we collect in the real-world is noisy, unstructured, and fairly diverse in its fields. In such situations, data in large quantities will be fairly unreliable. These challenges are mostly due to errors in measurement and quantification by instruments or simply due to human errors. For e.g., many customers might give wrong email ids when retailers collect it. At the time of sending different discount schemes to the customer, it will be difficult for them to send it to the appropriate people.

2. Distributed or Scattered Data

The data existing in the real world is stored in several different mediums. It could be on the internet, or even protected databases. To bring all the data to a single structure is a very beneficial data mining goal, but contains a lot of speed bumps in organizational terms [4]. For example, several different geo-located offices owned by the same organization may have their data saved in hundreds of different locations on protected databases. Hence, data mining demands the manpower, the algorithms, and the tools regarding that specific area [5, 6].

3. Complex Data Restructuring

The data that exists in the real world also has several different forms. There can be data in text form, numeric form, graphical form, audio form, video form and the list goes on. Extracting the required information from this data and compiling the required information from this diverse and heterogeneous medium of data can be complex. The data extracted from these sources might not be as relevant or can give unambiguous meaning to it.

4. Algorithm Performance

One of the most essential areas of data mining is algorithms. The performance of the data mining process ultimately depends on the mining methods and algorithms used. If these mining methods and algorithms are not up to the mark for the task assigned, the result will not be as required and will ultimately affect the end data. This additionally affect the complete campaign.

5. Background Knowledge Incorporation

Background knowledge is one of the essentials for a proper and perfect data mining technique. Background Knowledge enables the end data of the data mining procedure to be more accurate which is why it plays such an essential role. With background knowledge, predictive tasks can become actual predictions and descriptive tasks can produce more accurate results. However, collecting and implementing background knowledge is a time consuming and difficult process for data mining organizations.

6. Data Protection and Privacy

One of the most common issues for individuals and both private and governmental organizations is the privacy of data. The field and operations of data mining normally lead to serious data security and protection issues. A great example would be a retail company noting down the grocery list of a customer. This data can be a clear indication of customers' interest in several products. This is one of the many reasons hundreds of data mining companies around the world take the most security measures to secure the data being collected[4][5][6].

VI. TOP DATA MINING ALGORITHMS

1. C4.5:

C4.5 is an algorithm that is used to generate a classifier in the form of a decision tree and has been developed by Ross Quinlan. And to do the same, C4.5 is given a set of data that represents things that have already been classified.

C4.5 that is often referred to as a statistical classifier is an extension of Quinlan's ID3 algorithm. The decision trees that are generated by C4.5 can be further used for classification. The C4.5 algorithm has also been described as "a landmark decision tree program that is probably the machine learning workhorse most widely used in practice to date" by the authors of the Weka machine learning software.

2. k-means:

k-means clustering that is also known as the nearest centroid classifier or The Rocchio algorithm is a method of vector quantization, that is considerably popular for cluster analysis in data mining.

k-means is used to create k groups from a set of objects just so that the members of a group are more similar. It's a well known popular cluster analysis technique used for exploring a dataset.

3. Support vector machines:

When it comes to machine learning, support vector machines that are also known as support vector networks are supervised learning models that come with associated learning algorithms which then analyze data that are used for the analysis of regression and classification.

An SVM model is created that is a representation of the examples as points in space, that are further mapped so that the examples of the separate categories are then divided by a clear gap that is ought to be as wide as possible.

4. Apriori:

Apriori is an algorithm that is used for frequent itemset mining and association rule learning overall transactional databases. The algorithm is proceeded by the identification of the individual items that are frequent in the database and then extending them to larger itemsets as long as sufficiently those item sets appear often enough in the database. These frequent itemsets that are determined by Apriori can be used for the determination of association rules which then highlight general trends.

5. EM(Expectation-Maximization):

An expectation–maximization (EM) algorithm, when it comes to statistics is an iterative method that is used to find the maximum a posteriori(MAP) or maximum likelihood estimates of parameters in statistical models, that depends on unobserved latent variables.

6. PageRank(PR):

PageRank (PR) that was named after Larry Page who is one of the founders of Google is an algorithm that is used by Google Search to rank the websites in their search engine results. PageRank, that is the first algorithm that was used by the company is not the only algorithm that is being used by Google to order search engine results, but it is the best-known way of measuring the importance of website pages.

7. AdaBoost:

Adaptive Boosting or AdaBoost, that has been formulated by Yoav Freund and Robert Schapire is a machine learning meta-algorithm, that won the founders the 2003 Godel Prize for the same. The algorithm can be used in composition with many other types of learning algorithms to improve performance. AdaBoost is sensitive to noisy data as well as outliers.

8. kNN:

The k-nearest neighbors' algorithm (k-NN) is a type of lazy learning or instance-based learning and is considered as a non-parametric method that is used for classification and regression. In both the mentioned cases, the input consists of the k closest training examples in the feature space and the output depends on whether the algorithm is being used for classification or regression. This kNN Algorithm is considered and is also among the simplest of all machine learning algorithms.

9. Naive Bayes:

When it comes to machine learning, Naive Bayes classifiers that are considered to be highly scalable are known to be a family of simple probabilistic classifiers that are based on the application of Bayes' theorem with the help of strong independent assumptions between the features.

10. CART:

CART is an algorithm that stands for classification and regression trees. It is a decision tree learning technique that either outputs classification or regression trees and similarly like C4.5, CART is also a classifier. Many of the reasons that a user would use C4.5 for also apply to that of CART, since both of them are decision tree learning techniques and features like ease of interpretation and explanation are applied to CART as well.

VII. Data Mining Tools

Following are 2 popular Data Mining Tools widely used in Industry

1. R-language:

R language is an open source tool for statistical computing and graphics. R has a wide variety of statistical, classical statistical tests, time-series analysis, classification and graphical techniques. It offers effective data handling and storage facility.

2. Oracle Data Mining:

Oracle Data Mining popularly knowns as ODM is a module of the Oracle Advanced Analytics Database. This Data mining tool allows data analysts to generate detailed insights and makes predictions. It helps predict customer behavior, develops customer profiles, identifies cross-selling opportunities.

IX. CONCLUSION

It is impossible to imagine our society today without data mining. Both in the scientific and industrial world, the applications have become too widespread. In this small paper, a short overview was given of some new domains in which data mining can cause immense changes. However, there are still many problems to overcome, from which privacy protection draws the most attention. Privacy protection deserves certainly a solid amount of attention, but it should not lead to an exaggerated apprehension of data mining. After all, the possibilities and opportunities of data mining are too valuable, for example in the development cycle of new medicines. These techniques are still subject to further research, but we expect that they will make rapidly the transition into a business environment. Therefore data serve as a medium for any fact findings and the result that we want to get for a future course of action.

References

[1] Bharati M. Ramageri, "*Data Mining Techniques and Applications* ", Indian Journal of Computer Science and Engineering , Vol.**1** Issue**.4**, pp. **301-305**, **2012**.

[2] Jiawei Han & Micheline Kamber "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, **pp. 400-436,2003.**

[3] Sadiq Hussain, "*Survey on Current Trends and Techniques of Data Mining Research* ", London Journal of Research in Computer Science and Technology, Vol.**17** Issue.**1** , **pp. 07-13,2017.**

[4] Rakesh Kumar Saini, "*Data Mining tools and challenges for current market trends-A Review*," International Journal of Scientific Research in Network Security and Communication, Vol.**7,** Issue.**2**, **pp.11-14, 2019.**

[5] Bindushree V., Rashmi G.R., Uma H.R., “*Analysis of Text Recognition with Data Mining Techniques*,” International Journal of Scientific Research in Computer Science and Engineering, Vol**.7,** Issue.**6, pp.40-42, 2013.**

[6] Arun K Pujari, ”Data Mining Techniques”,University Press,**pp. 01-50,2013.**