IOSR Journal of Electrical and Electronics Engineering (IOSR-JEEE)
e-ISSN: 2278-1676, p-ISSN: 2320-3331, Volume 14, Issue 3 Ser. I (May. – June. 2019), PP 80-86
www.iosrjournals.org

Analyzing predictive data mining techniques through pre-processed data sets

Dr. SAWALE NAGESH PANDITRAO.
GUEST FACILITY
DEPT OF COMPUTER SCIENCE
P. G CENTER HALHALLI BIDAR
GULBARGA UNIVERSITY KALABURAGI.

Abstract:
The field of Data Mining includes a subfield known as Predictive Data Analytics. It would be easier for us to make rapid Decision Making forecasts if we do Predictive Data Analytics on enormous data sets. These forecasts will be based on the findings gained from live or sample data. Data mining, which is the process of extracting hidden predictive information from large databases, is an innovative and powerful new technology that has the potential to assist individuals and businesses in concentrating on the data that is the most relevant to their needs within their data warehouses. This study assists us in detailing how to carry out predictive data analytics in relation to data mining by making use of a variety of different technologies. R Studio, Weka, and KNIME are just a few of the Open Source Tools that may be utilised to assist in the process of carrying out Predictive Analytics. This document also provides a list of numerous predictive analytic tools, describing both their properties and how they should be used. The reader can also do a comparison or decide to employ a certain tool depending on the requirements, whichever comes first. The primary goal is to further the study of predictive data analysis and to offer the essential assistance in making snap decisions in any of the significant domains. Analyses of predictive data may be carried out in many different fields, including medicine, agriculture, the prediction of the conduct of children, the behaviour of customers in a certain industry, and many more. In this regard, the article provides further explanation on the tools that can be used to carry out predictive data analytics and also presents an overview of the significance of data mining. Variables are used as qualities in predictive data analysis, and these variables are referred to as predictors. Additionally, information on the Knowledge Discovery Process, which is a component of Data Mining, is provided in this study.

Keywords: *Data Mining, Predictive Analytics, Data Sets, Decision Making, Attributes, Predictor, Knowledge Discovery.*

I. INTRODUCTION

Data mining, which is the process of extracting hidden predictive information from large databases, is an innovative and powerful new technology that has the potential to assist individuals and businesses in concentrating on the data that is the most relevant to their needs within their data warehouses. Using a variety of data mining technologies, we are able to forecast future patterns of behaviour and trends, which enables organisations to make decisions that are proactive and driven by information. The outcomes of predictive analytics are mostly helpful in decision making, which is an essential component of any organisation or individual endeavour. Data mining goes farther than the normal retrospective tools used in decision support systems since it provides automated, prospective evaluations of events that are yet to occur. Data mining techniques can provide answers to business concerns that, in the past, were impossible to address due to the amount of time it would take. They explore databases in search of obscure patterns and discover knowledge about future events that experts would overlook because it defies their preconceived notions. We are able to make snap judgments, save a significant amount of time, and quickly apply them to the outcomes of the forecast. The most time-consuming part of carrying out predictive data analytics is the acquisition of data. The majority of businesses have already begun to gather and analyse enormous amounts of data. Data mining techniques may be swiftly deployed on current software and hardware platforms to increase the value of existing information resources. These techniques can also be integrated with newly released products and systems as they come online. The analysis of enormous datasets is within the capabilities of data mining tools when they are deployed on high-performance client/server or parallel processing systems.

Data Mining Functionalities-

Data mining may be carried out on many different kinds of databases and information repositories, and there are also many different sorts of patterns that can be mined from the data. The mining of data can, in the broadest sense, be divided into two categories:

1. Descriptive
2. Predictive

The duties of descriptive mining focus on the broad characteristics of the data stored in the database. Interfering with the process of creating predictions using the present are predictive mining tasks and methodologies. Because the users aren't sure what sorts of patterns they should utilise to make their data interesting in a number of instances, they could look for a variety of different kinds of patterns simultaneously. As a result, it is necessary to have a data mining system that is capable of running a wide variety of patterns in order to satisfy a variety of users' requirements or satisfy the needs of diverse applications..

Description of a Concept or Class:

Data can be connected with classes or concepts. As a result, we are able to provide concise descriptions of data to individual classes and ideas. These descriptions, which are known as class concept descriptions. These descriptions can also be obtained using the vice-format.

- Data characterization, by briefing the data of the class
- Data discrimination, by comparing the target class
- Both data characterization and discrimination.

An Examination of Associations

It is a collection of rules that illustrates the requirements about attribute values that take place concurrently in a given piece of data. Widespread application in the analysis of transaction data or market basket. Association rules that are limited to a single domain are referred to as single-dimensional association rules, whereas association rules that are applicable to several domains are referred to as multidimensional association rules.

Classification and Prediction-

It is the process of discovering a collection of models that separate or characterise the data classes or concepts, with the purpose of being able to utilise the model for predicting the class of an item when the class label that was maintained is unknown. This process is known as model discovery.

The derived model relies on the data object in situations when the class object may be determined. The resultant model can be expressed in a variety of ways, including (IF-THEN) rules, decision trees, mathematical formulations, and neural networks, among other possibilities. The prediction of the class label of data items may be accomplished by classification. Despite the fact that prediction is not the same as classification prediction, the focus of prediction is on identifying distribution trends based on the data that is available..

Clustering Analysis-

It examines data items without referring to a previously established class label. In most cases, the sequential format does not include any class labels at all because it is not clear who these labels belong with. These labels are produced as a result of clustering. The items have been grouped together using the strategy of maximising the interclass similarity as the guiding principle. Each cluster that is generated as a result of this procedure may be considered to be a class of those particular objects, from which certain rules can be deduced.

Outlier Analysis-

Outliers are a common name for data items that stand out from the rest of a set of data because of their unique characteristics. Errors in measurement or in the performance of the task produce them.

Outlier mining has wide application:

- Fraud detection
- Useful for identifying the behavior of customers
- Medical analysis

Comparative Studies of Evolution and Deviation-

The study of data evolution defines the functions of models or trends for those items whose behaviour shifts over the course of a specific period of time. Various characteristics of such analysis include time series data analysis, sequence or priority pattern matching, and similarity based data analysis. It includes characterisation, discrimination, association, classification, and grouping across time-related data.

Objective

1. To study of different trends of data mining and pre-processed data sets and its characteristics
2. Study on Analyzing predictive of data mining.

Classification Methods for Data Mining Systems

The process of data mining involves the interconnection of data, as well as database management systems, machine learning, statistical analysis, visualisation, and information science. In addition to this, it employs methodologies from a variety of different research areas, including neural networks, knowledge representation, and high-performance computing. The kinds of data that are to be mined determine which techniques from spatial data analysis, pattern recognition, image analysis, information retrieval, signal processing, bioinformatics, web technology, economics, psychology, computer graphics, or business are integrated into the data mining system. Because there are many different disciplines that contribute to data mining, the goal of its study is to develop a wide variety of data mining systems based on the many sources that are employed. Systems for data mining may be broken down into the following categories:

Fig :1 Data mining as a confluence of multiple disciplines

The advancement of information technology has led to the creation of a vast number of databases as well as enormous amounts of data in a variety of domains. The study of databases and information technology has resulted in the development of a method that can store and modify this valuable data for the purpose of facilitating subsequent decision making. The practise of extracting usable information and patterns from massive amounts of data is known as "data mining." There are a few other names for this process: the knowledge discovery process, knowledge mining from data, knowledge extraction, and data analysis.

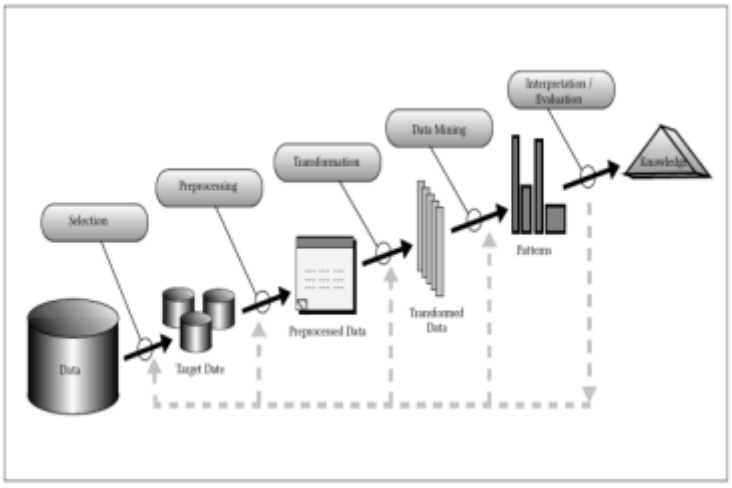

Figure 2. Knowledge discovery Process

Data mining is a method that uses logic to search through enormous amounts of data in order to locate data that is helpful. This procedure is used to find information. The objective of this method is to recognise patterns that had not been seen or recognised before. Once these patterns have been identified, they may be put to use in the process of making specific decisions about the growth of their enterprises.
Three steps involved are

- Exploration
- Pattern identification
- Deployment

Data Mining Algorithms and Techniques

For the purpose of knowledge discovery from databases, a wide variety of algorithms and methods, such as classification, clustering, regression, artificial intelligence, neural networks, association rules, decision trees, genetic algorithm, and the nearest neighbour method, amongst others, are utilised.

Classification

Classification is the data mining approach that is used the most frequently. It involves the utilisation of a collection of instances that have already been categorised in order to construct a model that is capable of classifying a huge population of records. Applications that monitor for fraudulent activity and credit risk lend themselves particularly well to this kind of research. This strategy usually makes use of classification algorithms that are based on either decision trees or neural networks. Learning and classification are both required steps in the process of classifying data. During the learning process, a classification algorithm is used to examine the training data. Test data are utilised in classification in order to provide an approximation of the correctness of the classification rules. In the event that the precision achieved is satisfactory, the rules may be implemented on the newly created data tuples. This would contain entire records of both fraudulent and lawful activity as assessed on a record-by-record basis, and it would apply to an application for the detection of fraud. The pre-classified samples are put to use by the classifier-training algorithm, which analyses them in order to find the necessary parameters for accurate discrimination. The parameters are subsequently encoded into a model that is referred to as a classifier by the algorithm.

Prediction

The term "prediction" refers to a broad field of research that encompasses a wide variety of topics, such as the detection of frauds and intrusions, the forecasting of the breakdown of machinery, and even the forecasting of the future of an organisation or a corporation. When integrated with the methodologies of data mining, prediction requires the completion of a variety of activities, including the study and generation of trends, classification, clustering, pattern identification, and relation. By doing a careful investigation and study of the patterns or happenings of the past,

Long Term Processing –

Data analytics and predictive analytics are solely reliant on the information and data that is processed over the course of a certain amount of time. It is necessary to first store the data for an extended period of time, and then to process the data in order to analyse it for patterns, classification, and categorization, as well as prediction. For example, in order to construct a pattern using predictive learning or sequential patterns, one must first store and evaluate the historical data and information instances before attempting to construct a pattern. The passage of time enables the collection of fresh facts and information, which is 3. The accuracy of a measurement on the occurrence of an event in the future can be improved by using predictive analytics.

Sequential Patterns –

Sequential patterns are created over a significant amount of time, during which it is possible to identify trends as well as actions or occurrences that are quite similar to those that occur on a consistent basis . The ability to recognise patterns and happenings that are quite similar is regarded as a very helpful tool. Take for instance a shopper at a grocery store who, over the course of a year or two, consistently acquires a number of items to add to their collection. The sequential pattern of what should be added to the grocery list may be determined using analytics performed on historical data. These analytics can be used to examine trends of supermarket purchases made throughout the year.

Decision trees –

The aforementioned methodologies are, in essence, connected to decision trees in some way. They can either be used to offer the criteria for selection or to provide support to the selection and use of certain data within the broader structure. Either way, they can be put to use in one of two ways. The construction of a decision tree begins with a question that can provide several answers or options, as this is required for the process to get under way. Each of the results, in turn, leads to another query that it is possible to classify into a result set that includes more than one choice. These additional questions lead to the classification of the data set in order to make the prediction based on the result set easier to do. detected and processed together with the

previous data and information, and the analytics are applied to the entire collection of data in order to accommodate the extra data.

Data Mining Applications

Mining for data is a relatively new technique that has not yet reached its full potential. Despite this, there are a number of different sectors that are already making frequent use of it. Retail outlets, medical facilities, financial institutions, and insurance firms are examples of some of these types of businesses. A good number of these companies are integrating data mining with a variety of other helpful techniques, such as statistics, pattern recognition, and other key methods. Data mining is a technique that allows users to discover patterns and relationships in large amounts of data that would otherwise be difficult to discover. This technology is widely used by many companies since it enables them to have a deeper understanding of their clientele and to make informed choices regarding their marketing strategies. The following is an outline of the challenges faced by businesses and the solutions discovered by employing data mining technologies.

DIFFERENT TRENDS OF DATA MINING AND PRE-PROCESSED DATA SETS

Many tough research challenges arise in the field of data mining as a result of the wide variety of data, data mining tasks, and data mining methodologies. For DM researchers as well as data mining system and application developers, one of the most essential jobs is to work on the creation of data mining methods, systems, and techniques that are efficient and successful in solving huge application challenges. The year 1980 marked the beginning of the period of data mining applications, which were mostly thought of by research-driven tools. In the world of data mining, there are a few different trends that can be seen in terms of the technologies and methods that are currently being created and investigated. The early day's trends in data mining are outlined in the following.

Data Trends

In the past, tabular data was stored in flat files, classical databases, and relational databases. Today, flat files are still utilised for this purpose. The most successful applications of data mining algorithms are those that make use of numerical data obtained from a single data source. Various data mining approaches have developed throughout time. Later on, with the combination of techniques used in statistics and machine learning, a variety of algorithms were developed in order to mine relational databases and non-numerical data.

Computing Trends

The world desperately needs more computing power. The programming languages of the fourth generation and the computer techniques that are associated with them have had a significant impact on the area of data mining. In the earlier days of data mining, the vast majority of algorithms relied solely on statistical methods. Later on, they progressed through the use of various computer methods such as artificial intelligence, machine learning, and pattern reformation.

Current Trends

These days, DM has gained a lot of popularity as a result of its enormous success in terms of the wide range of application successes it has achieved and the scientific advancement it has made. New obstacles have been presented to the world of data mining as a result of the ever-increasing complexities in a variety of disciplines and advancements in technology; in order to address these diverse challenges, some of the current trends include the following:

Mining of Data Using Distributed and Collective Resources (Cdm)

In the process of distributed data mining, the data being mined are situated in a variety of locations, both online and off. The primary purpose of the CDM is to efficiently mine dispersed data that is situated in a variety of different places. CDM offers a superior method for dealing with vertically partitioned datasets by making use of the concept of orthonormal basis functions. It then computes the basis coefficients in order to construct a global model of the data (Kargupta et. al., 2000).

Hypertext And Hypermedia Data Mining

Mining data that consists of text, hyperlinks, text markups, and numerous other kinds of hypermedia information is referred to as hypertext and hypermedia data mining. This type of mining can be defined as data mining. For mining hypertext and hypermedia data, some of the most significant data mining techniques are classification (supervised learning), clustering (unsupervised learning), semi-structured learning, and social network analysis.

This particular kind of data mining involves the application of limitations, which serve to direct the process.

In many cases, this is paired with the advantages of multidimensional mining in order to endow the process with an increased degree of potency (Han, Lakshamanan, and Ng, 1999). The usage of constraints may be broken down into a few different categories, each of which has its own set of defining qualities as well as distinct function.

Ubiquitous Data Mining

The proliferation of mobile computing devices like laptops, palmtops, cell phones, and wearable computers has made it easy to access vast amounts of data from virtually anywhere. The Ubiquitous computing settings are subsequently giving rise Ubiquitous Data Mining (UDM). The process of analysing data in order to get actionable intelligence from ubiquitous computing's collected data is referred to as UDM. It is possible that gaining access to and analysing data stored on a pervasive computing device will present a number of obstacles.

Multimedia Data Mining

Mining and analysis of many sorts of data, such as photographs, audio, video, and animation, is what is meant by the term "multimedia data mining." Among the data mining strategies that can be used on multimedia information are rule-based decision tree classification algorithms, such as those found in Artificial Neural Networks and Instance-based learning algorithms, Support Vector Machines, as well as association rule mining and clustering methods. It's a very young area of research, but there's a lot of hope for its prospects in the future.

Spatial Data Mining

The geographic data comprises astronomical data, natural resources data, satellite data and space craft data. When studying geographic data and other forms of data that are connected to it, some of the data mining techniques and data structures that are employed include the utilisation of spatial warehouses, spatial data cubes, spatial online analytical processing, and methods of spatial clustering. The majority of these data are of an image-oriented nature, and they have the potential to represent a significant amount of information if they are mined and evaluated correctly.

Time Series Data Mining

It focuses on the objective of recognising movements or components that exist within the data sets of stock prices, currency exchange rates, the volume of product sales, biological measures, meteorological data, and other types of information (trend analysis). Movements that are long-term or part of a trend, seasonal changes, cyclical variations, and random movements are some examples of these (Han and Kamber, 2001). Rule induction methods such as Version Space, AQ15, and C4.5 rules are now being used in a number of applications related to time series data mining..

Business Trends

Early data mining applications focused mainly on helping businesses gain a competitive edge. The exploration of data mining for businesses continues to expand as e-commerce and e- marketing have become mainstream elements of the retail industry. Today’s business/ industry must be more cost-effective, very faster and offer high value services that ever before. Due to customer’s expectations and constraints, data mining becomes a fundamental technology in supporting customer’s transactions more accurately. Most probably classification and prediction Techniques are used for supporting business decisions and progressed to Decision Support Systems (DSS) and very recently it has grown to Business Intelligence (BI) systems.

Future Trends

The field of data mining has been established itself as the primary subject of computer science and has showed interest potential for the future advances as a result of the significant success that has been seen across a variety of application fields that make use of data mining. The ever-increasing technology and new application areas bring new difficulties and possibilities for data mining on a consistent basis; some typical tendencies for the future of data mining include the following:

II. Conclusion

Mining data is important for a variety of business-related purposes, including but not limited to discovering patterns, making predictions, and learning new information. The use of data mining techniques and algorithms, such as classification and clustering, amongst others, assists in the discovery of patterns that may be used to determine the future growth trends of enterprises. Data mining has a wide application domain almost in every industry where data is generated. Because of this, data mining is considered to be one of the most important frontiers in database and information systems as well as one of the most promising interdisciplinary developments in information technology.

References

[1]. Salmin, Sultana et al. 2009. Ubiquitous Secretary: AUbiquitous Computing Application Based on WebServices Architecture , International Journal of Multimedia and Ubiquitous Engineering Vol. 4, No. 4,October, 2009

[2]. Hsu, J. 2002. Data Mining Trends and Developments:The Key Data Mining Technologies and Applications for the 21st Century, The Proceedings of the 19th Annual Conference for Information Systems Educators (ISECON 2002), ISSN: 1542-7382. Available Online: http://colton.byuh.edu/isecon/2002/224b/ Hsu.pdf

[3]. Kotsiantis, S., Kanellopoulos, D., Pintelas, P. 2004. Multimedia mining. WSEAS ransactions on Systems,No 3, s. 3263-3268.

[4]. T. M. Mitchell. 1982. Generalization as Search, Artificial Intelligence, 18(2), 1982, pp.203-226.

[5]. R. Michalski., I. Mozetic., J. Hong., and N. Lavrac. 1986. The AQ15 Inductive Leaning System: An Overview and Experiments, Reports of Machine Leaning and Inference Laboratory, MLI-86-6, George Maseon University.

[6]. J. R. Quinlan.1992.Programs for Machine Learning, Morgan Kaufmann.

[7]. Han, J., & Kamber, M. 2001. Data mining: Concepts and techniques .Morgan-Kaufman Series of Data Management Systems. San Diego: Academic Press.

[8]. Kargupta, H. et al, "Collective Data Mining," in Advances in Distributed Data Mining, Karhgupta and Chan, editors, MIT Press, 2000.

[9]. Kargupta, H. and A. Joshi, "Data Mining To Go: Ubiquitous KDD for Mobile and Distributed Environments," Presentation, KDD-2001, San Francisco, August 2001.

[10]. Data Mining :Concepts, Models and Techniques: Authors: Gorunescu, Florin

[11]. Han, J. &Kamber, M. (2012). Data Mining: Concepts and Techniques. 3rd.ed. Boston: Morgan, Kaufmann Publishers

[12]. Data mining techniques and applications – A decade review from 2000 to 2011

[13]. Shu-Hsien Liao, Pei-Hui Chu, Pei-Yuan Hsia, Department of Management Sciences, Tamkang University, Taiwan, ROC

[14]. International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.2, No.5, September 2012 : Data Mining and Its Applications for KnowledgeManagement : A Literature Review from 2007 to 2012. Tipawan Silwattananusarn, Dr. KulthidaTuamsuk Khon Kaen University, Thailand

[15]. REVIEW OF LITERATURE ON DATA MINING Mrs. Tejaswini Abhijit Hilage1 & R. V. Kulkarni 2 IJRRAS 10 (1)