STUDY AND EXAMINATION OF DATA MINING TECHNIQUES, TOOLS AND MACHINE LEARNING ALGORITHMS FOR EFFICIENT DATA ANALYTICS

Jacob Ajak Makuach Abuol (2k20/SE/63), Nabiyu Samuel (2k20/SE/84) *Delhi technological university, department of software engineering, Delhi, India Ms. Jyoti Patidar*

Abstract— **Data mining is the area of research, which means that useful information or knowledge is extracted from previous data. Data mining defines large amounts of data as a process of finding information such as super market data for various technologies used for data mining, such as science, research, medicine, media, web, entertainment and many other areas, which is implemented with various goods, data mining model data warehouses and online analytical resources. Data mining has made an immense advancement in recent year but the problem of lost data has remained a big challenge for data mining algorithms. This paper analysed the predictive and descriptive techniques such as classification, regression time series analysis, predication and clustering, summarization, association rules, sequence discovery techniques on the basis of algorithms which is used to predict previously unidentified class of objects. Not only that but also, this paper analyses various free and open sources data mining tools like Weka, R, etc.**

Our aim is to find most accurate tool and technique of classification process. Comparative analysis indicates that that we can achieve best result using various Combinations of tools and classification technique.

I. Introduction

Data mining is exploring hidden information from huge data sets [1]. In this case, there is a need to mainly focus on washing out the data so as to build it feasible for additional processing [2]. In various domains, data mining is widely used such as insurance, banking, retail, research, astronomy, medicine, forecast of rainfall and Government security

[3] Text Data Mining is a research area that includes many research areas, such as natural language processing, machine learning, information retrieval (Salton, 1989) and data mining [4]. with the large range of Data Mining technique, information or forms of data presentation it is essential to describe the restrictions of the relevance of certain methods according to the achieved objectives or provided data. It is also important to understand how the problem must be solved with the Data Mining such as clustering, classification, regression and so on. The development of information communication technology has generated a huge quantity of data from dissimilar sources [5], that is stored in dissimilar geological locations. Every database may possibly have its individual structure to store data [6]. Multiple mining data sources dispersed at different geological location to determine helpful patterns are significant important for decision making [7].

Data is irrelevant to each other from different Sources. Information that is generated from dissimilar sources is integrated, fresh and helpful facts may appear [8]. This paper primarily focuses on the predictive and descriptive techniques such as classification, regression, time series analysis, predication and clustering,

summarization, association rules, sequence discovery techniques on the basis of

algorithms which is used to predict previously unidentified class of objects.

Section I contains brief introduction to Data mining. Section II contains Data mining techniques and in section III Applications and comparative analysis of descriptive and predictive techniques.

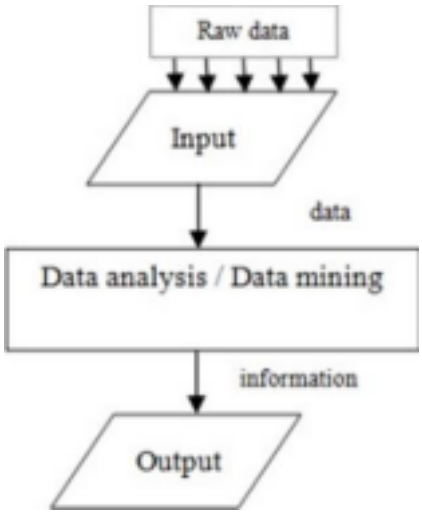

Fig 1: Knowledge discovery

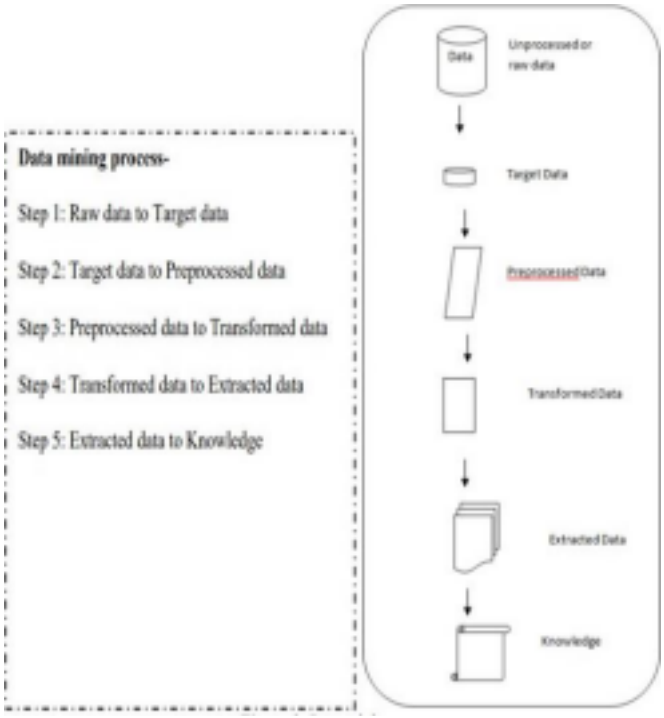

Fig 2: Data mining steps

II. Data Mining Techniques

The data mining process converts information from large data sets and transforms it into some sensible form for additional uses. So, it helps in achieving specific objectives. The goal of a data mining effort is usually to create a descriptive model or predictive model.

A. Predictive and Descriptive Data Mining techniques

Data mining is classified as primarily descriptive and predictive techniques, so that in order to fit a model for data, its various functions can be achieved. A predictive model predicts future values using different and historical data [9]. The prediction models include classification, regression, time series analysis and forecast. A descriptive model recognizes hidden patterns or relationships in data. It explores the properties of the data being examined. Descriptive models include clustering,

abbreviations, union rules, and series discovery [10].

B. Types of predictive techniques

Classification: This is the main data mining technique used, to develop a model that maintains a set of pre-classified examples that can categorize the population of large records.

It only used decision tree or neural network based classification algorithms. The general characteristics of classification work are classified as surveillance, providing new data to one of the categories dependent variables and well-defined classes [11]. Classification

techniques are used in customer segmentation, modelling business, credit analysis and many other applications. For example, classify countries classified by population, or classify bikes based on mileage.

Regression: Regression is another estimator data mining model also known as supervised learning technology.This technique analyses the dependency of certain attribute values, which depends on the values of other characteristics, which are present in the same item. The purpose value of regression technique is known. For example, you can estimate child's behaviour based on family history [12].

• Time Series data analysis: The time-series database uses the sequence of values or events received from the repeated capacity of the time. Prices are generally calculated at the same time interval as per hourly, daily, weekly.

A series database is a chain of events in any database, sometimes the actual thoughts of time [13]. For example, web page crossing sequences and customer purchase transactions are sequence data, but they cannot be time series data.

•Prediction: This technique reveals the relationship between independent variable and relationships between dependent and independent variables. Predicting is to predict the future situation, rather than getting a notification for natural disasters (floods, hurricanes, ice storms, etc.), pandemics, stock crashes etc in its applications. As another example, the number of sales of computer accessories can be estimated based on the number of computers sold in the last few months [14].

C. Types of Descriptive techniques

• Clustering: Clustering is a collection of identical data objects. Different object is another cluster; it is searching similarities between the data according to their features.

Clustering can be measured as identifying objects of the same classes. Using grouping techniques, we can identify solid and rare areas in the object space and take into account the general distribution pattern and the correlations in the data attribute [15]. The classification approach can also be used for efficient groups of categories of unique groups or objects, but this method is expensive so the use of clustering can be done in the form of pre processing approach to the specialty subset selection and classification [16]. For example, image processing, pattern recognition, city planning astronomy - the collection of stars,

galaxies, or super galaxies.

• Summarization: Summary is referred to as the intangible or generalization of data. Summarized technique map data in subset with simple description [17].The summarized data set gives an overview of all the objectives of the data with the collected information. Normal methods apply to normal and general deviation to analyse statistics, automated report generation and data visualization. For pre; Length can be measured as meter, centimetre or millimetre [18].

• Association: Association technology is used to remove the relations between properties and objects. In this technique, the occurrence of another model from the occurrence of a model means i.e., the object is due to the other and related causes and effect [19]. It is common for establishing a form of mathematical relationships between various mutually dependent variables of data mining; Association rules are useful for analysing and stimating client performance [20]. They also play an important role in the data analysis of the shopping cart, the grouping of products and the design of the catalogue and the design of the store. Federation rules being created by the programming system can be used to enable the machine to learn [21].

• Sequence Discovery: Highlight the relationship between data. It is a set of everything related to its individual timeline of events, for example, analysis of scientific experimentation, natural disaster and DNA sequence [22]

D. Comparative Analysis of Data Mining techniques

Techniques	Predictive types	Descriptive types
1	A predictive model predicts future values using different and historical data.	A descriptive model recognizes hidden patterns or relationships in data.
2	The main motive behind the predicted data mining is to prepare the model, which is to work for example assessment and classification.	The main purpose of descriptive data mining is to get an understanding of the analysis system by highlighting relationships and patterns in large data sets.
3	Based on data and analysis, the forecast record builds the model for record, and predicts the movement and properties of unknown.	Set the model or work important data is small, summary, educative differential forms.
4	The prediction models include classification, regression, time series analysis and forecast.	Descriptive models include clustering, abbreviations, association rules and series discovery.

E. The Comparative Study

The comparative study of data mining tools is carried out by selecting some open-source tools freely available on the internet and selecting of sample data sets from UCI machine learning repository. These data sets are then used with these tools in order to determine their performance and providing a complete analysis of their functionalities.

III.Experimental Analysis

The performance of these tools has been analyzed by first running them with several datasets available on UCI repository. Several different algorithms for classification and clustering were implemented in this analysis and performance of these algorithms was observed. In this section a sample of the experiment performed during the research is presented and conclusion of the results of different tools is discussed.

A. Accuracy for classification over audiology data

Data Set: Data set of (1) Audiology is used with data type multi variate, attribute type categorical, number of attributes 69, number of instances number of attributes, number of instances 226.(2)Zoo:data type: multivariate, attribute type: categorical, integer, number of attributes: 18, number of instances: 101 [24]Preliminaries: The classification was carried out of data set with percentage split methodology of 60% for training data and remaining 40% for the test data. The obtained measure of accuracy is used as the criterion for the performance analysis of the tools.

Accuracy% mapping	ZeroR	OneR
WEKA	27.27%	42.85%
KNIME	NA	NA
TANAGRAN	NA	NA

Accuracy% mapping	ZeroR	OneR
WEKA	38.23%	38.23
KNIME	NA	NA
TANAGRAN	NA	NA

IV. Observation

It has been observed that Weka has successfully run and implemented all the algorithms and produced appropriate results for the algorithms but with lowest accuracy that of ZeroR. Though ZeroR and OneR did not provide result, KNIME produced an accuracy of 89.36% with C4.5 over the zoo data set which is very close then 92.13% of that confirmed by C4.5 in Weka. Beside the non availability of ZeroR and OneR algorithm implementation, Tanagra's accuracy with C4.5 is satisfactory with result between 74.38% and 80.51% which is more stable then compared to 62.55% to 89.36% of what it is with KNIME.

Evaluation of Comparative Study

There are conclusions drawn from the study of these tools. The analysis of these tools has provided us with idea for the betterment of the whole data mining procedure. Even though these tools have proved to be appropriate for the specific data domains and specific data mining tasks such as classification, clustering, etc., the shortcomings, flaws or specificity of these tools have acted as a pullback from the

implementation of a general framework for data mining process. The major common drawback with these tools is that their processing of data, classification, clustering, prediction and inferring of rules all is based on the selection of the algorithm for data mining over a particular type of data set. If the selection of algorithm is not appropriate regarding the domain of data, then the produced patterns or predictions cannot be completely relied. For example, the classification over Audiology data set results with 84.41% correctly classified instances with Simple Logistic while with ZeroR it results in only 27.27%. If an inappropriate algorithm is used for future data value prediction, then it would produce incorrect results. Another issue with these tools is that the current state of art does not provide an automated mining technique. All the tasks such as classification and clustering are performed consecutively and are specific to an application [5]. A theoretical framework is required for implementation of unified theory where these data mining tasks can be unified and overcome with the shortcomings of these tools.

V. Conclusions

In this paper, we have given a brief introduction to data mining and various data mining techniques like classification, clustering, association rule, regression, anomaly detection and their features. Not only that, but also gone further to give brief discussion on machine learning. A comparative analysis of various data mining tools is also discussed in this paper. This paper shows various data mining technique can be applied for given domain and which data mining tool will help to analyse data by applying efficient algorithms. The best technique and tool of data mining can be determined by comparing the classification method and total accuracy. The discussion may be used to develop a new or modified data mining algorithm or software tool that can help industry to achieve best results.

References

[1] Deepashri.K. S Asst. Professor, Dept. of IS&E Adichunchangiri Institute of Technology, Chikmaglur, Karnataka, India 'Survey on Techniques of Data Mining and its Applications', International Journal of Emerging Research in Management &Technology, ISSN: 2278-9359 (Volume-6, Issue-2).

[2] Radhakrishnan Gopalapillaia* , Deepa Guptab, Sudarshan TSBa, 'Pattern Identification of Robotic Environments using Machine Learning Techniques', 7th International Conference on Advances in Computing & Communications, ICACC-2017, 22-24 August 2017, Cochin, India.

[3] Eunseog Youn a, Myong K. Jeong b,'Class dependent feature scaling method using naive Bayes classifier for text datamining', Article history: Received 18 September 2007 Received in revised for 1August 2008 Available online 24 December 2008.

[4] Mansi Gera, Shivani Goel," Data Mining Techniques, Methods and Algorithms: A Review on Tools and their Validity", International Journal of Computer Applications (0975 – 8887)Volume 113–No. 18, March 2015.

[5] Ayman E. Khedra, Mona Kadryb, Ghada Walidb, 'Proposed framework for implementing data mining techniques to enhance decisions in agriculture sector Applied case on Food Security Information Center Ministry of Agriculture', Egypt International Conference on Communication, Management and Information Technology (ICCMIT 2015).

[6] Vikas Gupta, Prof. Devanand 'A survey on Data Mining: Tools, Techniques, Applications, Trends and Issues', International International Journal of Computer Sciences and Engineering Vol.6(4), Apr 2018, E-ISSN: 2347-2693 © 2018, IJCSE All Rights Reserved 304 Journal of Scientific & Engineering Research Volume 4, Issue3, March-2013 1ISSN2229-5518.

[7] Catanzaro B, Sundaram N, Keutzer K. Fast support vector machine training and classification on graphics processors. In: Proceedings of the International Conference on Machine Learning, 2008. pp 104–111.

[8] Kochetov Vadim National Research Nuclear University MEPhI(Moscow Engineering Physics Institute),Moscow, Russian Federation Ko4etovvadim@gmail.com,"Overview of different approaches to solving problems of Data Mining", Procedia Computer Science 123 (2018)234–239.

[9] Ester M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996. pp 226–231.

[10] Ordonez C, Omiecinski E. Efficient disk based k-means clustering for relational databases. IEEE Trans Knowl Data Eng. 2004;16(8):909–21.

[11] Elkan C. Using the triangle inequality to accelerate k means. In: Proceedings of the International Conference on Machine

Learning, 2003, pp 147–153.