Journal of Innovative Engineering and Research (JIER)
Vol.- 4,Issue -1, April 2021, pp. 15-18 (4 pages)

Study on Data Mining Tools Open Source

SAMI ABDUL QADER MOHAMMED AL – ADEMI, DR. SANJAY SINGH BHADORIYA

Department of Computer Application, Dr. A. P. J. Abdul Kalam University, Indore (M. P.)- 452016
Corresponding Author Email: samialademi9 @gmail.com

Abstract— With Data mining, instruments and programming have raised that assistance in the investigation of the immense and expanding measure of information to get to information in different data sets, and these apparatuses work with work on most logical orders, including library and data sciences. Accordingly, this examination intends to consider what data mining is, its errands and applications. The study reached many results, the most important of which are: There is an advantage that characterizes some of the tools that are evident through use, which is the provision of a drag and drop model during the installation and construction process for data mining, which is available in four tools:KNIME, Orange, Weka, RapidMiner.

Index Terms— Data Mining, Information Mining, Web Mining, Bibliographic Mining.

I. INTRODUCTION

Interest in data mining began in 1989 during a workshop on knowledge discovery in databases. Piatesky-Shapiro, G. (Jan. 1991). Since then, this workshop was held continuously annually until 1994, and in 1995 the International Conference on Knowledge Discovery and Data Mining became one of the most important annual events, and then began planning a practical framework for data mining and knowledge discovery in two books: Knowledge Discovery in Databases, and Progress In knowledge discovery and data mining, after the year 2000 the possibility of storing huge amounts of data surpassed the ability of the human element to analyze and understand; There was no suitable tool for deriving information and knowledge from data, and specific models and rules could be created by means of data mining tools in light of the vast amount of data, which provided the necessary information for commercial activities, scientific discoveries, medical research and other fields.

The computer, and its prevalence and spread came from the expanding need for apparatuses that assistance in breaking down and understanding gigantic measures of information, and this information is created every day by different establishments like banks, insurance agencies, deals distribution centers and on the Internet, and this blast in information has likewise been joined by a tremendous expansion in the utilization of computers. System, scanners, advanced cameras, standardized identifications and the sky is the limit from there. With data mining, tools and programming have arisen that assistance in investigating the huge and expanding measure of information to get to information in different data sets, and these devices work with work on most logical orders, including library and data sciences.

II. PROCESS OF DATA MINING

There are hundreds of studies on data mining, but they are mostly focused on business and statistics, and there are few studies that deal with evaluating software and data mining tools and how to choose them, but they are general studies that are suitable for any software and not just data mining, and these studies focus on Commercial software due to the large number of services it provides and its difficulty in building and dealing with the huge amounts of data, and then some studies dealing with the evaluation of software and tools for data mining have been limited, which is the subject of the current study:

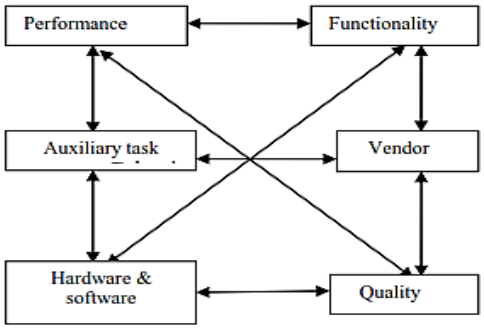
Androni, M. and Crisan, D. (2010).This study provides some of the available commercial data mining tools, along with some of the considerations related to the evaluation of data mining tools by companies that want to have such systems. Among the most important factors that companies must take into account are the amount of data available, how it can be stored, and the tasks of data mining that must be performed. It should be noted that the cost of the data mining system is important to the company, which has a limited impact on expanding the market for data mining products for companies.

Lyras, D., Panagiotakopoulos, T., Kotinas, I., Panagiotakopoulos, C., Sgarbas, K. and Lymberopoulos, D. Jun. 2014). This study aims to review, revise and test educational programs from several directions. The means of exploration techniques for educational data were used in the current study of the most popular evaluation criteria proposed by many researchers, and were tested and evaluated, taking into account the degree of impact on the efficiency of the educational program. By means of exploration techniques, especially forecasting and selection of advantages, the hidden relationship was investigated in the data collected from the experiments carried out in the Education Department of the University of Patras, which relate to the task of evaluating programs, and then the results of this study were presented and discussed in a quantitative and qualitative manner.

Bhargava, N., Aziz, A. and Arya, R. (2013)The number of data mining programs in the market is growing exponentially, so there is a need to select a standard for a software package that can be made available to intended users and organizations. With the continued increase in the number of programs and additional benefits included in the newer programs, it becomes more difficult to choose the appropriate program package, as a wrong decision may be reached with the loss of a lot of time and money, so many studies have been conducted around the world to evaluate the programs;

Journal of Innovative Engineering and Research (JIER)
Vol.- 4,Issue -1, April 2021, pp. 15-18 (4 pages)

However, the researchers did not reach the beneficiaries to generalize the selection and evaluation criteria. Improper software package selection can be extremely costly and negatively affect the business. The criteria that were followed in the study are as follows:



Criteria for selecting data mining programs

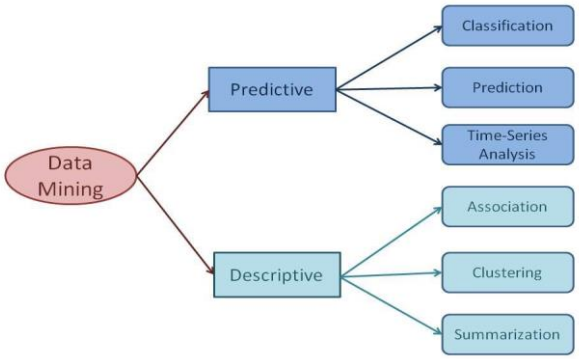
Qiu, M., Davis, S. and Ikem, F. (2004).Gatherings partition indistinguishable densities into various comparable subsets or gatherings that reflect areas of the informational index like models. This paper shows how a program assessment structure can fit the assessment of business data mining apparatuses to a particular climate of recipients. This examination applies the assessment of two principle business data mining devices, specifically SAS (EM) IBM DB2 Intelligent Miner (IM) and Enterprise Miner. For use in the university climate. Four rules have been utilized and (4) used to assess bunch procedures in data mining apparatuses:

1. Performance expected to be the capacity to deal with various information sources in an effective way, the rule of which is program establishment and homogeneous information access.
2. Function: Ability to implant a bunch of capacities, strategies and techniques for information mining. Standards: Diversity of calculations, strategy recently depicted.
3. Usage: It suits various levels and kinds of recipients without losing anything from the employment or pointlessness, and its measures: sorts of recipients, show of information.
4. Assisted task support: Allows the client to perform information cleaning, development, changing, showing and numerous different assignments that help information mining, and its guidelines: information sifting, and highlight induction.

III. DATA MINING TASKS

Data mining performs two fundamental activities as follows:

1. **Prediction:** Data mining plans to make expectations regarding the overall trademark or item ascribes of the obscure arrangement information, and the accessible learning model is utilized for forecast, grouping and relapse are two fundamental sorts of the expectation model, the first is utilized to anticipate the discrete or emblematic worth, while relapse is utilized to foresee nonstop qualities, for example The response to an inquiry concerning purchasing wares over the Internet is either X or Y, and this applies to the principal case, which is order, yet on account of gauging stock costs and patterns, it is through relapse undertakings. Anticipating models can distinguish market advantages and chances, and foresee the paces of utilization of land assets.
2. **Description:** The accessible potential information model that sums up connections is the narrative and logical job. Relationship investigation is regularly used to portray a model with solid social properties to determine significant models for discovering the connection between information. The inference of the properties of the equations communicates the overall attributes of the informational collection from the information stockroom, or the making of different highlights to recognize the qualities of a similar technique, for example, the deduction of highlights and recognize them from different cases. Regardless of its rationale, investigation of the gathering job can discover numerous significant connections, and this is known as the popular market bin examination, which was a distinct advantage in stores, where market bin investigation can assist with discovering deals of stores and their products. Padhy, N. Mishra, P. furthermore, Panigrahi, R. (June 2012)



Data mining tasks

The data mining classification can be divided into two sectors: direct and indirect data mining; the aim of direct data mining is to use the available data to create a model with a description of the variables; the goal of indirect data mining is not to choose a specific variable, but to build a relationship between all the variables. Classification, estimation and forecasting are part of direct data mining. The role of collecting, collating, describing and presenting is part of the indirect data mining. The role of the aggregator is not known in advance what knowledge is to be obtained, what can be obtained after analyzing the data, such as the beneficiary buying product (a) with product (b); as for the cluster, it is the grouping of similar records and putting them together in a group. The difference between the cluster and the classification is that the cluster does not depend on predefined classifications, nor is a trained group, while the description and presentation are a representation of the results of data mining. Weiping, F. and Yuming, W. (Dec. 2013).

IV. TYPES OF DATA MINING

- **Correlation analysis**

Correlation analysis: that is, the discovery of relevant and useful knowledge from a large group of data, and the basic

idea lies in a> b, where (f) expresses the set of traits, (b) represents the traits individually, and the rules interpret them if (and) its value is correct, Then (b) as a scalar value has the possibility and direction of the correct value in the database list. It can be explained that, after purchasing a commodity, how likely is it to continue purchasing good (B)? Jensen, D. and Neville, J. Correlation and Sampling in Relational Data Mining,(2001)

- **Decision Tree**

The decision tree is a series of nodes and branches, then the nodes are branched into sub-nodes by branches, where the nodes represent the features that must be considered in the decision-making process, and then the different values of the attributes come from the different branches; By using the decision tree model in making decisions, it is possible to search from the root to the leaves. The leaf nodes contain the results of each classification. Sharma, P., Bhartiya, R. (Dec. 2012)

- **Genetic Algorithm:**

Genetic algorithms search for possibilities to find the optimal process, resulting from specific or random groups, according to certain rules of the process to continue iterative calculation, such as selection, production, exchange and change, etc., which is the process of preserving good variables, eliminating bad variables, and directing the search to approach the optimal solution. According to the requirements of each person, the implementation of the genetic algorithm requires two data transformation processes, namely: decoding and encoding, where the coding consists in converting the parameters of the search distance to a chromosome or individuals from the genetic space; As for the decoding, it is represented in converting the chromosome or members of the genetic space into parameters of the search area. This has evolved the genetic algorithm based on simulation of genetics, to work directly on the structure of organisms, it has sufficient power without restrictions to do the process of derivation and function. Flockharta, I. and Radclieab, N. (1996)

- **Bayesian Networks:**

Virtual theory networks depend on a mathematical model for probabilistic inference, and probabilistic inference is done through some information to obtain probabilities for other variables, and Bayesian networks rely on the basis of probabilistic inference to solve the problem of uncertainty and incompleteness, and it has the best advantage to solve errors caused by difficult uncertainty and correlation, And it is widely used in many fields. Using Bayesian network architecture and conditional probability tables, it is possible to calculate the probabilities for specific node values after evidence is presented. Heckerman,D. (1997)

- **Rough Set Approach:**

Rough Set Approach is a mathematical method for addressing ambiguity and uncertainty using the Rough Set method that enables analysis of the decision table, evaluation of the significance of specific features, minimization of the set of characteristics and nuclear energy and elimination of additional characteristics from the decision table and classification rules that appear from the reduction table For decision-makers, the main idea of the Rough Set depends on the existing knowledge of a specific problem, by classifying actual data management, dividing the scope of the problem, minimizing data under the premise of retaining important information, reducing knowledge nuclear, assessing data independence, and deriving the concept classification rules. Pawlak, Z.(1884)

- **Neural Network:**

It is a dynamic system with a topological structure to direct the graph, it deals with information by responding to the continuous or intermittent state of input, and the neural network system consists of simple and large processing units, by linking to each other on a large scale and forming a complex network of systems. Although the structure and function of each cell is very simple, the behavior of the network system consists of a large number of colored and complex cells. The algorithm is suited for collecting data, which can present a lot of complex information and regular and organized data, to find the internal relationship between the data through the similarity of time and space. SINGH,Y. and Chauhan, A.

- **Statistical Analysis:**

It is an accurate method of data mining based on statistical probability theory, such as: regression analysis, factor analysis through models of objects and finding conclusions, and usually divided into the following steps: describing the nature of the analytical data, the research group of data relationships, model building, and data summary, And core group relationship, explain the validity of the model, and finally predict future development. SAS and SPSS are widely used as application programs for statistics. Friedman, J.

V. DATA MINING APPLICATIONS

1- Credit departments on advances: They depend on comparable people 'perceptions in buy, pay, and advance models. It is additionally conceivable to create constantly synopsis provides details regarding significant clients and Mastercards.

2- The general store: sorts out its merchandise as indicated by deals models and data about relationship between items, advertising to a particular gathering to discover clients to concede limits to them for a particular explanation, and to discover products that are sold together.

3- Intelligence Agency: audits spending models and travel information to recognize unusual conduct of its workers.

4- Clinician: Analyzes X-beam pictures to distinguish unusual examples.

5- Flight Reservation System: Uses data about venture out models and headings to expand seat use.

6- Banks: Data mining is utilized to get information that will assist with drawing in clients.

Albeit these applications have been utilized totally for quite a while, they depend on manual factual examination, and representatives have as of late began utilizing information mining innovation to investigate information, make equal connections, and make forecasts.

7- Applications for information mining in libraries:

I. Dealing with the library's property

II. Information bases of recipients

III. Human Resource Development

IV. Data administrations accessible in the library

V. Parts of spending

Open source data mining tools include:

- ❖ R IDE/Editors.
- ❖ Data Mining Software.
- ❖ Clustering.
- ❖ Association Rules.
- ❖ Sequence Analysis.
- ❖ Social Network Analysis.
- ❖ Process Mining.
- ❖ Spatial Data Analysis

VI. CONCLUSION

Data mining is a new type of intelligent information processing technology, and with the huge boom in information technology, we will see a horizontal and vertical expansion in the use of data mining applications, especially in military, security, intelligence and commercial applications, and data mining on mobile devices will be the future direction of data. The power of the Internet has overshadowed scientific datasets, the social hierarchy of the dataset, topology, geometry, and other characteristics in particular linking mining graphically and analyzing social networks. Data mining faces huge databases, so data mining algorithm must be efficient and scalable. Most of the current databases are relational; therefore, the emergence of other models of databases requires the ability to process types of data, and data mining specialists can speed up the process of data mining, that is, an interactive interface must be provided for beneficiaries suitable for expressing requirements and strategies. On the other hand, the interactive interface converts the various results to the user, i.e. a data mining system requires stronger interaction. At the same time, we find that the search for data mining may lead to an expansion of illegal data, and this constitutes a problem that must be solved.

REFERENCES

1. Piatesky-Shapiro, G., Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop, AI Magazine, 11: 5, pp. 68-70, Jan. 1991.
2. Jadhav, A., and Sonar, R., Framework for evaluating and selection of the software packages: A hybrid knowledge based system approach, The journal of system and software, 84, 8, pp. 1394-1407, 2011.
3. Goebel, M., and Gruenwald,L., A Survey of Data Mining and Knowledge Discovery Software Tools, SIGKDD Explorations, 1, 1, pp.20–33, 1999.
4. Ferguson, M. Evaluating And Selecting Data Mining Tools. InfoDB, 11, 2, pp: 1-10, 2017.
5. Mining Applications and Feature Scope International Journal of Computer Science, Engineering and Information Technology, 2, 3, pp. 23-40, 2015.
6. Weiping, F. and Yuming, W., The Development of Data Mining International Journal of Business and Social Science, 4, 16, 45-62, 2013.
7. Sharma, P., Bhartiya, R., Implementation of Decision Tree Algorithm to Analysis the Performance International Journal of Advanced Research in Computer and Communication Engineering, 1, 10, 172-184, 2012.
8. Zhao, Y., Chen, Y. and Yao, Y., User-Centered Interactive data Mining. Proceedings of the Sixth IEEE International Conference on Cognitive Informatics (ICCI'06), PP. 457-466.,2006.
9. Singh, Y. and Chauhan, A. Neural Networks In Data Mining Journal of Theoretical and Applied Information Technology, 5, 6, pp. 342-361, 2012.
10. Thuraisingham,B., A Primer for Understanding and Applying Data Mining. IT Pro IEEE Xplore, 2000.
11. Sharma, P., Bhartiya, R., Implementation of Decision Tree Algorithm to Analysis the Performance International Journal of Advanced Research in Computer and Communication Engineering, 1(10), Dec., 2012.