AN EFFICIENT ANALYSIS OF DATA MINING TOOLS AND TECHNIQUES AND THEIR COMPARISIONS

[1]Kv Naganjaneyulu, [2]S.Srisailam

[1]Professor & Principal,Department of CSE,Dhruva Institute of Engineering & Technology,Hyd

[2]M.Tech Student,Department of CSE,Dhruva Institute of Engineering & Technology,Hyd

([1]kvnaganjaneyulu75@gmail.com,[2]srisailam270@gmail.com)

Abstract -Data mining could be a method of exploring unknown patterns from large databases. This acts as a key to data discovery that provides a good support to business world and domain. To create this data discovery happening varied data processing tools are developed. These tools offer interface to urge information and to retrieve some attention-grabbing patterns out of it that are any helpful to achieve new data. There are forms of parameters outlined within the literature which give base for a tool to perform analysis and completely different tools are accessible to perform this analysis. this is often quite attention-grabbing to perform a comparative analysis of those tools and to watch their behavior supported some hand-picked parameters which is able to any be useful to search out the foremost applicable tool for the given information set and also the parameters.

Keywords-*Data mining; Data Mining Tools; Knowledge Discovery; WEKA; KNIME; Rapid Miner,Orange,Apache Mahout.*

I. INTRODUCTION

Data mining is that the method of finding patterns from great amount of information by applying some techniques. this is often used as Associate in Nursing instrument for data discovery in databases to be employed in decision making process. Massive organizations use it primarily for locating new ways in which to extend their profits and to reduce value. Data processing analyzes the info and helps to name the hidden factors in order that helpful patterns and data will be generated. For an instance, business organizations will analyze the client's behavior toward specific product by analyzing the historical knowledge and this helps the organization to seek out the dynamical behavior of the customer with the passage of your time, like, to seek out the trends in modification, to seek out the degree of modification etc. These sorts of findings is certainly facilitate any organization to require future choices in relevance that product [1][2]. Data processing tools square measure the code which gives automatic implementation of information mining techniques on the info and provides programme to use machine learning algorithms [2]. These tools will handle large quantity of information and supply relevant results expeditiously. Varied tools square measure discovered with completely different parameters in keeping with meet the various sorts of needs. The handling of information, programme, missing values, finding error rate and plenty of additional parameters create these tools completely different from one another. These parameters will be accrued or decreased in keeping with the requirement of user. These tools square measure having options of handling complicated still as unstructured knowledge [3]. Corporations bought data processing tool to make their own customize mining solutions. Several data processing tools square measure accessible with their strengths and limitations in context to parameters like interfaces, algorithms, accuracy of results, mining techniques, knowledge set size etc. These tools square measure any categorized into 3 classes i.e. Dashboards, Traditional data processing tools and Text Mining tools. Traditional data processing tools principally employed by corporations for business analytics purpose. These tools work on databases accessible with the corporate. There tools apply pre-defined algorithms on knowledge for locating the invisible pattern and results. These tools give broad knowledge classes to come up with clear reports. As an instance, a information of sales will show monthly sales results and reports with the assistance of ancient data processing tools. These tools square measure accessible each in Windows and operative system versions of operating systems and square measure principally used for on-line Analytical process (OLAP)[4]. a number of these tools square measure rail , R studio, fast laborer, SQL and D2K [5]. Dashboards square measure put in on pc to watch information info and reflects the updates and changes onscreen concerning business info and performance. These square measure principally employed by corporations that wish to see its sales from historical purpose of read with the assistance of historical knowledge i.e. knowledge Warehouse. Dashboards square measure simple to grasp and it give leads to the shape of charts and bar-graphs to produce summary concerning company's performance.

All details associated with profits and loss of company square measure visible to the manager on one screen interface and also the whole task is performed by dashboard options mechanically. The leading dashboards give the snap of actual performance of tools and conjointly show the recent happenings [6]. The business intelligence dashboards are called enterprise dashboards [7]. These have the power to tug the important time knowledge from multiple sources. Oracle[6] and Microsoft[8] square measure among the leading vendors of business intelligence dashboards[10]. Text mining is analyzing the text to extract info that will be helpful for explicit purpose. It deals with language text and lexical usage to seek out helpful info. Text mining tools simply access databases, scanned contents and embrace handling of structured and unstructured knowledge. Text analytic code modification unstructured knowledge into numerical values in order that it will link with structured knowledge and notice the result with ancient data processing tools.

Apache mahout[9] is a tool which can handle structured and unstructured information. There are some text mining tools that square measure open sourced like orange[11], NLTK[12], Voyant[13] and ALchemy API[14]. IBM company build smarter Apps with ALchemy language[15] for linguistics text mining[16] exploitation tongue Processing[17]. This application facilitate company to know worlds spoken communication, reports and photos. These tools square measure increasingly adding new options to satisfy the quick ever-changing necessities of the user and to handle the information quality in a very higher method. it's quite troublesome to feature all the options in one tool therefore there square measure totally different classes of tools introduced [2][18].

II. DATA MINING TECHNIQUES

There are several techniques of knowledge mining like classification, regression, clustering, summarisation that have their own characteristics and limitations. Classification [2] classifies information into completely different categories. There are several classification algorithms like call tree [19], Naive Bayes[20], Generalized linear Model [21] and Support Vector Machine[22]. The classification is performed primarily on the idea of parameters i.e. accuracy and confusion matrix [23][24]. this system give varied applications within the field of client interest, social network, medical and health care and plenty of additional [25].

Regression[26] is employed to map the link between 2 variables. This is often conjointly drawn within the map kind and may be accustomed check the result by examination the gap of knowledge points from regression line[2]. Profit, sq. footage, temperature, sales and distance are expected through regression. There are two formula's used for regression statistics i.e. Root Mean sq. Error (RMSE) and Mean Absolute Error [26][27]. In Clustering [2], another data processing technique; one performs the distribution of knowledge supported completely different classes. This system provides the solid information from huge amount of data sets. There are completely different strategies employed in bunch like partitioning technique, ranked technique, density based mostly technique, grid based mostly technique, model based technique and constraint based technique [28]. There are varied applications of bunch within the field of promoting, biology, fraud detection, similar land identification [3]. In Summarization [2] one will create a compact description of any information. Summarization is finished within the variety of table. The summarization provides the link between completely different sort of information sets [29]. There are two approaches for automatic summarization i.e. extraction and abstraction. Extraction technique work on existing words, phrases or sentences within the original text to make the outline. Theoretic technique use natural language generation techniques [30][31].

III. PARAMETERS

Parameters offer data concerning the analysis of techniques and tools. In data processing to look at the output we'd like parameters. It's worth offer information for decision making[32]. The performance analysis in data processing tools is completed by completely different parameters. It offer data concerning however the input vary and additionally offer accuracy concerning the results [33]. There are a unit numerous parameters used for testing however best parameter offer accuracy concerning mining patterns. Some common parameters used for comparison area unit developer, programming language, movability, interface, platform, visual image, accuracy and time taken [34].The values of those parameters area unit taken manually.

There is a unit some distinctive parameters altogether data processing tools. As an example, rail containing parameters for analysis i.e. properly classified instances, incorrectly classified instances, alphabetic character statistics, mean absolute error, root mean square error, relative absolute error, root relative square error[35].

The properly classified instances offer data concerning the accuracy in classification of categories. The F-measure combines preciseness and recall mean. The accuracy addresses the standard or state of being correct or correct worth of calculation. Alphabetic character statistics offer measuring concerning multiclass and unbalanced category. It tells however your classifier performs with the input. Mean absolute error measures the accuracy for continuous variables. Root Mean square error measures the common magnitude of error [36]. The Orange tool uses parameters for analysis like check and score. Check score offer accuracy estimation through cross validation. Second predictions that show predictions of models for an input dataset.

Third confusion matrix which offer data concerning classifier analysis. Fourth , ROC analysis that show the receiver in operation characteristics curve supported the analysis of classifier. Fifth, lift curve that construct and show the curve from the analysis of classifier[37]. The MATLAB tool uses parameters for analysis like accuracy, execution time and observation speed. The observation speed could be a distinctive parameter for evaluation[34]. The Rapid miner tool used parameters like accuracy , precision, Recall, AUC(Optimistic), AUC(neutral). The Rapid miner additionally contain performance vector for predicting the performance values [40]. The KEEL tools have distinctive options off/on line run of experiment setup that is new in data processing tools. Another distinctive parameters area unit sequence or path analysis and error rate [42].

IV. DATA MINING TOOLS

4.1 WEKA

Weka is a Java based mostly free and open source package licensed under GNU GPL and on the market to be used on Linux, Macintosh OS X and Windows. It contains a set of machine learning algorithms for data processing. It packages tools for information pre-processing, classification, regression, clustering, association rules and image. Individual may be a easy graphical interface for two-dimensional image of well-mined information. It enables you to import the data from numerous file formats, and supports accepted algorithms for various mining actions like filtering, clustering, classification and attribute choice. However, once handling giant information sets, it's best to use a CL based mostly approach as individual tries to load the total information set into the most memory, inflicting performance problems. This package conjointly provides a Java Appetiser to be used in applications and can connect with databases using CJD.

4.2 APACHE MAHOUT

Mahout is a machine learning algorithms which can help in classification and frequent pattern mining. It may be employed in a distributed mode that helps simple integration with Hadoop. Mahout is presently being used by a number of the giants in the tech industry like Adobe, AOL, Drupal and Twitter, and it's conjointly created a bearing in analysis and teachers. It may be a good selection for anyone searching for simple integration with Hadoop and to mine immense volumes of data.

V. CONCLUSION

In our paper more analysis has been done in comparison of data mining tools.Implementation of new algorithms are needed for rule mining to perform better decision making.An enhanced classification technique like Rough set theory to be used for getting a better results in rule structuring algorithm.The three main algorithms like KNN,Naive Bayes and decision tree can approximate the same amount of time with the defined set of parameters. There is a scope to check the efficiency of these algorithms by taking new parameters

REFERENCES

[1]. H. Jiawei , M. Kamber, J. Pei, Data mining concepts and techniques, 3rd ed., Morgan Kaufmann Elsevier: USA , 2012.

[2]. I. H.Witten, E. Frank, M. A.Hall, Data Mining practiced machine learning tools and techniques, 3rd ed., Morgan Kaufmann Elsevier: USA,2011.

[3]. 12 data mining tools and techniques [Online]. Available: https://www.invensis.net/blog/data-processing/12-data-mining-tools-techniques [Cited 2015 November 18].

[4]. OLAP Tools (Online Analytical Processing)[Online]. Available :http://www.informationbuilders.com/olap-online-analytical-processing-tools

[5]. 10 most popular analytic tools in business[Online]. Available from:http://analyticstraining.com/2011/10-most-popular-analytic-tools-in-business .[Cited 2011 January 15].

[6]. Defining dashboards, visual analysis tools and other data presentation media[Online]. Available from:http://www.dashboardinsight.com/articles/digital-dashboards/fundamentals/what-is-a-dashboard.aspx .[Cited 2011 November 28].

[7]. Enterprise Dashboard Digest[Online].Available from: http://enterprise-dashboard.com

[8]. Building and Using Dashboards[Online].Available from: https://docs.oracle.com/cd/E28280_01/bi.1111/e10544/dashboards.htm#BIEUG682

[9]. What is Apache Mahout[Online]. Available from: https://mahout.apache.org/

[10]. Teacher Dashboard[Online].Available from: http://www.teacherdashboard365.com/

[11]. Orange: Data mining Fruitful and Fun[Online].Available from: http://orange.biolab.si/

[12]. Natural language Toolkit[Online].Available from: http://www.nltk.org/

[13]. Voyant [Online] . Available from: http://voyant-tools.org/

[14]. Alchemy API Tools[Online].Available from: http://www.alchemyapi.com/developers/tools

[15]. Alchemy Language[Online].Available from: https://www.ibm.com/watson/developercloud/alchemy-language.html

[16]. A. Stavrianou, P. Andritsos, N. Nicoloyannis, Overview and Semantic Issues of Text Mining, SIGMOD Record.2007 September

[17]. Introduction to Natural Language Processing[Online] Available from:http://blog.algorithmia.com/introduction-natural-language-processing-nlp.[Cited 2016 August 11].

[18]. Predictive Analytics [Online].Available from:http://www.predictiveanalyticstoday.com/top-software-for-text-analysis-text-mining-text-analytics/

[19]. Decision Tree[Online].Available from: https://www.mindtools.com/dectree.html

[20]. 6 easy steps to learn Naive Bayes Algorithm[Online].Available from: https://www.analyticsvidhya.com/blog/2015/09/naive-bayes-explained/

[21]. D. Kroese , J. Chan, "Generalized Linear Models," Springer,2013.
[22]. P. Lad, A. Somani, K.E. Krishnan, A. Gupta and V. Kartik," High-Throughput Shape Classification Using Support Vector Machine," IEEE.2016.
[23]. Confusion Matrix[Online].Available from: http://www2.cs.uregina.ca/~dbd/cs831/notes/confusion_matrix/confusion_matrix.html
[24]. R. Kumar and R.Verma ,"Classification Algorithm for data mining :A survey,"IJIET,2012.
[25]. G.Keseavaraj, S.Sukumaran,"Study on classification techniques on data mining," 4th ICCCNT ,IEEE, 2013.
[26]. M.Rathi,"Regression modeling technique o data mining for prediction," ICT ,Springer,2010.
[27]. S.Gupta,"A regression modeling technique on data mining. International journal of computer Application",2015 April.
[28]. D.Singh and A.Gosain ,"A comparative analysis of distributed clustering Algorithm : A survey," International symposium on computational Business Intelligence, IEEE,2013.
[29]. M. Hu and B.Liu,"Mining and summarizing customer reviews," KDD-04 tenth ACM SIGKDD International conference on knowledge discovery and data mining,ACM,2004.
[30]. Top 10 challenging problems in Data mining[Online].Available from: http://www.dataminingblog.com/top-10-challenging-problems-in-data-mining/
[31]. A.Kumar, AK. Tyagi and SK. Tyagi,"Data mining: Various issues and challenges for future," IJETA,2014
[32]. H.Nasereddin," NEW TECHNIQUE TO DEAL WITH DYNAMIC DATA MINING IN THE DATABASE," IJRRAS,.December 2012.
[33]. DK. Singh, V.Swaroop,"Data Security and Privacy in Data Mining: Research Issues & Preparation. International Journal of Computer Trends and Technology,"2013.
[34]. Shuang, Cong. "the Neural Network Theory and Application by Matlab Tool Box [M]." Hefei: Publishing Company of University of Science and Technology of China .
[35]. M.Hall, E.Frank , G.Holmes, B.Reutemann , IH Witten,"The WEKA Data Mining Software: An Update," SIGKDD Explorations,2009.
[36]. https://weka.wikispaces.com/Optimizing+parameters
[37]. J.Demšar and B.Zupan,"Orange: Data Mining Fruitful and Fun - A Historical Perspective",2012
[38]. M.Berthold, N.Cebron, F.Dill, T.Gabriel, T.Kotter, T.Meinl, P.Ohl, C.Sieb, K.Thiel and B.Wiswedel,"KNIME: The Konstanz Information Miner,"Springer,2008.
[39]. E.Loper and S.Bird ,"NLTK: The Natural Language Toolkit,"2002.
[40]. Z.Haofeng,"RapidMiner: A Data Mining Tool Based on Association Rules," Springer,2001.
[41]. A.Kusiak,"Rough set theory: A data mining tool for semiconductor manufacturing," JANUARY,2001.
[42]. J.Alcalá-Fdez,"KEEL: a software tool to assess evolutionary algorithms for data mining problems,"Springer,2008.
[43]. S.Christa, K.Madhuri, V Suma," A Comparative Analysis of Data Mining Tools in Agent Based Systems,"2010.
[44]. G.Smith , J.Whitehead, M.Mateas,"Tanagra: A Mixed-Initiative Level Design Tool,"ACM, 2010
[45]. R.Mikut and M.Reischl,"Data mining tools. Research gate,"2011.
[46]. Shelly,"Performance Analysis of various data mining classification Technique on healthcare data,"2011.
[47]. A.Wahbeh.,"A Comparison Study between Data Mining Tools over some Classification Methods," International Journal of Artificial Intelligence,2012
[48]. D.Jain,"A Comparison of Data Mining Tools using the implementation of C4.5 Algorithm ,"International Journal of Science and Research Vol3,2014.
[49]. Salma ,"Rule based complaint detection using Rapid Miner," RCOMM; 2013,Volume: 141 - 149,2013.
[50]. R.Arun and J.Tamilselvi,"Data Quality and the Performance of the Data Mining Tool",2015.
[51]. H.Odan, A.Daraiseh,"Open source Data Mining Tools," IEEE,2015.
[52]. C.Shah, A.Jivani,"Comparison of data mining classification algorithms for breast cancer prediction,"4th ICCCNT ,IEEE,2013.
[53]. P.Kakkar, A.Parashar," Comparison of different clustering Algorithm using WEKA tool," International Journal of Advanced Research in Technology, Engineering and Science, 2014.
[54]. S.Bavisi, Ȧ.J and L.Lopes,"A Comparative Study of Different Data Mining Algorithms,"International Journal of Current Engineering and Technology,2014
[55]. P.Gonc ¸ Jr. A, R.Barros and D.Vieira," On the use of data mining tools for Data preparation in classification problems," ACIS 11th International Conference on computer and information science ,IEEE ,2012.
[56]. N.Chauhan and N.Gautam," Parametric comparison of data mining tools," IJATES,2015.
[57]. A.Gupta, N.Chetty , S.Shukla,"A classification method to classify High Dimensional data",IEEE,2015.
[58]. M.Hassan , ME.Shahab , EMR.Hamed.,"A comparative study of classification algorithm in E-health Environment," IEEE.2016.
[59]. S.Singh, Y.Liu, W.Ding and Z.Li,"Evaluation of data mining tools for Telecommunication Monitoring Data using design of experiment," IEEE ,2016.
[60]. Information of dataset[Online].Available from https://archive.ics.uci.edu/ml/datasets/iris
[61]. WEKA dataset [Online]