ANALYSIS OF DATA MINING TECHNIQUES AND THEIR COMPARISONS

Yash Goel

ABSTRACT

Data mining is a method of extracting patterns from large datasets. This is a technique through which we discover data which is very important for the business. To extract data various types of tools has been developed. To grab this data These tools, offer an interface to extract data and to recover some attention-catching examples out of it that is any assistance to accomplish new information. There are types of boundaries laid out inside the writing which give a base for a device to perform investigation and totally various devices are open to play out this research. this is frequently very eye-catching to play out a near investigation of those instruments and to watch their conduct upheld some hand-picked boundaries which can any be helpful to look out the premier material device for the given data set and furthermore the boundaries.

1. INTRODUCTION

Data mining is the strategy for discovering designs from a lot of data by applying a few strategies. this is regularly utilized as an Associate in Nursing instrument for information disclosure in databases to be utilized in the dynamic procedure. Monstrous associations use it principally for finding new manners by which to stretch out their benefits and to lessen esteem. The information handling examinations the data and assists with naming the concealed elements all together that supportive examples and information will be produced. For an occurrence, business associations will break down the customer's conduct toward explicit item by examining the authentic information and this encourages the association to search out the dynamical conduct of the client with the entry of your time, as, to search out the patterns in change, to search out the level of adjustment and so forth. These sorts of findings are certainly facilitating 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 management of data, program, missing qualities, discovering error rate and a lot of extra boundaries make these devices totally unique in relation to each other. 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]. Partnerships purchased information preparing apparatus to cause their own to redo mining arrangements. A few information handling instruments square measure available with their qualities and constraints in setting to boundaries like interfaces, calculations, the precision of results, mining methods, information set size and so on. These instruments square measure any sorted into 3 classes for example Dashboards, Traditional information handling devices and Text Mining apparatuses. Customary information preparing apparatuses chiefly utilized by organizations for business examination reason. These apparatuses chip away at databases open with the corporate. Their instruments apply pre- characterized calculations on information for finding the undetectable example and results. These instruments give expansive information classes to concoct clear reports. As an occurrence, a data of deals will show month to month deals results and reports with the help of antiquated information preparing apparatuses. These devices square measure open each in Windows and usable framework variants of working frameworks and square measure essentially utilized for on-line Analytical procedure (OLAP)[4]. some of these devices square measure rail, R studio, quick worker, SQL and D2K [5]. Dashboards square measure put in on pc to watch data information and mirrors the updates and changes onscreen concerning business information and execution. 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.

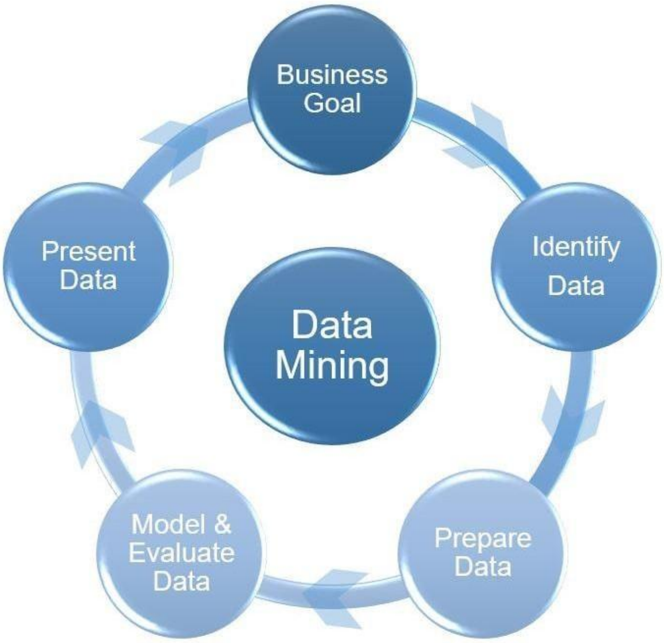

Fig 1 Datamining Operations

All details associated with benefits and loss of organization square measure obvious to the director on one screen interface and furthermore the entire assignment are performed by dashboard choices precisely. The main dashboards give the snap of the real execution of apparatuses and conjointly show the ongoing happenings [6]. The business knowledge dashboards are called undertaking dashboards [7]. These have the power to tug the important time knowledge from multiple sources. Oracle[6] and Microsoft[8] square measure among the main merchants of business insight dashboards[10]. Text mining is breaking down the content to remove data that will be useful for an unequivocal reason. It manages language text and lexical utilization to search out supportive data. Text mining instruments just access databases checked substance and grasp treatment of organized and unstructured information. 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 device which can deal with organized and unstructured data. There are some content mining instruments that square measure publicly released like orange [11], NLTK[12], Voyant[13] and Alchemy API[14]. IBM organization assemble more intelligent 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].

2. DATA MINING TECHNIQUES

There are a few procedures of information mining like order, relapse, bunching, outline that have their own attributes and restrictions. Classification [2] classifies information into completely different categories. There are several classification algorithms like call tree [19], Naive Bayes[20], Generalized linear Model [21] what's more, Support Vector Machine[22]. The characterization is performed basically on the possibility of boundaries for example precision and disarray network [23][24]. this framework gives differed applications inside the field of customer intrigue, interpersonal organization, clinical and medicinal services and a lot of extra [25]. Relapse [26] is utilized to plan the connection between 2 factors. This is regularly conjointly drawn inside the guide kind and perhaps acclimated check the outcome by assessment the hole of information focuses from relapse line[2]. Benefit, sq. the recording, temperature, deals and separation are normal through relapse. There are two equation's utilized for relapse insights for example Root Mean sq. Blunder (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 differed uses of the bundle inside the field of advancing, science, misrepresentation location, comparable land distinguishing proof [3]. In Summarization [2] one will make a minimal depiction of any data. Rundown is done inside the assortment of the table. The synopsis gives the connection between totally extraordinary kind of data sets [29]. There are two methodologies for programmed outline for example extraction and reflection. Extraction strategy deal with existing words, expressions or sentences inside the first content to make the layout. The hypothetical strategy utilizes normal language age methods [30][31].

3. 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 offers information for decision making [32]. The performance analysis in data processing tools is completed by completely different parameters. It offers 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 normal boundaries utilized for examination territory unit engineer, programming language, portability, interface, stage, visual picture, exactness and time took [34]. The estimations of those boundaries zone unit taken physically. There is a unit of some particular boundaries by and large information preparing apparatuses. For instance, rail containing boundaries for examination for example appropriately ordered examples, inaccurately arranged occasions, alphabetic character measurements, mean outright blunder, root mean square mistake, relative supreme mistake, root-relative square error [35]. The appropriately grouped examples offer information concerning the precision in arrangement of classes. The F-measure joins exactness and review mean. The precision tends to the norm or condition right or right worth of computation. Alphabetic character insights offer to quantify the multiclass and lopsided classification. It tells anyway your classifier performs with the information. Mean total mistake gauges the exactness of consistent factors. Root Mean square mistake gauges the basic extent of blunder [36]. The Orange device utilizes boundaries for examination like check and score. Check score offer exactness estimation through cross-approval. Second forecasts that show expectations of models for an info dataset.

Third disarray grid which offers information concerning classifier examination. Fourth, the ROC examination that shows the recipient in activity attributes bend upheld the investigation of classifier. Fifth, lift bend that develops and show the bend from the examination of classifier [37]. The MATLAB apparatus utilizes boundaries for investigation like precision, execution time and perception speed. The perception speed could be an unmistakable boundary for evaluation [34]. The Rapid digger device utilized boundaries like exactness, accuracy, Recall, AUC(Optimistic), AUC(neutral). The Rapid excavator also contain a presentation vector for foreseeing the exhibition esteems [40]. The KEEL devices have unmistakable alternatives off/on line run of investigation arrangement that is new in information handling devices. Another particular boundary zone unit arrangement or way examination and mistake rate [42].

4. DATA MINING TOOLS

4.1 WEKA

Weka is a Java based for the most part free and open source bundle authorized under GNU GPL and available to be utilized on Linux, Macintosh OS X and Windows. It contains a lot of AI calculations for information handling. It bundles instruments for data pre-preparing, order, relapse, bunching, affiliation rules and picture. Individual might be a simple graphical interface for two-dimensional picture of very much mined data. It empowers you to import the information from various record arrangements, and supports acknowledged calculations for different mining activities like separating, bunching, characterization and trait decision. Be that as it may, when dealing with monster data sets, it's ideal to utilize a CL based generally approach as individual attempts to stack the all out data set into the most memory, exacting execution issues. This bundle conjointly gives a Java Appetizer to be utilized in applications and can interface with databases utilizing CJD.

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 result in rule structuring algorithm. The three fundamental calculations like KNN, Naive Bayes and choice tree can rough a similar measure of time with the characterized set of boundaries. There is an extension to check the effectiveness of these calculations by taking new boundaries.

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]