Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 11, November-2018

Comparative Analysis of Various Tools for Data Mining and Big Data Mining

Ahamed Lebbe Sayeth Saabith
Centre for Information Communication Technology
Faculty of Science, Eastern University
Vantharoomulai, Sri Lanka

Mr. M M Mohamed Fareez
Finance Department
Eastern University Sri Lanka
Vantharoomulai, Sri Lanka

***Abstract*—Data mining and knowledge discovery has emerged to extract useful, interesting, and unknown patterns and knowledge from huge amount of database. Big data is the term used to delineate massive amounts of information of both structured and unstructured data types. Data mining techniques can be classified as classification, association, clustering, anomaly detection, regression analysis, prediction, and tracking patterns. Data mining tools which are helpful to achieve above data mining techniques. This research analysis various data mining and big data mining tools with different perspectives. This research will help for researchers to select appropriate data mining tool or tools for their research.**

Keywords—Big data; association; clustering; anomaly detection; regression

I. INTRODUCTION

Data mining is an essential step of knowledge discovery process by analyzing the varieties of data from miscellaneous perspectives and summarizing it into useful knowledge {1,7,12,13,14]. Data mining is widely used in various application domains such as Future Healthcare, Market Basket Analysis, Manufacturing Engineering, Education, Customer Relationship Management, Fraud Detection, Intrusion Detection, Lie Detection, Customer Segmentation, Financial Banking , Corporate Surveillance, Research Analysis, Criminal Investigation, Bio Informatics., and Science Exploration [7,8,13,15,16,25,31,32,35]. In today's digital world, we are surrounded with big data that is forecasted to grow 40% per year into the next decade. The data could be anything from a real time transaction, climatic situations, computers, and mobile logs, posts or tweets from social media and more and more. If the data is impossible keep in a single machine store and process, then such data could be named as **Big Data**. Data mining techniques can be classified as follows:

Classification: Classification is the most commonly used data mining technique which cover a set of pre classified samples to create a model that can classify the big data [7,8,9].

Association: This technique helps to find the association between two or more items. It helps to know the relations between the different variables in databases [17,18]

Clustering: Clustering is the division of data into groups of similar objects. Representing the data by fewer clusters necessarily loses certain fine details, but achieves simplification. Clustering can be viewed as a data modeling technique that provides for concise summaries of the data [2,3].

Anomaly Detection: Anomaly detection is defined as the process of finding the patterns in a given dataset whose behavior is abnormal or unexpected.

Regression Analysis: The process of identifying the relationship and the effects of this relationship on the outcome of future values of objects is called regression

Prediction: It discovers relationship between dependent and independent variables

A. Traditional and Big Data Mining Data Life Cycle

The traditional data mining life cycle can be categorized in two methodologies which are **CRISP-DM** and **SEMMA** methodology. The CRISP-DM methodology that stands for **Cr**oss **I**ndustry **S**tandard **P**rocess for Data Mining, is a cycle that depicts commonly used approaches that data mining experts use to tackle problems in traditional Business Intelligence data mining. CRISP-DM life cycle consist the components are Business Understanding, Data Understanding, Data reparation, Modeling, Evaluation, and Deployment. The following figure illustrates the CRISP-DM life cycle.

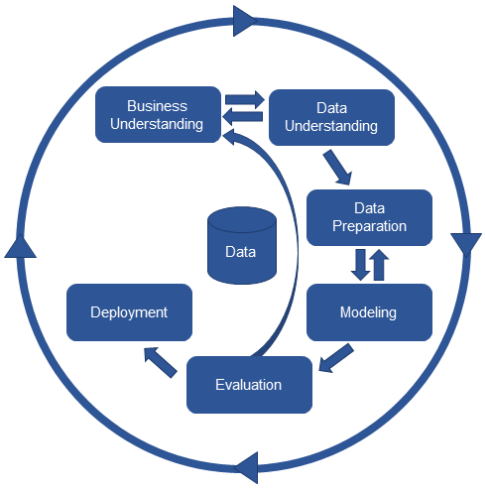

Figure 1: CRISP-DM

SEMMA is the another methodology was developed by SAS for data mining modeling that stands for **S**ample, **E**xplore, **M**odify, **M**odel, and **A**sses.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 11, November-2018

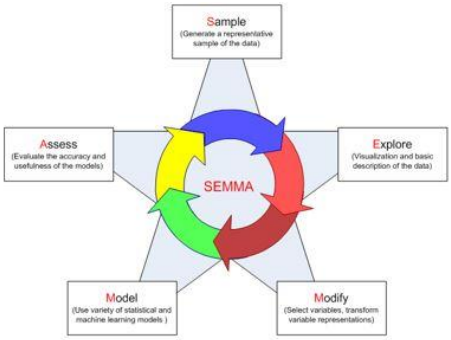

Figure 2: SEMMA

In today's big data context, the previous approaches are either incomplete or substandard. The Big Data analytics lifecycle can be divided into the following nine stages.

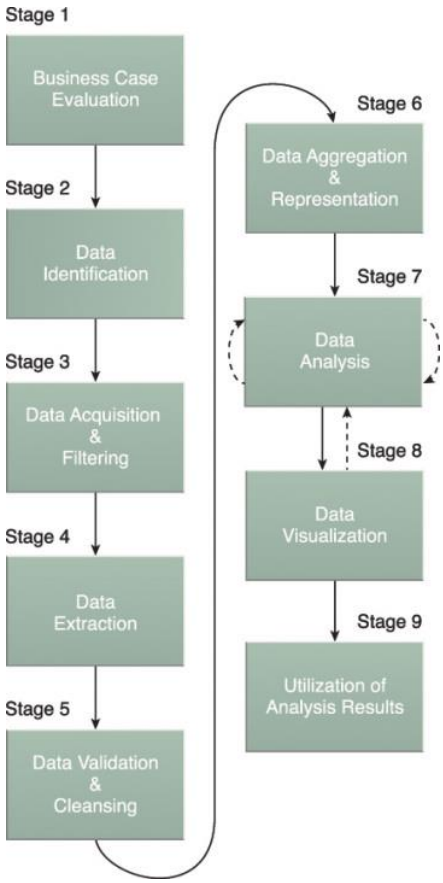

Figure 3: Big Data Mining Architecture

B. *Data Mining Tools*

The implementation of data mining techniques requires the use of powerful software tools. Today number of data mining tools are available with different categories, the choice of the most suitable tool becomes increasingly difficult. This paper attempts to survey the availability of the traditional data mining and big data mining software in a several categorizations such as commercial and open source, business size, platform, deployments, data mining tasks and methods, and visualization.

II. LITERATURE REVIEW

A comparative analysis of data mining tools and to observe their behavior based on some selected parameters which will further be helpful to find the most appropriate tool for the given data set and the parameters [1]. M.Hall et al, expressed the importance of WEKA tool which is an open source implemented in Java language. WEKA is used for implementing the most of the data mining techniques [2]. In [3], this research focused on comparison of various data mining tools based on traditional data mining tools, dashboards, text mining, and standalone application [3]. This study compared four open source Data Mining tools which are KNIME, Orange, Rapid Miner and Weka. The research objective is to reveal the most accurate tool and technique for the classification task. Analysts may use the results to rapidly achieve a good result [4]. In this study, various frequently used open-source data mining tools and tools with open-source algorithms implementations are selected and compared against user groups, data structures, algorithms included, visualization capabilities, platforms, programming languages, and import and export options. In addition, evaluation of publicly available datasets has been performed by using selected tools [5]. Wang et al. (2008) in their comparison of leading data mining software packages, compared them against several software different ways, such as portability, reliability, efficiency, human engineering, understanding, modifiability, price, training and support [6]

III. METHODOLOGY

A. *Traditional Open Source Data Mining Tools*

- ***Orange:*** Orange is an opensource tool for data analysis and visualization. Data minng is done through python or visual programming which has components for machine learning feature selection, and text mining
- ***R***: R is free open source software programming language and software environment for statistical computing and graphics. The R language is widely used among data miners for developing statistical software and data analysis. Ease of use and extensibility has raised R's popularity substantially in recent years.
- ***Weka:*** Weka, an open source data mining software, is a collection of machine learning algorithms for data mining tasks such as Data Pre – Processing, Data Classification, Data Regression Data Clustering, Data Association Rules, and Data Visualization. The algorithms can either be applied directly to a data set or called from your own JAVA code.
- ***Shogun:*** Shogun is a free open source software toolbox written in C++. It offers lots of algorithms, and data structure for machine learning problems. The Shogun focus on Support Vector Machine (SVM), regression, and classification data mining problems.
- ***RapidMiner:*** RapidMiner operates through visual programming and is capable of manipulating, analyzing and modeling data. RapidMiner makes

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 11, November-2018

data science teams more productive through an open source platform for data preparation, machine learning, deep learning, text mining, predictive analytics and model deployment.

- ***TANAGRA:*** Tanagra is one of the free open source software for academic and research purposes which is provides various data mining methods from exploratory data analysis, statistical data mining, machine learning, and deep learning.
- ***ELKI:*** This is an open source (AGPLv3) data mining software written in Java. The focus of ELKI is research in algorithms, with an emphasis on unsupervised methods in cluster analysis and outlier detection. In order to achieve high performance and scalability, ELKI offers data index structures such as the R*-tree that can provide major performance gains.

Table 1: Traditional Open Source Data Mining Tools Comparison

Software	URL	Business Size	Platform	Category	Data visualization	Language
Orange	https://orange.biolab.si	S, M, L	Windows, macOS, Linux	Data mining	Yes	Python
R	https://www.r-project.org/	S, M, L	Windows, macOS, Linux	Predictive Analysis	Yes	R programming
Weka	https://www.cs.waikato.ac.nz/ml/index.html	S, M, L	Windows, macOS, Linux	Data mining	Yes	Java
Shogun	http://shogun-toolbox.org/	S, M, L	Windows, macOS, Linux	Data Analysis	No	C++
DATAMELT	https://jwork.org/dmelt/	S, M, L	Windows, macOS, Linux, Android	Data Mining, Data Analysis	Yes	Java
RapidMiner	https://rapidminer.com/	S, M, L		Predictive Analysis	Yes	Java
TANAGRA	http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html	S, M, L	Windows	Data Mining	Yes	
ELKI	https://elki-project.github.io/	S, M, L	Linux	Most Data mining Techniques	Yes	Java

B. Traditional Commercial Data Mining Tools

- ***Sisence:*** Sisense is a business intelligence platform that lets you join, analyze, and picture out information they require to make better and more intelligent business decisions and craft out workable plans and strategies.
- ***Neural Designer***: This is a desktop application for data mining which is uses neural network and machine learning
- ***SharePoint:*** SharePoint is a Microsoft-hosted cloud service that empowers companies to store, access, share, and manage documented information from all devices.
- ***Cognos:*** IBM Cognos is a set of smart self-service capabilities that enable them to quickly and confidently determine and make decisions on insight. The engaging experience provided by Cognos Analytics encourages business users to make and/or configure dashboards and reports on their own – while providing IT with a proven and scalable platform that can be deployed either on premises or in cloud.
- ***Borad:*** Board is a Management Intelligence Toolkit that combines compact software. BOARD enables users to collect and gather data from almost any source, as well as create full self-service reporting. These reports can be delivered in different formats if needed, like CSV, HTML and more. Features of business intelligence (BI) and corporate performance management (CPM) into a comprehensive and compact software.

C. Big Data Mining Tools

- ***Sisence:*** Sisense is a business intelligence platform that lets you join, analyze, and picture out information they require to make better and more intelligent business decisions and craft out workable plans and strategies.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 11, November-2018

Table 2: Traditional Commercial Data Mining Tools Comparison

Software	URL	Business Size	Features		Category	Data visualization	Free Trial
Sisense	https://www.sisense.com/?utm_medium=cpc&utm_source=financesonline&utm_campaign	S, M, L	• Ad-hoc analysis • In-house set up • Centralized data hub • Non-programming/SQL writing business query • Data collection, filtering, consolidation & storage • Data connectors • Data export to various formats • Scalable data handling • Scalable analytics • Drag-and-drop functionality • No restrictions on data size • Embeddable widgets & dashboards • Widgets library • Apps & sites integration • Single-Sign-On • Visualizations • Metrics identification	Windows Linux Android iPhone/iPad Mac Web-based	Business Intelligence	Yes	Yes 30 Days
SharePoint	https://products.office.com/en-us/sharepoint/collaboration	S, M, L	• Access Services • Compliance • Customized Web Parts • Library • Durable Links • Encrypted Connections • Fast Site Collection Creation • Identification of Sensitive Content • Information Rights Managements • Image/Video Preview • Mini Roles • Large File Support • Mobile Support • OneDrive Control • Business Intelligence • Search • Site Pages Pinning • SMTP Ports • SMTP Encryptions • WOPI • Side Folders View	Windows Android iPhone/iPad Mac Web-based Windows Mobile	Business Intelligence	Yes	No

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 11, November-2018

Cognos	https://www.cs.waikato.ac.nz/ml/index.html	M, L	• Smart search works in context • Personalized experience • Scheduling and alerts • Interactive content available online or offline • A complete web-based experience • Easy upload of personal/external data • Report directly off a data source • Effortlessly combine data sources • Data models can be automatically generated based on keywords • Dashboards created using drag and drop on mobile device or desktop • Best automatic visualizations	Windows Linux Mac Web-based	BI	Yes	Yes
Neural Designer	https://www.neuraldesigner.com/	S, M, L	• High Performance Computing • Easy to use • Advanced Analytics • Unlimited network architecture • Pre-analysis tools • Neural network equation exporting • Many different training algorithms • Extensive testing analysis method	Desktop Cloud Server	Neural Network, and Machine Learning	Yes	Yes up to 5000 instances
Board	https://jwork.org/dmelt/	S, M, L	• Data discovery and analysis • Planning • Simulation • Reporting • Dash boarding • Predictive and advanced analytics • Score carding • Mobile • MS Office integration	Windows Android iPhone/iPad Web-based Windows Mobile	Data Mining, Data Analysis	Yes	No

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 11, November-2018

Table 3: Big Data Mining Tools Comparison

Software	URL	Business Size	Deployment	Big Data Features	Free Version
SPSS	https://www.ibm.com/analytics/spss-statistics-software	S, M, L	Installed - Windows Linux/Unix	• Collaboration • Data Mining • Predictive Analysis	No
MangoDB	https://www.mongodb.com/	S, M, L	Cloud, SaaS, Web Installed - Mac Installed - Windows	• Data Visualization • Data Warehousing • High Volume Processing	Yes
ElasticSerach	https://www.cs.waikato.ac.nz/ml/index.html	S, M, L	Cloud, SaaS, Web	• No-Code Sandbox • Predictive Analytics • Templates • Data Visualization	Free Trial
Cyfe	https://www.cyfe.com/	S, M	Cloud, SaaS, Web	• Data Visualization	Yes
SAP HANA	https://www.sap.com/products/hana.html	L	Cloud, SaaS, Web	• Data Mining • Predictive Analysis • Data Warehousing	Yes
DATA HERO	https://datahero.com/	S, M, L	Cloud, SaaS, Web Mobile - Android Native Mobile - iOS Native	• Collaboration • Data Blends • Data Cleansing • Data Mining • Data Visualization • Data Warehousing • High Volume Processing • No-Code Sandbox • Predictive Analytics • Templates	Free Trial
Cloudera Enterprise	https://www.cloudera.com/products.html	S, M, L	Windows	Yes	No

II. CONCLUSION

This study compared the Traditional free datamining tools described in the different perspectives such as business size, category, platform, data visualization, and which language used for developed the tools, Traditional Commercial Data Mining tools, and Big data mining. Traditional free datamining tools described in the different perspectives such as business size, category, platform, data visualization, and which language used for developed the tools. Traditional Commercial Data Mining tools described in the different perspectives such as Official URL of the tools, business size, features, category, data visualization, and whether free trial available or not. Big Data Mining tools described in the different perspectives such as Official URL of the tools, business size, deployment, what are the big data features exist, and whether free version available or not. This study will help to choose correct data mining tools for upcoming researchers who are going to do the research under the data mining and machine learning.

III. REFERENCES

[1] Dr. Anil Sharma, Balrajpreet Kaur, A RESEARCH REVIEW ON COMPARATIVE ANALYSIS OF DATA MINING TOOLS, TECHNIQUES AND PARAMETERS, International Journal of Advanced Research in Computer Science, Volume 8, No. 7, July – August 2017

[2] M.Hall, E.Frank , G.Holmes, B.Reutemann , IH Witten,"The WEKA Data Mining Software: An Update," SIGKDD Explorations,2009.

[3] Mrs. Parminder Kaur, Dr. Qamar Parvez Rana, Comparison of Various Tools for Data Mining, International Journal of Engineering Research & Technology (IJERT) , Volume 3, Issue 10 - 2014

[4] Luís C. Borges , Viriato M. Marques , Jorge Bernardino Comparison of data mining techniques and tools for data classification, C3S2E '13 Proceedings of the International C* Conference on Computer Science and Software Engineering Pages 113-116

[5] Dakić Dušanka et al, A Comparison of Contemporary Data Mining Tools, XVII International Scientific Conference on Industrial Systems (IS'17), Novi Sad, Serbia, October 4. – 6. 2017.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 11, November-2018

[6] Wang J, Hu X, Hollister K, Zhu D. (2008) "A comparison and scenario analysis of leading data mining software". Int J Knowl Manage 2008, 4:17–34.
[7] M. Antonie, A. Coman, and O. R. Zaiane, “Application of Data Mining Techniques for Medical Image Classification,” in Proceedings of the second international Workshop on Multimida Data Mining (MDM/KDD’2001), 2001, pp. 94–101.
[8] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan, “Web usage mining: Discovery and applications of usage patterns from web data,” Acm Sigkdd …, vol. 1, no. 2, pp. 12–23, 2000.
[9] J. Han and M. Kamber, Data Mining, Southeast Asia Edition: Concepts and Techniques. Morgan kaufmann, 2006.
[10] J. Hipp, U. Güntzer, and G. Nakhaeizadeh, “Algorithms for association rule mining - a general survey and comparison,” ACM SIGKDD Explor. Newsl., vol. 2, no. 1, pp. 58–64, 2000.
[11] C. Zhang and S. Zhang, Association rule mining: models and algorithms, vol. 2307. Springer-Verlag, 2002.
[12] R. Agrawal, T. Imieliński, and A. Swami, “Mining association rules between sets of items in large databases,” ACM SIGMOD Rec., 1993.
[13] R. Agrawal and R. Srikant, “Fast algorithms for mining association rules,” Proc. 20th int. conf. very large data bases, VLDB, 1994.
[14] I. Witten and E. Frank, Data Mining: Practical machine learning tools and techniques. 2005.
[15] B. Ambulkar and V. Borkar, “Data Mining in Cloud Computing,” in MPGI National Multi Conference, 2012, pp. 23–26.
[16] R. S. Petre, “Data mining in cloud computing,” Database Syst. J., vol. 3, no. 3, pp. 67–71, 2012.
[17] M. J. Zaki, “Scalable algorithms for association mining,” Knowl. Data Eng. IEEE Trans., vol. 12, no. 3, pp. 372–390, 2000.
[18] M. J. Zaki and K. Gouda, “Fast vertical mining using diffsets,” in Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’03, 2003, p. 326.
[19] J. Han, J. Pei, Y. Yin, and R. Mao, “Mining frequent patterns without candidate generation: A frequent-pattern tree approach,” Data Min. Knowl. Discov., vol. 8, no. 1, pp. 53–87, 2004.
[20] C. Borgelt, “Keeping things simple: Finding frequent item sets by recursive elimination,” in Proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations, 2005, pp. 66–70.
[21] Z.-H. Deng and S.-L. S. Lv, “Fast mining frequent itemsets using Nodesets,” Expert Syst. Appl., vol. 41, no. 10, pp. 4505–4512, 2014.
[22] T. Krishna, “Effectiveness of various FPM Algorithms in Data Mining,” ijcsit.org, vol. 02, no. 01, pp. 01–05, 2014.
[23] S. Patel Tushar, P. Mayur, L. Dhara, K. Jahnvi, D. Piyusha, P. Ashish, P. Reecha, S. P. Tushar, P. Mayur, and L. Dhara, “An Analytical Study of Various Frequent Itemset Mining Algorithms,” Res. J. Comput. Inf. Technol. Sci., vol. 1, no. 1, pp. 2–5, 2013.
[24] S. Pramod and O. P. Vyas, “Survey on frequent itemset mining algorithms,” Int. J. Comput. Appl., vol. 1, no. 5, 2010.
[25] P. Prithiviraj and R. Porkodi, “A Comparative Analysis of Association Rule Mining Algorithms in Data Mining: A Study,” Open J. Comput. Sci. Eng. Surv., vol. 3, no. 1, pp. 98–119, 2015.
[26] M. Tiwari, M. B. Jha, and O. Yadav, “Performance analysis of Data Mining algorithms in Weka,” IOSR J. Comput. Eng. ISSN, pp. 661–2278, 2012.
[27] M. M. Trivedi, “REVIEW AND ANALYSIS OF VARIOUS EFFICIENT FREQUENT PATTERN ALGORITHMS,” Int. J. Technol. Res. Eng., vol. 2, no. 2, pp. 139–143, 2014.
[28] K. Garg and D. Kumar, “Comparing the Performance of Frequent Pattern Mining Algorithms,” Int. J. Comput. Appl., vol. 69, no. 25, pp. 21–28, 2013.
[29] G. Sinha and S. M. Ghosh, “Identification of Best Algorithm in Association Rule Mining Based on Performance,” Int. J. Comput. Sci. Mob. Comput., vol. 3, no. 11, pp. 38–45, 2014.
[30] M. B. Nichol, T. K. Knight, T. Dow, G. Wygant, G. Borok, O. Hauch, and R. O’Connor, “Quality of anticoagulation monitoring in nonvalvular atrial fibrillation patients: Comparison of anticoagulation clinic versus usual care,” in Annals of Pharmacotherapy, 2008, vol. 42, no. 1, pp. 62–70.
[31] L. C. Yu, C. L. Chan, C. C. Lin, and I. C. Lin, “Mining association language patterns using a distributional semantic model for negative life event classification,” in Journal of Biomedical Informatics, 2011, vol. 44, no. 4, pp. 509–518.
[32] Q. Zhao and S. S. Bhowmick, “Association rule mining: A survey,” Nanyang Technol. Univ. Singapore, 2003.
[33] A. M. Said, P. D. D. Dominic, and A. B. Abdullah, “A comparative study of fp-growth variations,” Int. J. Comput. Sci. Netw. Secur., vol. 9, no. 5, pp. 266–272, 2009.
[34] J. Han, J. Pei, and Y. Yin, “Mining frequent patterns without candidate generation,” in ACM SIGMOD Record, 2000, vol. 29, no. 2, pp. 1–12.
[35] O. R. Zaiane, M. El-Hajj, and P. Lu, “Fast parallel association rule mining without candidacy generation,” in Proceedings 2001 IEEE International Conference on Data Mining, 2001, pp. 665–668.
[36] C. Borgelt, C. Borgelt, R. Kruse, and R. Kruse, “Induction of Association Rules: Apriori Implementation,” in 15th Conference on Computational Statistics Physica Verlag, Heidelberg, Germany 2002, 2002, vol. 1, pp. 1–6.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)