International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 04 | Apr 2019 www.irjet.net p-ISSN: 2395-0072

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

**Mrs. G. SangeethaLakshmi[1], Ms. M. Jayashree[2]**

*[1]Asst Prof,Department of Computer science and Application, DKM College for Women (Autonomous), Vellore.*
*[2]Research scholar, Department of Computer Science, DKM College for Women (Autonomous), Vellore, TamilNadu.*

---

**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 datamining and big data mining tools with different perspectives.This research will help for researchers to select appropriate datamining tool or tools for their research.

*Keywords—Big data; association; clustering; anomalyDetection.*

## 1. INTRODUCTION:

**Data mining involves six common classes:**

Anomaly detection (outlier/change/deviation detection) – The identification of unusual data records, that might be interesting or data errors that require further investigation.

Association rule learning (dependency modelling) – Searches for relationships between variables. For example, a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes.

This is sometimes referred to as market basket analysis.

Clustering – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.

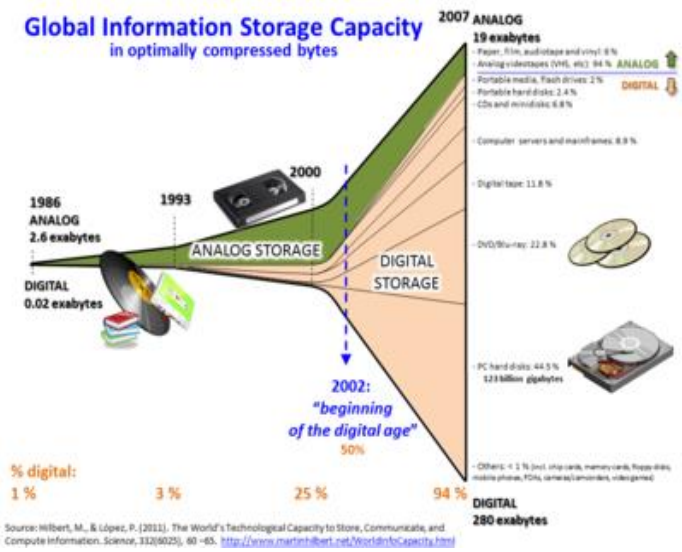
Classification – is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".

Regression – attempts to find a function which models the data with the least error that is, for estimating the relationships among data or datasets.

Summarization – providing a more compact representation of the data set, including visualization and report generation.

The rapid and inevitable development of technology is causing a substantial global increase in the volume of data. Such data mean better information, and information is wealth. This is because information makes it possible for mankind to have a safer and better future, which is the primary goal of scientists and researchers. Due to this incredible amount of information that can be obtained from Big Data, humanity is able to make considerable progress in diverse fields ranging from health and safety to education and economy.

The analysis and modeling of big data are not new subjects for actuaries, bankers, and insurers; DM helps them overcome many difficulties in their aim to manage money more effectively, control the system, reduce or transfer potential risks, understand client requirements, improve funds management, increase market share, and reduce or transfer potential risks . Specifically, DM can be used in the banking and insurance industries to determine default risks and risk groups, specify the correct insurance options for individual customers, increase customer satisfaction, and identify credit card fraud.




**Big data can be described by the following characteristics:**

**Volume-**The quantity of generated and stored data. The size of the data determines the value and potential insight, and whether it can be considered big data or not.

**Variety-**The type and nature of the data. This helps people who analyze it to effectively use the resulting insight. Big data draws from text, images, audio, video; plus it completes missing pieces through data fusion.

**Velocity-**In this context, the speed at which the data is generated and processed to meet the demands and challenges that lie in the path of growth and development. Big data is often available in real-time. Compared to small data, big data are produced more continually. Two kinds of velocity related to big data are the frequency of generation and the frequency of handling, recording, and publishing.

**Veracity-**It is the extended definition for big data, which refers to the data quality and the data value.The data quality of captured data can vary greatly, affecting the accurate analysis.

Data must be processed with advanced tools (analytics and algorithms) to reveal meaningful information. For example, to manage a factory one must consider both visible and invisible issues with various components. Information generation algorithms must detect and address invisible issues such as machine degradation, component wear, etc. on the factory floor.

## 2. 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. 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. In this research focused on comparison of various data mining tools based on traditional data mining tools, dashboards, text mining, and standalone application. 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 . 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.

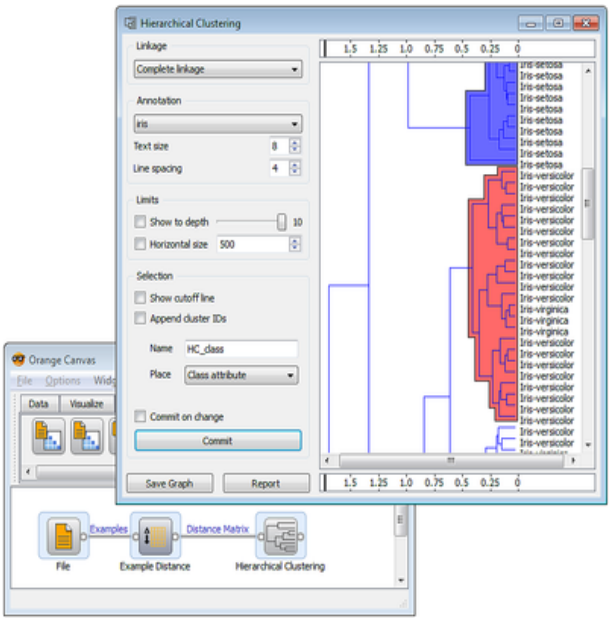
In addition, evaluation of publicly available datasets has been performed by using selected tools . Wang et al. (2008) in their comparison of leading data mining software packages, compared them against several software different ways, such as portability, reliability, efficiency, human engineering, understanding, modifiability, price, training and support.

## 3. METHODOLOGY

### A. TRADITIONAL OPEN SOURCE DATA MINING TOOLS

***Orange:*** Orange is an open source tool for data analysis and visualization. Data mining is done through python or visual programming which has components for machine learning feature selection, and text mining. Python is picking up in popularity because it's simple and easy to learn yet powerful.

Hence, when it comes to looking for a tool for your work and you are a Python developer, look no further than Orange, a Python-based, powerful and open source tool for both novices and experts.




***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 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.

Besides data mining it provides statistical and graphical techniques, including linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, and others.In addition to data mining, RapidMiner also provides functionality like data preprocessing and visualization, predictive analytics and statistical modeling, evaluation, and deployment. What makes it even more powerful is that it provides learning schemes, models and algorithms from WEKA and R scripts.

***Weka:*** Weka, 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.

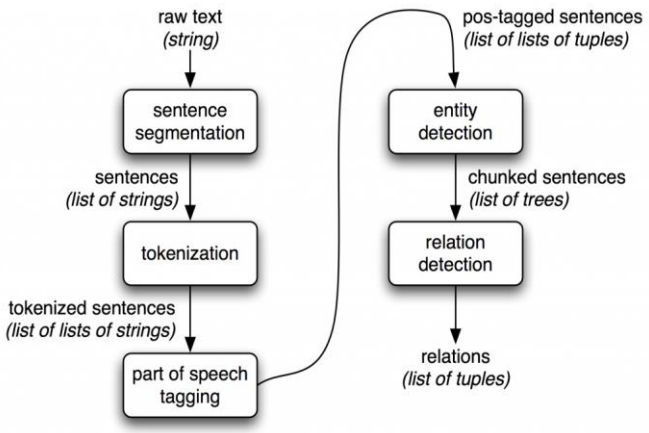
***Rapid Miner:*** Rapid Miner operates through visual programming and is capable of manipulating, analyzing and modelling data.

Data preprocessing has three main components: extraction, transformation and loading. KNIME does all three. It gives you a graphical user interface to allow for the assembly of nodes for data processing. It is an open source data analytics, reporting and integration platform. KNIME also integrates various components for machine learning and data mining through its modular data pipelining concept and has caught the eye of business intelligence and financial data analysis.

Written in Java and based on Eclipse, KNIME is easy to extend and to add plugins. Additional functionalities can be added on the go. Plenty of data integration modules are already included in the core version.

The algorithms that have similar accuracy rates were compared again with different statistical criteria such as ROC (receiver operating characteristic), precision, recall, F-measure, and the root mean squared error (RMSE) to achieve the best results. As a result, the most appropriate algorithm for this dataset is found as the logistic regression algorithm.

The aim of this study is to use DM classification algorithms to investigate the effects of certain demographic and socioeconomic characteristics on the probability of individuals' default risk, as well as to predict their future payment challenges by determining individual attributes using a logistic regression classification algorithm.



## 4. 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 fordata 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-servicecapabilities that enable them to quickly andconfidently 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 createfull 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.

## 5. 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.

**KEEL:**KEEL stands for "Knowledge Extraction based on Evolutionary Learning," and it aims to help uses assess evolutionary algorithms for data mining problems like regression, classification, clustering and pattern mining. It includes a large collection of existing algorithms that it uses to compare and with new algorithms. Operating System: OS Independent.

**MAHOUT:**This Apache project offers algorithms for clustering, classification and batch-based collaborative filtering that run on top of Hadoop. The project's goal is to build scalable machine learning libraries. Operating System: OS Independent.

## 6. CONCLUSION:

This study compared the Traditional free data mining 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 data mining 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.

## REFERENCES:

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

[2] M.Hall, E.Frank, G.Holmes, B.Reutemann , IH Witten,"TheWEKA 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.

[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 andalgorithms, vol. 2307. Springer-Verlag, 2002.

[12] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," ACM SIGMODRec., 1993.