ISSN: 0971-1260
Vol-22-Issue-3-July-2019

Big Data Mining and Tools: A Comparative Analysis

Surbhi,
School of Computer Science & Engineering, Lovely Professional University, Phagwara, India
surbhi.23540@lpu.co.in

Balraj Singh,
School of Computer Science & Engineering, Lovely Professional University, Phagwara, India
balraj.13075@lpu.co.in

Shivraj Puggal,
School of Mechanical Engineering, Lovely Professional University, Phagwara, India
shivraj.16125@lpu.co.in

Abstract- With the development in information, it is primary to extract knowledge from the huge data archives. Subsequently, Data Mining has become a crucial element to examine this huge measure of data and describe the productive ends. With the help of various data mining tools, the process of finding relationships and patterns in the data can be automated. These results can be then used in decision support system. This paper talks about big data, its features, big data mining, its points of interest and issues identified. This paper likewise gives the far reaching and hypothetical investigation of five data mining tools and also characterizes the technical highlights, advantages and constraints for each data mining tool.

Keywords: Big Data; Data Mining Tools; Weka; KNIME; KEEL; RapidMiner

1. INTRODUCTION

There has been a major development in measure of data and information which is put away in electronic format since recent couple of years. The size of database has been currently incrementing and has come to terabytes. This monstrous pace of information augmentation is beyond the estimations. Along these lines the primary issue with data is recovery of data which finds its answer in this process. Data mining is used for extracting helpful data from huge masses of data [1].

Data mining is concerned with data science which involves control and request for data by applying real and numerical speculations. Because of the accessibility of diverse, immense, rich datasets, the capacity to divide significant information hidden in it , data mining is essential.Thus data mining is investigating the

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

enormous observational datasets to discover un-anticipated connections and to outline the information in the manners that are both comprehensible and helpful to information proprietor. [61]

This paper is categorized into five sections. First section describes Big Data and next one construe Data mining definition, advantage and its issues. In one of the sections the focus is on the open source tools for data mining. The last section explains comparative study of different tools and final section is the conclusion.

2. BIG DATA

Big data is a trademark which is used to characterize a tremendous amount of unstructured and structured data that is colossal to the point that it is difficult to utilize customary programming strategies. In these enterprise arrangements the data is more or it moves quickly or it surpasses handling limits. Big data can possibly support organizations, increment activities and make increasingly canny, faster choices. The term big data refers to the innovation that associations use to deal with a lot of data [2].

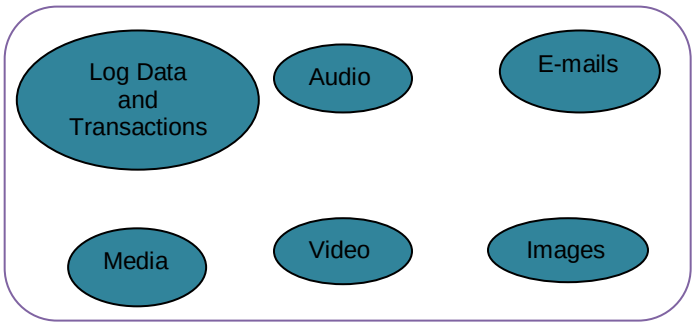

Figure I: Big Data Sources [2]

The 3V's are:

Volume: The measure of information. Maybe this trademark is for the most part connected with enormous information. Volume alludes to the massive amount of information that associations are attempting to harness to improve basic leadership over the endeavor. Data volumes keep on ascending at a remarkable rate. Volume of information continues changing with the time.

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

Variety: Various kinds of information and information sources. Variety is tied in to dealing with the multifaceted nature of various information types including structured, semi-structured and unstructured information. With the explosion of gadgets and social collaboration advancements, information is being generated in different forms including content, web information, tweets, sound, video, log documents, and that's just the beginning.

Velocity: Data in motion. The speed, at which data is produced, processed and analyzed.

There are two more V's nowadays:

Variability: The change in data structures and how user understands data.

Value: The worth is additionally changing step by step with the mind blowing development of information.

3. DATA MINING FOR BIG DATA

3.1 Definition of Data Mining

Data Mining speaks to a procedure created to look at a lot of information routinely gathered. The term likewise refers to the collection of all the devices used to play out this procedure. Information gathered from different regions for example, advertising, wellbeing and correspondence are utilized in data mining. This finds the concealed examples, predicts the data that encourages with arrangement outside their expectations [3]. The objective of data mining is to take data from datasets in human-justifiable structures.

Data Mining is used for the specific classes of tasks [2]:

1) Classification: Classification is a procedure of disentangling the information as per various cases. The classification task is described by the well characterized classes and a training set of renamed occurrences.

2) Estimation: It manages endlessly important results. Given some information that we go through to estimate the increment for some uncertain constant factors for example salary, tallness or Master card balance.

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

3) Prediction: It is an articulation about the manner in which things will happen later on frequently yet not constantly found information. It might be some statement where the result is normal.

4) Association Rules: It is a rule which proposes certain afflictive connections among a lot of objects in a database.

5) Clustering: It can be considered the most significant problematic learning issue. As each other issue of this sort, it manages to find a structure in the group of unlabeled data.

3.2 Data Mining advantages

Advantages of data mining in different applications [4]:

1) Banking: It supports banking sector in the procedure of looking through an enormous database to find once in the past unknown examples; program the way toward finding expository data. Data mining gauges levels of advances and phony master cards use, anticipating master card spending by new clients.

2) Production: It helps in estimating the key disappointments and to find the key factors for controlling the streamlining of assembling limit.

3) Marketing: It quickens promoting area by naming client statistic, utilized to anticipate which client will answer to mailing or buy a specific item and it is particularly steady in business development.

4) Computer hardware and software: It is used for foreseeing disk failures and potential security destructions.

3.3 Issues in Data mining

It is not at all smooth. These algorithms are compliscated. The data is loaded from the heterogeneous sources. These components create some problems.

The Various issues are [16]:

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

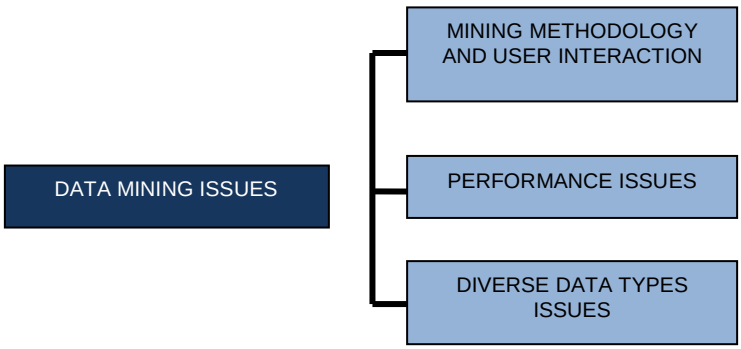

Figure II: Data Mining Issues

1) Mining methodology and User interaction Issues

- Mining heterogeneous knowledge in database
- Interactive knowledge mining at multiple abstraction levels
- Ad-hoc mining and data mining query language
- Picturing and presentation

2) Performance Issues

- Proficiency and Scalability
- Incremental, distributed and parallel

3) Diverse data type Issues

- Management of complex and relational data
- Data mining from diverse information systems

4. OPEN SOURCE TOOLS FOR DATA MINING

Data mining has a lot of usages stretching from displaying of product or things, man-made thinking assessment, natural sciences, bad behavior assessments to raised level government information. As a result of its expansive use and complexity associated with building data mining applications, data mining tools have been made over decades. Each tool has its own focuses and obstructions. [7].

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

Data mining supplies various mining techniques to remove data from the databases. These tools predict future examples, works on, empowering business to make active and data driven decisions. The improvement and usage of data mining counts requires use of powerful programming tools. As the amount of available mechanical assemblies continues propelling the choice to find the most sensible tools ends up being logically inconvenient [8].

The top five tools present for data mining are as below:

4.1 WEKA

It is an assortment of machine learning algorithms for data mining assignments. These computations can either be applied legitimately to data collection or it can be called from Java code. It contains a collection of few tools for perception and calculations of information and displaying, together with GUIs for simple access.

1) Technical

- Released on 1997
- Last version is WEKA 3.6.11
- GNU
- Platform independent.
- Java Programming

2) Advantages

- Development of different machine learning techniques [9].
- It loads data file of different formats. It is free, open source, extensible and can be incorporated into other packages of java.

3) Limitations

- Worse connectivity to Non-Java databases and Excel spreadsheet.
- Much weaker in classical statistics.

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

- No scaling facility for parameters
- Parameter optimization cannot be done automatically.

4.2 KEEL

It is an application foundation of machine learning programming devices. It has a collection of libraries for preprocessing and post-handling procedures for data controlling, soft computing strategies in knowledge of removing and learning, and giving logical and examining techniques.

1) Technical

- Released on 2004
- Latest version is KEEL 2.0
- GNU
- Platform Independent
- Supported by java language.

2) Advantages

- It includes grouping, regression, clustering and pattern matching.
- It contains a collection of hybrid algorithms, knowledge extraction models and preprocessing methods [10].

3) Limitations

- Efficiency is restricted by supported algorithms.

4.3 R

It is utilized among analysts and information miners for mounting measurable programming and information investigation. One of the qualities of R is the straightforwardness with which production of quality plots can be done, like scientific images and formulas where required.

1) Technical

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

- First released on 1997
- Latest version is 3.1.0
- GNU
- CrossPlatform
- C, Fortran and R

2) Advantages

- Able to create a machine learning program
- Numerical programming is better
- Better graphics.
- Importing and exporting of data from spreadsheet is easy

3) Limitations

- Less specialized towards data mining.
- Sheer learning curve, unless accustomed with array languages

4.4 KNIME

It is an integration, open source data analysis and reporting platform. It depends on Eclipse and through its particular APIs it is effectively extensible.

1) Technical

- Latest version is KNIME2.9
- GNU
- Compatible with Linux and Windows
- Java.

2) Advantages

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

- It joins all examination modules of the notable. Weka information mining condition and extra modules permit R-contents to run; offering access to immense libraries of factual schedules [9].
- No installation required.
- Compatible with data visualization and analysis programs.

3)Limitations

- Error measurement methods are less.
- There is no descriptor selection method.
- No automatic parameter optimization.

4.5 RAPIDMINER

It is a tool created by the organization of a similar name that gives an incorporated situation to machine learning, data mining, prescient investigation, content mining and business examination. RapidMiner utilizes a customer-Server model with the server offered as SAAS.

1) Technical

- Latest Version is RapidMiner 6.
- AGPL
- CrossPlatform
- Language Independent.

2) Advantages

- Full facility for model evaluation.
- Offers numerous procedures

3) Limitations

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

- RapidMiner is a data mining software package that is suitable for individuals, comfortable to work with database documents, for example, in scholarly or business settings. The product requires the capacity to control Sql articulations and records.
- If RapidMiner Studio is actualized in UI mode and an Operator gets an Example Set which beats the size of the permitted information pushes at that point Process will be halted and a comparing data air pocket will be shown. Right now you can either down example your information or redesign your present permit.

- **COMPARATIVE STUDY**

The best five data mining instruments were picked and expository examination was made by considering specialized particulars and feature [12] [13] [14] [15].

Table I: Technical Overview

Name	Release	Latest Release	Version	License	Operating System	Programming Language
RapidMiner	2006	21/11/2013	6.0	AGPL	CrossPlatform	Platform Independent
Knime	2004	6/12/2013	2.9	GNU	Linux and Windows	Java
Weka	1993	24/04/2014	3.7.11	GNU	CrossPlatform	Java
Keel	2004	5/06/2010	2.0	GPL V3	CrossPlatform	Java
R	1997	10/04/2014	3.1.0	GNU	CrossPlatform	C, Fortran and R Programming

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

The above table gives the specialized outline of the apparatuses which incorporates name of hardware and portrayal of discharge date, most recent form, permit, working framework and language.

Table II: Analysis of features

Tool	Type	Highlights
RapidMiner	Statistical Analysis Predictive analytics.	• 20 new capacities for taking care of data and examination, including aggregation functions. • Direct file operatives • It contain macro viewer for macros and real time values.
Knime	Enterprise Reporting and Business Intelligence	• Interactive user interface, scalable and Highly extensible • Well Defined plug-in APIs • Automatic caching, data handling and visualization of data • Worksheets import and export.
Weka	Machine learning	• 49 tools for data pre-processing, 76 regression and classification algo's, 8 algo's for clustering, 15 evaluators of attributes, 10 search algo's for feature selection. • 3 association rules algorithm.
Keel	Machine learning	• Clustering, classification, regression and association discovery methods. • Data and Discovery visualization, user-friendly GUI
R	Statistical computing	• Outlier detection, Data Exploration, clustering ,Text mining, classification, SNS analysis ,regression, graphical representation of geo spatial datasets. • Handling of data and error handling.

5. CONCLUSION

Data mining is most significant and most promising interdisciplinary improvements in data innovation. The tools for data mining present interactive GUIs. The main focus of these tools is on usability and interactivity.

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

Through the use of interfaces and augmentation, extensibility is also supported. Flexibility is improved either through GUI based visual programming or by scripting languages prototyping. This paper quickly survey big data, its mining, advantages and issues of data mining, and the different open source tools and their characteristics.

Of the five data mining tools that have been inspected, Knime would be advised for all the individuals from someone new to such softwares to the ones who are exceptionally talented. The tool is extremely useful as it contains many built- in highlights and it contains some additional features that can be downloaded from the outside libraries. Based on the investigation, Weka would be recognized an extremely close to Knime due to its many built- in highlights that require no programming expertise. Conversely, Rapid Miner would be viewed as well-suited for advance clients, due to the extra programming aptitudes that are required and the restricted visualization support that is given. It may be very well summed up from the above tables that data mining is the fundamental idea to all tools; RapidMiner is the main tool which is independent of language restrictions and has statistical and predictive abilities. So, it tends to be effectively utilized and executed on any framework.

REFERENCES

[1]S.H. Begum, “Data Mining Tools and Trends – An Overview”, International Journal of Emerging Research in Management &Technology. ISSN: 2278-9359, 2013.

[2]B. Thakur and M. Mann, “Data Mining for Big Data: A Review”, Volume 4, Available: www.ijarcsse.com.

[3]K. Vikram and N. Upadhayaya, “Data Mining Tools and Techniques: a review,” Computer Engineering and Intelligent Systems, Vol 2, No.8, pp.31-39, Available: www.iiste.org, 2011.

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

[4] S. P. Deshpande and Dr. V. M. Thakare, "Data Mining System And Applications: A Review ," International Journal of Distributed and Parallel systems (IJDPS) Vol.1, No.1, September, pp.32 -44. Available: http://airccse.org/journal/ijdps/papers/0910ijdps03.pdf, 2010.

[5] H. David, M. Heikki and S. Padhraic, "Principles of data mining", Prentice hall India, pp.1.

[6] S.R.Mulik and S.G.Gulawani:"PERFORMANCE COMPARISON OF DATA MINING TOOLS IN MINING ASSOCIATION RULES", International Journal of Research in IT, Management and Engineering (IJRIME), Volume1Issue3 ISSN: 2249- 1619, 2004.

[7] R. Mikut and M.R. Wiley, "Interdisciplinary Reviews: Data Mining and Knowledge Discovery" , Volume 1, Issue 5, pages 431–443, September/October, 2011.

[8] I.H. Witten and E. Frank, "Data Mining: Practical machine Learning tools and techniques", 2nd addition,Morgan Kaufmann, San Francisco.

[9] Alcala-Fdez, J.,L., del Jesus, M.J., Ventura, s., Garrell, J.M, Otero, J., Romero,C., bacardit, j., Rivas, V.M., Fernandez, J.C., Herrera., F., : "KEEL: A software tool to Assess Evolutionary Algorithms to Data mining Problems", Soft computing 13:3,pp 307-318,2005.

[10] I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz and T. Euler, "YALE: Rapid Prototyping for Complex Data Mining tasks", in Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD-06), pp. 935-940, 2006.

[11] A.Komathi, Ramya , M. Shanmugapriya and V. Sarmila, "A Novel Comparative Study on Data Mining Tools" DOI: 10.15680/IJIRCCE.2016.0411020, 2016.

[12] "KNIME | Open for Innovation." [Online]. Available: https://www.knime.com/. [Accessed: 05-Dec-2019].

[13] "Lightning Fast Data Science Platform for Teams | RapidMiner©." [Online]. Available: https://rapidminer.com/. [Accessed: 05-Dec-2019].

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

[14] "Weka 3 - Data Mining with Open Source Machine Learning Software in Java." [Online]. Available: https://www.cs.waikato.ac.nz/~ml/weka/. [Accessed: 05-Dec-2019].