ISSN No. 0976-5697

Volume 8, No. 5, May – June 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Assessment of the prominent Open Source and Free Data Mining Tools

Afreen Samad
Department of Computer Science & Engineering
Jamia Hamdard (Hamdard University)
New Delhi, India

Syed Imtiyaz Hassan
Department of Computer Science & Engineering
Jamia Hamdard (Hamdard University)
New Delhi, India

.*Abstract:* The fundamental thought process of Data mining is to analyze the information from alternate point of view then name it and typify it keeping in mind the end goal to secure helpful data by utilizing their different new procedures and tools. Today, the different information mining apparatuses accessible that analysts requirements for assessing their information. Alongside the expanding significance of this science, there was fast increment in the quantity of free and open source apparatuses created to actualize its ideas. It wouldn't be anything but difficult to choose which device plays out the coveted undertaking better, in addition to we can't depend exclusively on depiction given by the seller .Distinctive devices incorporated into data mining like, WEKA, Tanagra, Rapid Miner, Orange, KNIME etc are reviewed in this paper and their advantages and disadvantages are presented and furthermore think about their features. For the scientists this comparative review would be helpful to make a best determination of the device

Keywords: Data Mining Tools, WEKA tool, R Tool, Rapid Miner, KMINE, Orange, Tanagra, Scikit-learn

I. Introduction

Presently a day's different instruments for information mining are accessible either as open-source or business programming. It incorporates extensive variety of programming items, from agreeable issue autonomous information mining suites, to business focused information distribution centers with coordinated information mining capacities and to early research prototFree and freely accessible programming instruments for DM have been being developed for as long as 20 years. The objective of these apparatuses is to encourage the somewhat muddled information investigation handle and to offer every single intrigued specialist a free other option to business information examination stages. They do as such essentially by proposing incorporated conditions or concentrated bundles on top of standard programming dialects, which are regularly open source types for recently created strategies. Specialists needs these sort of hardware for investigation their information. Bringing about decision to be made by the client, which apparatus to choose from all the accessible open source and free instruments to fit his needs? Plainly, every client searches for various components to be accessible in his favored device. From the client perspective, his best device will picture information and apply sought DM assignment on client accessible information while running on client current stage effectively. Moreover, it might be famous which impact on the accessibility of support and answers for the issues that may show up while utilizing the apparatus. Additionally, if the client is propelled he may need to stretch out and add usefulness to the apparatus.

This paper concentrates on the survey of a few such devices that have developed more productive and valuable throughout the years, some even tantamount or preferable in specific angles over their business partners. Specifically, the primary attributes of Rapid Miner [1], R [2], WEKA [3], Orange [4], KNIME [5], and scikit-learn [6] and Tanagra will be illustrated and analyzed.

II. Classification of Data Mining Tools

There are mostly three unique classifications of information mining apparatuses. Conventional information mining apparatuses, Application based devices/Commercial based programming and electronic information mining instruments. Portrayal of each is as per the following:

1) ***Traditional information mining devices***: Some mining projects are work as conventional approach to gather and break down information which utilized by different organization for basic leadership procedure of huge informational indexes. Dominant part of these is upheld by windows and UNIX forms. In any case, some product had practical experience in a solitary working framework and some of the time taking care of with just a single database sort.

2) ***Application based devices***: An applications which demonstrates the business situated interface for information execution. In this chronicled information are spoken to as a reference and check the present patterns with a specific end goal to see the adjustments in the business. In this way, application based instruments are anything but difficult to utilize and helps in managerial work and give administrations to organization execution.

3) ***Web based information mining devices***: This sort of devices is called content mining instrument in view of its capacity to mine different sort of content from any composed assets. And furthermore help for checking and changing over information in chose organize which is perfect with any tools.[7]

III. DATA MINING TOOLS

Some of apparatuses which are accessible in market are portrays as takes after like: WEKA, R, Orange, Rapid Miner and Tanagra, KMINE and so forth..

A. WEKA Tool

WEKA is an information mining framework created by the College of Waikato in New Zealand that actualizes information mining calculations. WEKA gives 3 approaches to utilize the programming: the GUI, a Java API, and an command line interface (CLI) .WEKA Explorer preprocessing, order, bunching, affiliation, trait determination, representation devices and the meaning of information source, information readiness, machine learning calculations, and representation. The Experimenter is utilized for the most part for examination of the execution of distinctive calculations on the same dataset. Learning stream is similar to Rapid Miner's administrator worldview it could be said that it enables one to determine the dataflow utilizing suitably associated visual segments. In spite of the fact that not as outwardly engaging and extendable as Rapid Miner, WEKA's learning stream still does the employment. WEKA underpins many model assessment methods and measurements, however needs numerous information study and representation techniques. It is likewise more arranged towards grouping furthermore, relapse issues and less towards spellbinding insights and bunching strategies, albeit a few changes were made as of late regarding bunching. The support for enormous information, content mining, and semi-administered learning is presently constrained, while profound learning strategies are as yet not considered. WEKA is a gathering of machine learning calculations for information mining assignments and appropriate for growing new machine learning plans. WEKA is a java based programming ability of working under different working frameworks. With the Java-based form, the device is extremely modern and utilized as a part of a wide range of uses counting perception and calculations for information investigation and prescient displaying. Its free under the GNU General Public Permit, which is a major in addition to contrasted with Rapid Miner, since clients can tweak it anyway they please.

WEKA bolsters many model assessment methods and measurements, however needs numerous information overview and representation techniques. It is additionally more arranged towards order also, relapse issues and less towards spellbinding measurements and bunching strategies, albeit a few enhancements were made as of late as for grouping. The support for huge information, content mining, and semi-directed learning is at present restricted, while profound learning techniques are as yet not considered.

Highlights

- This open source data mining tool is java based.
- For beginners it is suitable tool as it is easy to use and has capability to run various algorithms
- It is stage autonomous.
- Various data mining including: Data preprocessing, Classification rules, relapse, Clustering, affiliation rules, perception, highlight choice and enhancing the learning disclosure.
- WEKA has 49 Data preprocessing devices, 76 Classification/relapse calculation, 8 Clustering calculation, 3 calculation for discovering affiliation rules, 15 property/subset evaluator in addition to 10 look calculation for highlight choice [8].
- There are different worked in elements.
- There is no programming and coding dialect required.

Favorable circumstances

- Easy to control the information.
- Provide access to SQL databases.
- It gives two choices to the client to communicate through Explorer and Command line [9].
- Specially utilized for information mining.
- It gives different machine learning calculations to information mining undertakings.
- It bolsters different standard Data mining assignments that include: Data preprocessing, Clustering and Classification, Regression, Visualization and Feature determination [10].

Hindrances

- Memory is constrained and has lesser execution [11].
- Data perception and information overview is constrained.
- Worse appropriate choice for the vast datasets as they generally dealt with.
- Lacking in the portrayal to the consequence of handling.
- Limited capacity to segment dataset to preparing and test set [12].
- It doesn't acknowledge information in each organization (information arrange requirements).
- Not great in interfacing with other programming [11].

B. R Tool

The open-source instrument and programming dialect of decision for analysts, R, is likewise a solid choice for DM errands. R has been being developed throughout the previous 15 years and is the successor of S, a factual dialect initially created by Bell Labs in 1970s. The source code of R is written in C++, Fortran, and in R itself. It is a translated dialect and is generally advanced for grid based counts, tantamount in execution to economically accessible MATLAB and openly accessible GNU Octave. The fundamental dialect is stretched out by a bunch gathering of bundles for a wide range of computational assignments. The apparatus offers just a basic GUI with order line shell for input. It is surely not an easy to use condition since all summons should be entered in the R dialect. The expectation to learn and adapt is steep, and albeit straightforward undertakings, for example, drawing charts and enlightening measurements can be adapted rather simple, the dialect's maximum capacity is hard to ace. Some propelled clients now and then offer accommodating reference cards with the rundown of the most noteworthy capacities [13]. From DM client's point of view, R offers exceptionally quick executions of many machine learning calculations, practically identical in number to Rapid Miner and WEKA (from which a substantial number of calculations is obtained), and furthermore the full prospect of measurable information representations strategies. It has particular information sorts for taking care of enormous information, underpins parallelization, information streams, web mining, chart mining, spatial mining, and numerous other propelled errands, including a constrained support for profound learning strategies.

R's primary issue is its dialect, which, in spite of the fact that profoundly extendable, is likewise a troublesome one to learn altogether enough to end up noticeably gainful in DM. Headway for DM undertakings toward that path is the Rattle bundle (creator Dr. Graham Williams and other patrons) that offers a not too bad GUI for R. Shake, which is being developed from 2006, is like WEKA's Pilgrim in the feeling of ease of use. It loads discrete bundles from R upon demand for a particular investigation. Shake utilizes a portion of the R's standard usage of DM techniques and furthermore extra bundles. The main issue with Rattle is that it can't utilize the majority of the R's calculations or WEKAs DM usage. In any case, Shake is easy to understand and very well known in DM group.

C. *Rapid Miner*

Rapid Miner (beforehand: Rapid-I, YALE) is a developed, Java-based, general DM device right now being developed by the organization Rapid Miner, Germany. Past variants (v. 5 or lower) were open source. The most recent one (v. 6) is exclusive for the time being, with a few permit choices (Starter, Individual, Professional, and Enterprise). The Starter form is free with constraints just in regard to most extreme designated memory estimate (1 GB) and information documents (.csv, Excel). The device has turned out to be extremely well known in a few late years and has a expansive group bolster. Rapid Miner offers a coordinating situation with outwardly engaging and easy to understand GUI. Everything in Rapid Miner is centered around procedures that may contain sub processes. Forms contain administrators as visual segments. Administrators are executions of DM calculations, information sources, and information sinks. The dataflow is developed by intuitive of administrators and by associating the sources of info and yields of relating administrators. Rapid Miner likewise offers the alternative of application wizards that develop the procedure consequently in light of the required venture objectives (e.g. coordinate showcasing, agitate examination, notion investigation). There are instructional exercises accessible for some particular errands so the instrument has a steady expectation to learn and adapt. In spite of the fact that Rapid Miner is very intense with its essential set of administrators, the expansions make it considerably more helpful. Well known augmentations incorporate arrangements of administrators for content mining, web mining, time arrangement investigation, and so forth. The vast majority of the administrators from WEKA are likewise accessible through augmentation, which builds the quantity of executed DM techniques. The device has not very many inadequacies. The most imperative one is the move to a novel model of business. It stays to be seen whether the move to restrictive permit will constrain the number of its clients, yet it may not be useful. The support for profound learning techniques and a portion of the more propelled particular machine learning calculations (e.g. to a great degree randomized trees, different inductive rationale programming calculations) is at present constrained. Be that as it may, enormous information investigation by means of Hadoop bunch (Radoop) is bolstered.

Highlights

- It is stage free.
- It has similarity with different databases like prophet, MySQL, Excel, SPSS, Microsoft SQL server and so forth.
- It gives Drag and Drop interface to outline the investigation procedure.
- It underpins and acknowledges new information drivers.
- It gives more than 500 administrators to all machine learning methods, and furthermore joins learning plans and properties evaluators of the WEKA learning condition [14].
- It enable client to work with various sizes and sorts of information sources.

Preferences

- It has gigantic adaptability.
- It gives the mix of greatest calculation of such instruments.
- Easy to troubleshoot the mistakes.

Disservices

- Limited dividing capacities for dataset to preparing and testing sets.

D. *KMINE*

KNIME (Konstanz Information Miner) is a general purpose DM apparatus in light of the Eclipse stage, created and kept up by the Swiss organization KNIME.com AG. Its advancement begun in 2004 at the College of Konstanz, Germany, and the underlying rendition was discharged in 2006. KNIME is open-source, however business licenses exist for organizations requiring proficient specialized support. As indicated by the authority site, KNIME is utilized by more than 3000 associations in more than 60 nations, and there is by all accounts a extensive group bolster. The device sticks to the visual programming worldview show in most DM apparatuses, where building squares are set on a canvas and associated with get a visual program. In KNIME, these building squares are called hubs, and as indicated by the official site, more than 1000 hubs are accessible through the center establishment and different augmentations. Hubs are composed in a chain of command what's more, can be looked by name inside a natural interface. Every hub is recorded in detail, and the documentation is naturally appeared inside the interface once the hub is chosen. An extensive vault of illustration work processes is accessible to encourage speedier learning of the device. One of the most prominent qualities of KNIME is the joining with WEKA what's more, R. Despite the fact that augmentations must be introduced to empower the combination, the establishment itself is paltry. WEKA mix empowers utilizing all the usefulness accessible in Weka as KNIME hubs, while R mix empowers running R code as a stage in the work process, opening R perspectives and learning models inside R. A few other fascinating free expansions are additionally accessible, e.g. JFreeChart augmentation that empowers progressed outlining, OpenStreetMap augmentation that empowers working with geological information, and so forth. There are likewise business augmentations for more particular functionalities. By and large, KNIME is by all accounts one of the best decisions for a client keen on an absolutely visual programming worldview with a requirement for an expansive assortment of hubs.

KNIME is capable instrument for investigative undertaking, extricating information and learning from the web communities. The KNIME base form as of now consolidates hundreds of handling hubs for information I/O, preprocessing and purifying, demonstrating, examination and information mining and different intelligent perspectives, for example, disperse plots, parallel directions and others[15].In KNIME, portrayal of information sources and sinks, mining algorithm, transformations, visualizations ,etc characterized by set of hubs called "work process" and every hub has its particular information and yield ports that relies on upon the usefulness of the hub [16]. For both basic and complex information sorts, KNIME enables progressive examination to find slants and foresee future outcomes. KNIME utilizes for educating and additionally look into which permits to coordinate the new calculation and instruments in a more straightforward way.

Highlights

- Available to everybody i.e., enable clients to utilize the all around characterized hub API to include restrictive expansions.
- Intuitive UI.
- It is anything but difficult to utilize and handle distinctive capacities.
- KNIME modules cover a wide assortment of functionalities like, I/O, information control, sees, hilting and so forth to better comprehend your information.
- It gives the clients to make information streams or pipeline outwardly, clients can specifically execute a few or all investigation steps, concentrate the outcomes, models, and community oriented understandings [10].
- For cross approval and autonomous approval, it gives usefulness to spare parameters.

Favorable circumstances

- The significant advantage of this is anything but difficult to utilize module [16].
- It in light of the hub work which incorporates more than 100 hubs to analyze the information [17].
- It gives information stream handle by moving new hubs.

Disservices

- Less appropriate alternative for vast complex work processes.
- Partitioning capacity is restricted for dataset [12]

E. ***Orange***

Orange is a Python-based apparatus for DM being created at the Bioinformatics Laboratory of the Faculty of Computer and Information Science at the University of Ljubljana. It can be utilized either through Python scripting as a Python module, or through visual programming. Its visual programming interface, Orange Canvas, offers a organized perspective of upheld functionalities gathered into nine classes: information operations, representation, order, relapse, assessment, unsupervised learning, affiliation, representation utilizing Qt, and model executions. Functionalities are outwardly spoken to by various gadgets (e.g. perused document, discretize, prepare SVM classifier and so forth). A short depiction of every gadget is accessible inside the interface. Writing computer programs is performed by setting gadgets on the canvas and interfacing their sources of info and yields. The interface is exceptionally cleaned and outwardly engaging, offering a lovely client encounter. One obvious drawback of Orange is that the number of accessible gadgets appears to be constrained when contrasted with different devices, for example, Rapid Miner or KNIME, particularly due to the absence of combination with WEKA. Still, the scope of standard information mining systems is very great. Besides, there are a number of intriguing gadgets right now being developed that can be found in the "Model" class, so it is sensible to expect that the list of capabilities will be extended later on.

F. ***Scikit-learn***

Scikit-learn is a free bundle in Python that expands the usefulness of NumPy and SciPy bundles with various DM calculations. It additionally utilizes the matplotlib bundle for plotting diagrams. The bundle continues making strides by tolerating profitable commitments from numerous patrons and is upheld by both INRIA and Google Summer of Code. One of its fundamental solid focuses is a well written online documentation for the greater part of the actualized calculations. Elegantly composed documentation is a necessity for any donor and is esteemed more than a great deal of inadequately archived calculation usage. The bundle bolsters the vast majority of the center DM calculations. Be that as it may, a few noteworthy DM calculation bunches have been overlooked as of now, including arrangement standards and affiliation rules. On the other hand, the bundle is solid in capacity based strategies counting many general direct models and different sorts of SVM executions. It is likewise very quick in spite of being written in a deciphered dialect. This is essentially on the grounds that the benefactors are made a request to upgrade the code in different perspectives, for example, calling cluster based NumPy number crunching calculation or composing wrappers for existing C/C++ executions in Cython. In spite of its points of interest, the utilization of Scikit-learn still requires that one is a talented software engineer in Python in view of its charge line interface. This will degrade practically anybody not versed in this dialect in light of the fact that there are different devices that don't have this suspicion.

G. ***Tanagra***

Tanagra is open source information investigation programming for scholarly furthermore, inquire about purposes which proposes a few information mining techniques from exploratory information examination, factual learning, machine learning and database range [18]. The principle reason of Tanagra is to give stage to specialists and understudies to utilize information mining programming in simple route by fitting in with the present standards of the product improvement and permitting to examine either genuine or engineered information. The second design is to propose an engineering enabling the clients to add to include their own information mining strategies it looks at their exhibitions. It acts more as an exploratory stage so as to do the fundamental work, apportioning them to manage the repulsive piece of the information administration. Last reason for existing is to give the bearing to fledgling designers in diffusing a conceivable strategy for building this sort of programming. It can be considered as educational device for getting the hang of programming strategies since it grants to get to the source code, to look example of the programming how it is fabricated, the issues to maintain a strategic distance from, key strides of the venture, apparatuses utilized and code libraries utilized for the venture.

IV. CONCLUSION

A few DM devices were exhibited in this work. By and large conclusion is that there is no single best device. Each instrument has its solid focuses and shortcomings. By and by, Rapid Miner, R, WEKA, and KNIME have the majority of the fancied attributes for a completely useful DM stage what's more, in this way their utilization can be prescribed for the greater part of the DM tasks. This similar review will make things simpler to the learner in the determination of information mining apparatuses as indicated by their territories. In future, we will discover the answer for overcome from the restrictions of such apparatuses to make them best in each angle.

V. REFERENCES

[1] M. Hofmann and R. Klinkenberg, RapidMiner: Data Mining Use Cases and Business Analytics Applications, Boca Raton: CRCPress, 2013.

[2] Y. Zhao, R and Data Mining: Examples and Case Studies, San Diego: Academic Press, 2012.

[3] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The WEKA data mining software: an update,” SIGKDD Explorations, vol. 11, no. 1, pp. 10–18, 2009.

[4] J. Demšar, T. Curk, and A. Erjavec, “Orange: Dat Mining Toolbox in Python,” Journal of Machine Learning Research, vol. 14, pp. 2349−2353, 2013.

[5] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, et al., “KNIME: The Konstanz Information Miner”, in Data Analysis, Machine Learning and Applications (Studies in Classification, Data Analysis, and Knowledge Organization), Springer Berlin Heidelberg, pp. 319–326, 2008.

[6] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al., “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830,2011.

[7] Mrs. bharti M. Ramgeri, “ Data Mining Techniques and application”, Indian Journal of Computer Science and Engineering,Vol. 1 No. 4,pp 301-305.

[8] S. Srivastava, WEKA: A Tool for Data Preprocessing, Classification, Ensemble, Clustering and Association Rule mining, International Journal of Computer Applications, 88(10), February 2014.

[9] S.K. David, Amr T.M. Saeb, K.A. Rubeaan, Comparative Analysis of Data Mining Tools and Classification Techniques using WEKA in Medical Bioinformatics, Computer Engineering and Intelligent System, 4(13), 2013.

[10] K. Saravanapriya, A Study on Free Open Source Data Mining Tools, International Journal of Engineering and Computer Science, 3(12), December 2014.

[11] S. Singhal, M. Jena, A study on WEKA tool for Data Preprocessing, Classification and Clustering, International Journal of Innovative Technology and Exploring Engineering (IJITEE), 2(6), May 2013.

[12] S. Sarumathi, N. Shanthi, S. Vidhya, M. Sharmila , A Review: Comparative Study of Diverse Collection of Data Mining Tools, International Journal of Computer, Electrical, Automation, Control and Information Engineering, 8(6), 2014.

[13] Y. Zhao, R Reference Card for Data Mining, Available at: [last accessed 2014-02-23] http://www.rdatamining.com/docs/Rrefcard data-mining.pdf

[14] H. Solanki, Comparative Study of Data Mining Tools and Analysis with Unified Data Mining Theory, International Journal of Computer Applications, 75(16), August 2013.

[15] M. Vijayakamal, M. Narendhar, A Novel Approach for WEKA & Study On Data Mining Tools, International Journal of Engineering and Innovative Technology (IJEIT), 2(2), August 2012.

[16] L. Kataria, Implementation of Knime-Data Mining Tool, International Journal of Advanced Research in Computer Science and Software Engineering, 3(11), November 2013.

[17] S. Gunnemann, H. Kremer, R. Musiol, R.Haag, T. Seidl, A Subspace Clustering Extension For the KNIME Data Mining Framework, 2012 IEEE 12th International Conference on Data Mining Workshops.

[18] Y. Ramamohan, K. Vasantharao, c. Kalyana chakravarti, and A.S. K.Ratnam, “ A Study of Data Mining Tools in Knowledge Discovery Process”, International Journal of Soft computing and Engineering (IJSCE), Vol.2, Issue -3,July 2012,pp 191- 194.