Gurukul International Multidisciplinary Research Journal (GIMRJ)*with*
International Impact Factor 8.357
Peer Reviewed Journal
https://doi.org/10.69758/GIMRJ/2504I5VXIIIP0055

e-ISSN No. 2394-8426
Monthly Issue
APR-2025
Issue–IV, Volume–XIII

A Study: WEKA and different Data mining tools

Aniruddha S Holey[1] **Dr.Swati S Sherekar**[2]

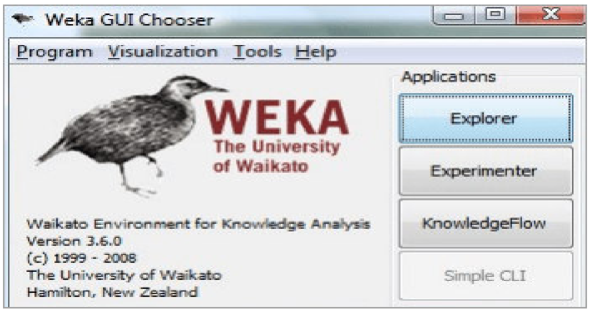
[1]P.G.Department of CSTH.V.P.M.,Amravati, Maharashtra, India *Email:*[*aniruddha.holey@gmail.com*](mailto:aniruddha.holey@gmail.com)
2Department of CSE,SGB Amravati University, Amravati, Maharashtra, India
Email:ss_sherekar@rediffmail.com

Abstract: Whenever datasets grow in size, it is now necessary to analyze those using data mining and machine learning approaches. Data mining as well as machine learning researchers struggle with how to use machine learning algorithms, how much additional efforts required for data processing, and which tools work best in certain situations. This article includes an introduction to several tools for data mining and an analysis of each.

Keywords: Data Mining, Machine Learning, WEKA, Data Mining tools.

I -Introduction

To transform data into business intelligence, the data mining process uses association rules, classification, decision trees, clustering techniques, the K-Mean algorithm, and regression. Orange, Rapid Miner, WEKA, JHep Work, and KNIME are the top open source data mining software but WEKA is good as compare to others tools. [1,2].A Java-based open-source data mining tool called Waikato Environment for Knowledge Analysis was created by the University of Waikato, New Zealand. It is free software distributed under the terms of the GNU General Public License. It is compatible with adverse range of operating systems, namely Windows, Linux, and Mac.Both a GUI and a Command Line Interface are offered by WEKA for running these algorithms on your dataset. WEKA is Java primarily based package for playacting completely different information mining task like Information preprocessing, Classification, Clustering, Association Rule Mining and Visualization. Four different forms of graphical user interfaces, including explorer, experimenter, knowledge flow, and workbench, are supported by WEKA [2]. It is the only toolkit that has attracted such wide acceptance and endured for such a long time, WEKA is a landmark system in the history of the data mining and machine learning researchers. The dataset is split into two groups by the WEKA tool: the training dataset and the test dataset. The training dataset is then used to build a predictive model and the test dataset is used to evaluate the performance of the model [5].In this paper first part contains the introduction about WEKA ,second part include Literature review, third section contains different data mining tools, fourth section contains the analysis of different data mining tools and fourth section conclusion.



Waikato Environment for Knowledge Analysis
Version 3.6.0
(c) 1999 - 2008
The University of Waikato
Hamilton, New Zealand

Fig.1WEKAGUI Launcher

II-Literature review

Numerous authors have researched WEKA tools in the area of data mining and compared them to other tools.

FatmaAIbrahim[2]:was discussed about the steps for knowledge discovery from databases, data mining and WEKA and also covered some common classification algorithms used in WEKA tool for data classification such as Decision Tree, K-nearest neighbor, Naive Bayes, Support Vector Machine. PankajSingh[3]:state that when employing a large data set, the K-Means algorithm produces high-quality clusters. Simple K-Mean algorithms outperform Hierarchical Clustering in terms of performance. High variance in density makes the density-based clustering algorithm unsuitable for the data, while noisy data makes the hierarchical clustering approach more sensitive. K mean algorithm is a faster clustering algorithm than others. Stephen Garner [4]:includes a variety of machine learning techniques from the supervised and unsupervised learning disciplines. Other industries that make use of these technologies include education, machine learning research, and agriculture. It has lessened the complexity associated with feeding real-world data into various machine learning systems and assessing their results. K.P.S.Attwal [5]: This paper also demonstrates how to utilize WEKA filters to split the dataset into training and test categories, and it also illustrates how the model can be used to predict the class of those tuples and objects whose class label is unknown. In the form of Java API, WEKA saves pre-written packages, classes, and interfaces together with their corresponding methods, fields, and constructors. The default file format for WEKA is ARFF (Attribute-Relation File Format), however it also supports files in CSV, C4.5, and other formats. It is simple to convert data from an Excel worksheet to the csv/ARFFformat. J.E.Gewehr et al [6]: The use of several data formats that are pertinent to bio informatics is made simple by BioWeka, as we learn in this paper while using WEKA. Using the normal process in WEKA, anything from classification to validation can be done with such data without adding extra work. Additionally, WEKA now includes certain bioinformatics-specific methods due to BioWeka. Mark Hell [7] describes the learning techniques that WEKA's latest version has added, together with the LogitBoost classifier. With the addition of a central log file, WEKA 3.6 has also improved logging. This file includes a number of plug-in techniques and records any data written to any graphical logging panel in WEKA, as well as any output to standard out and standard error. Two new supporting GUIs, the SQL viewer and Bayes network editor, are available via the WEKA Tools menu.

III-Different Data Mining Tools:

Software used for data mining includes Orange, Rapid Miner, Weka, JHepWork, and KNIME etc.

Orange: The Orange is a component-based data mining and machine learning software suite that includes Python bindings and libraries for scripting, as well as user- friendly yet potent, quick and adaptable visual programming front exploratory data analysis and visualization. With a complete set of components for data preparation, feature filtering, modeling, model evaluation, and exploration approaches. It is described by the University of Ljubljana in Switzerland and was created using Python and C++. [1, 8]

Rapid Miner: It is one of the top predictive analysis systems created by the Rapid Miner firm. Java is the programming language used to create it. It offers an integrated platform for predictive analysis, machine learning, deep learning, text mining, multimedia mining, feature engineering, data stream mining and tracking drifting notions, ensemble technique development and distributed data mining. It offers the server in both on-premises and private/public cloud configurations. Its foundation is a client/server model. It is formally known as YALE (Yet another Learning Environment). [1, 8, 9]

JHep Work: It is a free and open-source, Java based data analysis framework; this tool is primarily intended for scientists, engineers, and students. It was developed in an effort to create a data analysis environment using open source packages with a user-friendly interface and to produce a tool that is competitive with proprietary software. It is designed specifically for interactive scientific charts in 2D and 3D, and it includes Java-based libraries for mathematical functions, random numbers, and other data mining algorithms. Although jHepWork is based on the high-level programming language Jython, Java code can also be used to access its graphic and numerical libraries. [1]

Konstanz Information Miner (KNIME): A platform for data integration, processing, analysis, and exploration is called KNIME (Konstanz Information Miner), and it is free and open source. Users are given the ability to graphically build data flows or pipelines, execute only some or all of the analytic processes, and then study the models, interactive views, and findings. It is built on Eclipse and uses its extension technique to support plug-ins, giving it additional capability. It is written in Java programming language, pioneered by University of Konstanz, Germany. [1, 11]

DBMiner: Introduced by Intelligent Database System Research Laboratory, Simon Fraser University, Canada.[10] It is a data mining system for interactive mining of multiple-level knowledge in large relational databases. The system implements a wide spectrum of data mining functions, including generalization, characterization, discrimination, association, classification, and prediction. By incorporation of several interesting data mining techniques, including attribute-oriented induction, progressive deepening for mining multiple- level rules and meta-rule guided knowledge mining, the system provides a user-friendly, interactive data mining environment with good performance. DBminer performs interactive data mining at multiple concept levels on any user-specified set of data in a database using an SQL-like Data Mining Query Language, DMQL, or a graphical user interface. [1, 8]

TANAGRA: It is free data mining software for academic and research purposes. It proposes several data mining methods from exploratory data analysis, statistical learning, machine learning and databases area. TANAGRA is more powerful, it contains some supervised learning but also other paradigms such as clustering, factorial analysis, parametric and nonparametric statistics, association rule, feature selection and construction algorithms. TANAGRA is an "open source project" as every researcher can access to the source code, and add his own algorithms, as far as he agrees and conforms to the software distribution license. TANAGRA is to propose to researchers an architecture allowing them to easily add their own data mining methods, to compare their performances. Tanagra can be considered as a pedagogical tool for learning programming techniques.[1,8,9]

IV-Analysis:

In this paper study the different Data Mining tools, which is used for the generating the knowledge. The comparison is based on the Programming Language ,License, availability such as Open Source or proprietary software, area, portability and tools wise supporting data mining task.

Table-I: Comparison of different Data Mining Tools

Parameter	**WEKA[2,4,7]**	**Orange[1,8]**	**RapidMiner(YALE)[1,8,9]**	**TANAGRA[1,8,9]**	**DBMINER[1,8,10]**	**JHepWork[1]**	**KNIME [1,11]**
CompanyName	University of Waikato, New Zeland	University of Ljubljana, Switzerland	Rapid Miner Germany	LumiereUniversityLyon2,France	Intelligent Database System Research Laboratory, Simon Fraser University, Canada	-	University of Konstanz, Germany
ProgrammingLanguage	Java	C++,Python	Java	C++	Data MiningQuery Language(DMQL)	Jythone, Java-based libraries	Java

License	GNU	GNUGPLV3	AGPL	GNUGPL3	NSERC	GNU	GNU
Availability	OpenSource	Open Source	Open Source	Open Source	Proprietary software	Open Source (Not completely free for Commercial usage)	Open Source
Areas	Education, Agriculture ,Business, Medical	Datapreparation,Featurefiltering,Modeling,Modelevaluation,andExplorationapproaches	IT Industry, Academic and Research Purpose Artificial Intelligence	Education, Research	Relational databases and Data warehouses	Scientists, Engineers, and Students	Manufacturing , Finance, Health, Retail, Government.
Portability	Cross Platform	Cross Platform	Cross Platform	Cross Platform	Windows/NT system	Cross Platform	Cross Platform
Supporting Data Mining task	Processing, Classification, Clustering, Regression, Visualization and Feature Selection.	Preprocessing,Discutization,PredictiveModeling,DataDescriptionMethod	DataLoading, Transformation,Preprocessing,Visualization,Modeling, Evaluation,Deployment	Clustering,Parametric&Non-ParametricStatisticassociationrules,FeatureSelection,Construction algorithm.	Characterization, Discrimination,Association,Classification,Prediction,Time-Series Analysis,Clustering,	Scientific computation, data analysis and data visualization	Integration, Processing,Analysis,Exploration.

V-Conclusion:

In this article, we talk about eight data mining tools, some of which are free source and others of which are proprietary. We have compared the tools based on a number of factors, including the programming language, the tools' accessibility, the areas in which they are most frequently used, the software's portability, and, finally, which data mining tasks they support. From these factors, we can determine how well the tools function and whether they are helpful to researchers. Most of the tools used in this study are really functional and offer excellent resources for students and researchers; we can also choose the best tool for data mining from the provided analysis. When it comes to the work that will be done in the future, various data sets will be used with best tools to test the accuracy and effectiveness of the dataset.

REFERENCES

[1] SaurabhA.Ghogare,"Introduction to WEKA & Study on DataMining tool with its Comparative Analysis,"61stIETE Annual Convention 2018 on "Smart Engneering for Sustainable Development" Special Issue of IJECSCSE,ISSN:2277-9477, pp.187-191,2018.

[2] Fatma A. Ibrahim, Omar A. Shiba, "Data Mining: WEKA Software (an Overview)," Journal of Pure and Applied Science,ISSN:2521-9200, Vol.18 No. 3pp. 54-58,2019.

[3]Pankaj Singh, Sudhakar Singh, RakhiGarg, Devisha Singh, "Comparative Study of Data Mining Algorithms through Weka," International Journal of Emerging Research in Management &Technology ISSN: 2278-9359 (Volume-4, Issue-9) 2015.

[4] StephenGarner,"WEKA:TheWaikatoEnvironmentforKnowledge Analysis,"ResearchGATE,MAY 1995.

[5] KanwalPreet Singh Attwal , Amardeep Singh Dhiman, "EXPLORING DATA MINING TOOL - WEKA AND USING WEKA TO BUILD AND EVALUATE PREDICTIVE MODELS," Advances and Applications in Mathematical Sciences Mili Publications Volume 19, Issue 6, Pages 451-469April 2020.

[6] Jan E. Gewehr, Martin Szugat, Ralf Zimmer, "BioWeka—extending the Weka framework for bioinformatics," Oxford University Press Vol. 23 no. 5 2007, pages 651–6532007.

[7] Mark Hall Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer Peter Reutemann, Ian H. Witten, "The WEKA Data Mining Software: An Update," SIGKDD ExplorationsVol. 11 Issue 1, pages 10–18.

[8] Y. Ramamohan, K. Vasantharao, C. Kalyana Chakravarti, A.S.K.Ratnam, "A Study of Data Mining Tools in Knowledge DiscoveryProcess" InternationalJournalofSoft Computingand Engineering(IJSCE)ISSN:2231-2307,Volume-2, Issue-3,July 2012

[9] Nurdatillah Hasim,Norhaidah Abu Haris," A Study of Open-Source Data Mining Tools for Forecasting", IMCOM '15, January 08 - 10 2015, BALI, Indonesia

[10] JiaweiHan,SonnyHanSengChes,NebojsaStefanovic,OsmarR.Zaiane,"DBMiner:asystemfordataminingin Relational Database and Data warehouse" https://www.researchgate.net, September 1999.

[11] Michael R. Berthold, Nicolas Cebron, Fabian Dill, Thomas R. Gabriel, Tobias K¨ otter,Thorsten Meinl, Peter Ohl, Kilian Thiel and Bernd Wiswedel, "KNIME – The Konstanz Information Miner Version 2.0 and Beyond", SIGKDD Explorations, Volume 11, Issue 1, pages 26-31.