An Overview of General Data Mining Tools

Bala Brahmeswara Kadaru[1] , Munipalli UmaMaheswararao[2]

[1,2] Computer Science and Engineering, GEC, Gudlavalleru, AP, India.

---***---

Abstract— Data Mining has emerged to meet th e requirement of quick and accurate information support for decision making process. The idea is to extract the data from the database for the operational use. Data Mining is analysis of data to identify relationship between different data elements or entities. The process of data mining can also involve correlation or association between two or more data elements, entities or events. Data Mining tools which are helpful and marked as the important field of data mining Technologies. describes the characteristics of most used software tools for general data mining that are available today: Weka, Rapid Miner, IBM SPSS, Tanagra, KNIME, Orange, R. To analyse, manage and make a decision of such type of huge amount of data we need techniques called the data mining which will transforming in many fields using these tools. Data mining tools emphasizes the quality of Weka,RapidMiner, IBM SPSS, Tanagra, KNIME, Orange and R platforms, but also acknowledges the significant advancements made in the other tools.

Keywords— Weka, Rapid Miner, IBM SPSS, Tanagra,KNIME,Orange, R

1.INTRODUCTION

Data mining refers to extracting or mining" knowledge from large amounts of data. The term is actually a misnomer. Remember that the mining of gold from rocks or sand is referred to as gold mining rather than rock or sand mining. Thus, data mining" should have been more appropriately named knowledge mining from data", which is unfortunately somewhat long. Knowledge mining", a shorter term, may not reflect the emphasis on mining from large amounts of data. Nevertheless, the mining is a lively term which characterizes the process that finds a small set of precious nuggets from a great deal of raw material. . There are many other terms carrying a similar or slightly different meaning to data mining, such as knowledge mining from databases, knowledge extraction, data/pattern analysis, data archaeology, and data dredging[1].

Data mining as a synonym for another popularly used term, Knowledge Discovery in Databases", or KDD. Data mining as simply an essential step in the process of knowledge discovery in databases Knowledge discovery as a process consists of an iterative sequence of the following steps:

Data cleaning (to remove the noise or irrelevant data)[2].

Data integration (where the multiple data sources may be combined)[3].

Data selection (where data relevant to specific analysis task are to be retrieved from the database).

Data transformation (where the data is to be transformed or consolidated into different forms .

1.1 Importance of Data Mining

Recently the Data mining is fascinating the attention in the information industry in the recent years is due to the wide availability of the huge amounts of data and the imminent need for using such type of data into a useful information and knowledge. The information and the knowledge is to be gained and also can be used for the applications ranging from business management, production control, and market analysis, to engineering design and science exploration.

Data mining is to be viewed as a result of the natural evolution of the information technology. An evolutionary path has been witnessed in the database. In the Data collection and the database creation, the data management (including data storage and retrieval, and database transaction processing), and the data analysis and the understanding (involving data warehousing and data mining). For instance, the early development of data collection and database creation mechanisms served as a prerequisite for later development of effective mechanisms for data storage and retrieval, and query and transaction processing.

1.2 Data Mining Techniques:

1. Association.
2. Classification. ...
3. Clustering. ...
4. Prediction. ...
5. Sequential patterns. ...
6. Decision trees. ...
7. Combinations. ...
8. Long-term (memory) processing.

II. LITERATURE SURVEY

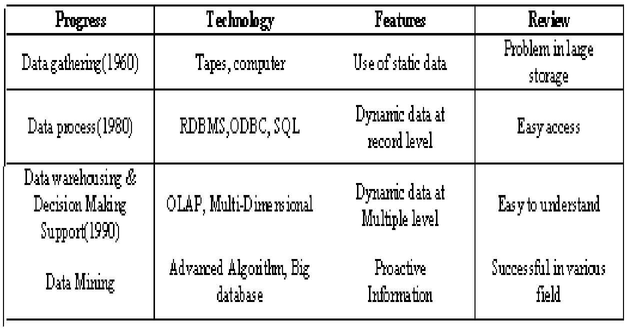
Presently the Data mining is not about the tools or the database software that which we are using. We can perform the data mining with the comparatively modest database systems and the simple tools, including the creating and writing your own, or using off the shelf software packages. There are also some Complex data mining benefits from the past experience and the algorithms defined with the existing software and the packages, with certain tools gaining a greater affinity or reputation with different techniques.
Today we are having a entirely new range of the tools and the systems available, including the combined data storage and the processing systems. And also we can mine the data with the various different data sets, including, traditional SQL databases, raw text data, key/value stores, and the document databases. We are using the Clustered databases, such as the Hadoop , Cassandra, CouchDB, and Couchbase Server, store and providing the access to the data in such a way that it does not match the traditional table structure.

For instance, audit interrogation tools can be used to highlight fraud, data anomalies, and patterns. An example of this has been published by the United Kingdom's Treasury office in the *2002–2003 Fraud Report: Anti-fraud Advice and Guidance*, which discusses how to discover fraud using an audit interrogation tool. We also have the Additional examples of using the audit interrogation tools to identify the fraud are found in David G. Coderre's 1999 book, *Fraud Detection*.

Internal auditors can use spreadsheets to undertake simple data mining exercises or to produce summary tables. In this Some of the desktop, notebook, and the server computers which are using operating systems such as Windows, Linux, and Macintosh can be imported directly into Microsoft Excel. By Using the pivotal tables in spreadsheet, the auditors can review the complex data in to a simplified format and drill down where the necessary to find the underlining assumptions or the information.

While we are evaluating the data mining methods, the companies and the industries may decide for acquiring several tools for different purposes, rather than purchasing the one tool that meets all needs. Although we are acquiring several tools is not a mainstream approach, and a company may choose to do so if, for example, it installs a dashboard to keep managers informed on to the business matters, and a full data-mining suite to be capture and can build the data for its marketing and sales arms, and an interrogation tool to the auditors so they can identify fraud activity.

Table 1 : Evolution of Data Mining

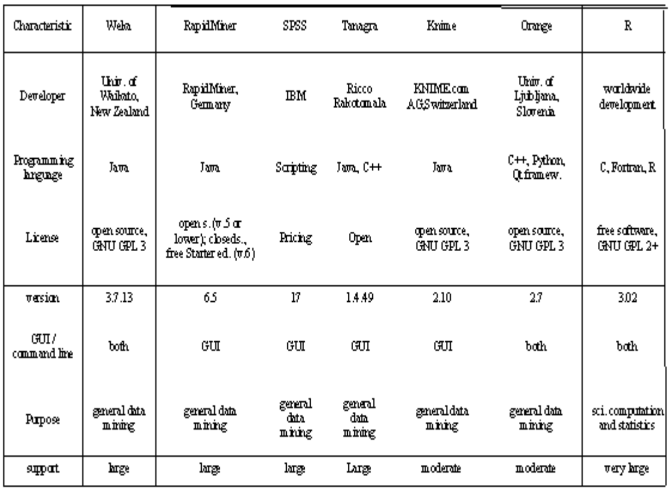


Progress	Technology	Features	Review
Data gathering(1960)	Tapes, computer	Use of static data	Problem in large storage
Data process(1980)	RDBMS,ODBC, SQL	Dynamic data at record level	Easy access
Data warehousing & Decision Making Support(1990)	OLAP, Multi-Dimensional	Dynamic data at Multiple level	Easy to understand
Data Mining	Advanced Algorithm, Big database	Proactive Information	Successful in various field

iii. Data mining tools

Most of the modern DM tools are effectively softwarebased dataflow architectures. Some of the tools (e.g. RM, and KNIME) are said to be graphical integrated environments that enable visual component placement, connection, and dragging. The most common paradigm for such components is: pull the data in, transform, and push it further down the pipeline. The components we are having only pulls the data in once all of its prerequisites are to be met. The under-the-hood implementation of such components is irrelevant for the average user, but more advanced users usually have the possibility of tackling with the code in order to improve it. Other tools (e.g. R) are just plain extensions of the underlying language in the form of specialized packages and/or GUI add-ons.General characteristics of the six DM tools are listed in Table II. All of the tools have implementations for Windows, Linux, and Mac OS X operating systems.

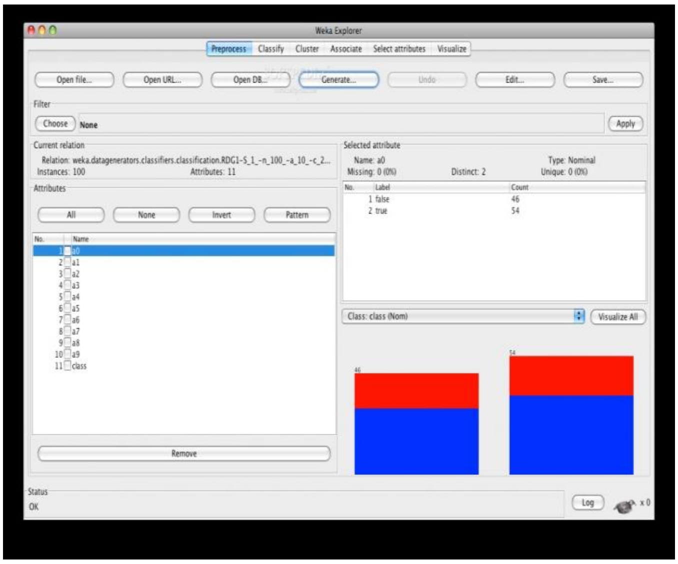
Table II. GENERAL CHARACTERISTICS OF THE DATA MINING TOOLS



Characteristic	Weka	RapidMiner	SPSS	Tanagra	Knime	Orange	R
Developer	Univ. of Waikato, New Zealand	RapidMiner, Germany	IBM	Ricco Rakotomala	KNIME.com AG,Switzerland	Univ. of Ljubljana, Slovenia	worldwide development
Programming language	Java	Java	Scripting	Java, C++	Java	C++, Python, Qt framew.	C, Fortran, R
License	open source, GNU GPL 3	open s. (v.5 or lower); closeds., free Starter ed. (v.6)	Pricing	Open	open source, GNU GPL 3	open source, GNU GPL 3	free software, GNU GPL 2+
version	3.7.13	6.5	17	1.4.49	2.10	2.7	3.02
GUI / command line	both	GUI	GUI	GUI	GUI	both	both
Purpose	general data mining	general data mining	general data mining	general data mining	general data mining	general data mining	sci. computation and statistics
support	large	large	large	Large	moderate	moderate	very large

WEKA

Weka, or the Waikato Environment is to be used for the Knowledge Analysis, which is licensed under the GNU general public license. Weka stems from the University of Waikato and is a collection of packages for the machine learning which is Java based. Weka is providing an API,by using this the developers may use the Weka from their projects. Weka is widely adopted in academic and business and has an active community (Hall et al., 2009). Weka's community is contributing many add-in packages such as the k-anonymity and l-diversity for privacy preserving the data mining and bagging and boosting of the decision trees. Tools are said to be downloaded from a repository and via the package manager. Weka is java based and extensible. Weka is also providing the .jar files which may be built into any Java application permitting custom programming outside of the Weka environment. The basic Weka environment with sample data is illustrated as Figure 1. For the big data processing, the Weka is having its own packages for the map reduce programming for maintaining independence over the platform but also provides the wrappers for the Hadoop. Weka has workflow support via its Knowledge Flow utility[4].



Filter: None

Relation: weka.datagenerators.classifiers.classification.RDG1-S_1_-n_100_-a_10_-c_2...
Instances: 100 Attributes: 11

No.	Name
1	a0
2	a1
3	a2
4	a3
5	a4
6	a5
7	a6
8	a7
9	a8
10	a9
11	class

Name: a0 Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count
1	false	46
2	true	54

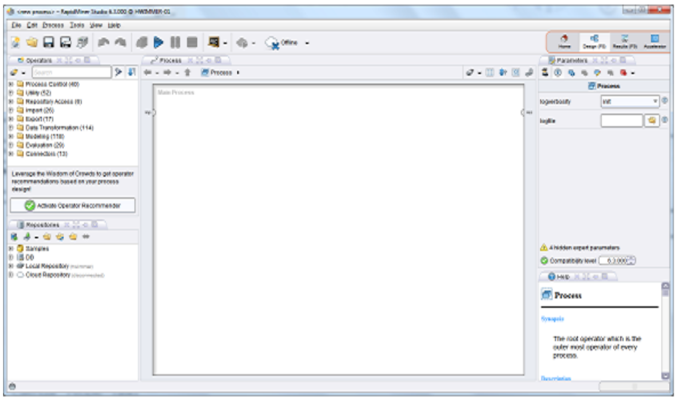
Class: class (Nom)

OK

Figure 1: The Weka Environment

Rapid Miner

RM, formerly called as Yale, has been morphed into a licensed software product which is opposed to open source; Although, the RM community edition is still free and also open source. RM have the Capability to perform the process control ,connect to a repository, to import and export the data, data transformation, modeling, and Evaluation. While we are having many features which are available in the open source version certain features are said to be not enabled. One such example is data sources. The open source version of the RM only supports the CSV and MS Excel and has no access to the database systems. Aside from data connectivity, memory access is limited to 1GB in the free starter version. RM is full-featured with the ability to visually program control structures in the process flows. In Addition to that, the modeling covers the essential methods such as the decision trees, neural networks, logistic and the linear regression, support vector machines, Naïve Bayes , and the clustering. In some cases, such as the k-means clustering, the multiple algorithms are to be implemented leaving the data scientist with the options. In this the Big data processing, and the RM's Radoop, is not available in the free edition. Finally, the ability to create workflows is well implemented in the RM environment which is shown as Figure 2[5].



Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

4 hidden expert parameters

The root operator which is the outer most operator of every process.

Figure 2: The Rapid Miner Environment

IBM SPSS

In the IBM SPSS Modeler it said to be a set of data mining tools which enable us to quickly develop the predictive models using business expertise and to deploy them into the business operations for improving decision making. And also Designed the industry-standard CRISP-DM model, IBM SPSS Modeler will be supporting the entire data mining process, from data to get better business results.

IBM SPSS Modeler is offering a variety of the modeling methods which are taken from the machine learning, artificial intelligence, and statistics. The methods available on the Modeling palette which allows you to derive the new information from your data and to develop the predictive models. Each method is having different strengths and is best suited for particular types of problems only.

As a data mining application, the IBM SPSS Modeler will be offering a strategic approach for finding the useful relationships in large data sets. In contrast for more traditional statistical methods, we do not unnecessarily need to know what are we looking for when you want to start. We can explore our data, fitting for different models and investigating the different relationships, until we find some useful information[6]

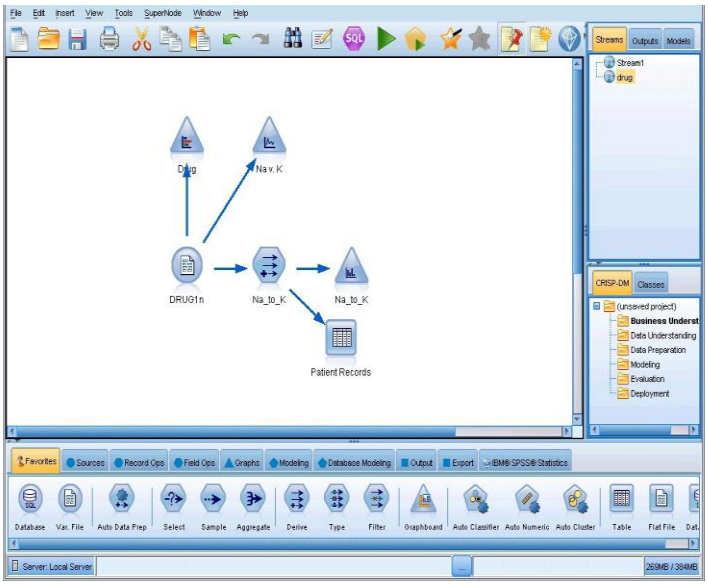

File Edit Insert View Tools SuperNode Window Help
Streams Outputs Models
Stream1
drug
Drug
Na v. K
DRUG1n
Na_to_K
Na_to_K
Patient Records
CRISP-DM Classes
(unsaved project)
Business Underst
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Favorites Sources Record Ops Field Ops Graphs Modeling Database Modeling Output Export IBM® SPSS® Statistics
Database Var. File Auto Data Prep Select Sample Aggregate Derive Type Filter Graphboard Auto Classifier Auto Numeric Auto Cluster Table Flat File Dat.
Server: Local Server
269MB / 384MB

Figure 3: The IBM SPSS Modeler Environment

Tanagra

The Tanagra is said to be an open source environment for teaching and the research and is also the successor to the SPINA software (R Rakotomalala, 2009). Capabilities aslo include Data source (reading of data), Visualization, Descriptive statistics, Instance selection, Feature selection, Feature construction, Regression, Factorial analysis, Clustering, Supervised learning, Meta-Spv learning (i.e bagging and boosting), Learning assessment, and also the Association Rules. Tanagra is said to be designed for the research and teaching; conversely, use in for profit activities is permitted based on the license agreement. One statement in the license agreement usually addresses the commercial use. The translated statement "The software is to be primarily for teaching and research. Anyone using this still can load and use, including the profit, without the payment and the royalties." Tanagra is full featured with multiple implementations of various algorithms (3 for A-Priori alone). Developed in Delphi, extending will prove difficult. Additionally, the capabilities for the big data processing are not mentioned. Finally, the workflows are possible via the diagram menu where the tasks are added and processed in a order. Figure 4 illustrates the Tanagra environment[7].

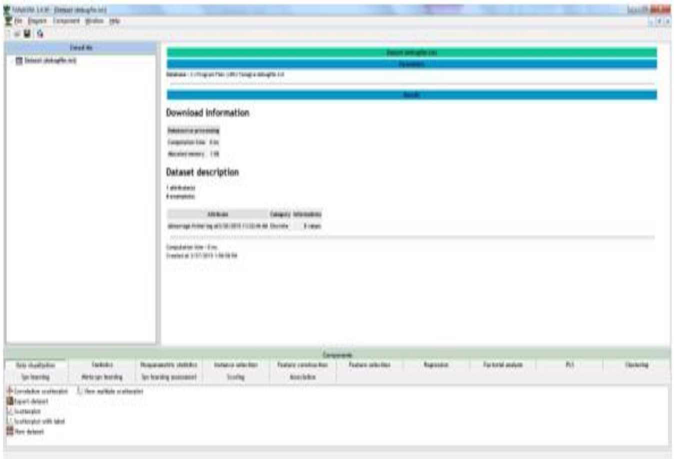


Figure 4: Tanagra Environment

KNIME

KNIME is known as Konstant Information Miner which it had its beginnings at the University of Konstanz and also has since developed into a full-scale data science tool. There are multiple versions of the KNIME each with the added capabilities. Much like the Rapid Miner, advanced capabilities and the tools will come at a smart price. It has different Functionalities which include uni variate and multivariate statistics, data mining, time series analysis, image processing, web analytics, text mining, network analysis, and social media analysis. Commercial extensions as well as the open source community are providing the extensions that may be purchased or can be downloaded. KNIME also provides an open API and is based in the Eclipse platform which facilitates developers by extending the functionalities. Additionally, we support for the Weka analysis modules and R scripts can be downloaded. KNIME boasts over 1000 analytics routines, either natively or through the Weka and R (KNIME.org, 2015). Big data processing is said to be not included in the free version but may be purchased as the KNIME Big Data Extension. Support for the workflows is to be built in to all the versions and illustrated in Figure 5 [8].

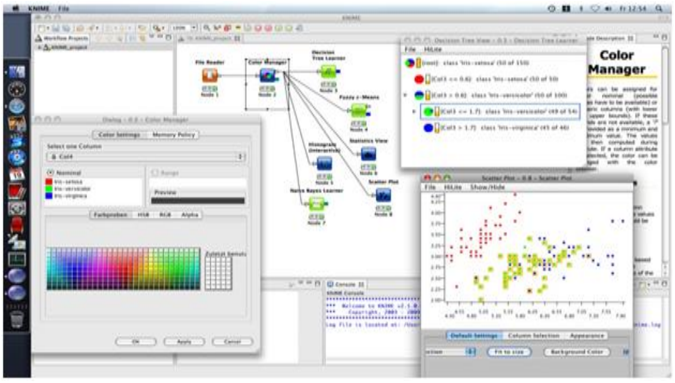


Figure 5: KNIME Environment

Orange

Orange is an open source data mining, visualization environment, analytics, and scripting environment. Figure 1 shows the Orange environment. Widgets are to be used as the building blocks to create workflows with in the Orange environment. Widgets are can be categorized as Data, Visualize, Classify, Regression, Evaluate, Associate, and Unsupervised. Data widgets are enable for data manipulation such as discretization, concatenation, and merging. Visualization widgets can perform the graphing such as the plotting, bar graphs, and the linear projection. Classification widgets are at the heart of the Orange functionality and can be employed for the multiple decision trees such as C4.5 and CART, k-nearest neighbour, support vector machines, Naïve Bayes, and the logistic regression. Regression widgets have the logistic and linear regression as well as the regression trees. Evaluation widgets contain the standard evaluations such as ROC curves and confusion matrices. Associate widgets are having the association rule mining (ARM) capabilities while unsupervised capabilities which include k-means clustering, principle component analysis (PCM), as well as a host of the other capabilities. The Orange environment, paired with its array of the widgets, supports most common data science tasks. Support for big data processing is missing; on the other hand, The Orange supports scripting in Python as well as the ability to write extension in C++. Finally, We will be creating the workflows which is a supported feature via linking the widgets together to form a data science process. Figure 6 illustrates the Orange environment [9].

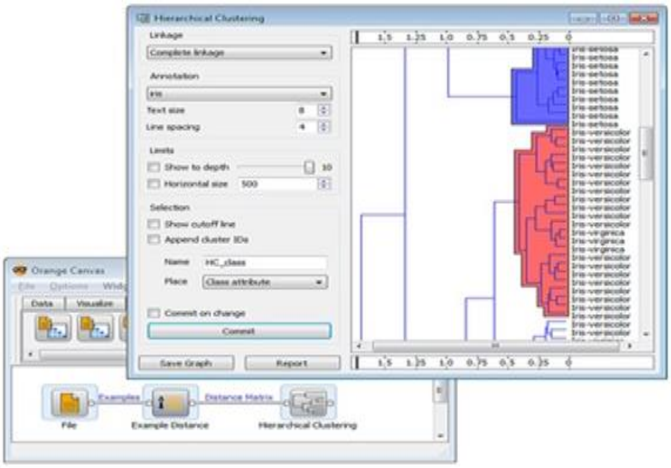

Linkage: Complete linkage
Annotation: Iris
Text size: 8
Line spacing: 4
Limits
□ Show to depth 10
□ Horizontal size 500
Selection
□ Show cutoff line
□ Append cluster IDs
Name: HC_class
Place: Class attribute
□ Commit on change

Figure 6: ORANGE Environment

R

R is a free and open source package for statistics and graphing. R is traditionally command line; however, there are many feely available open source tools that integrate into R. One such example is R Studio which provides a graphical user interface for R. R can be employed for a variety of statistical and analytics tasks including but not limited to clustering, regression, time series analysis, text mining, and statistical modeling. R is considered an interpreted language more so than an environment. R supports big data processing with RHadoop. RHadoop connects R to Hadoop environments and runs R programs across Hadoop nodes and clusters. Natively, visual features are not available making creating workflows challenging, especially for a novice; still, its broad community provides many graphical utilities such as R Studio shown as Figure 7 [10].

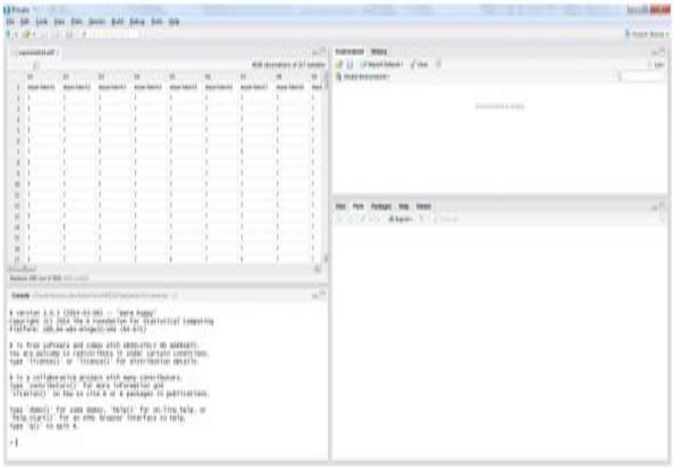
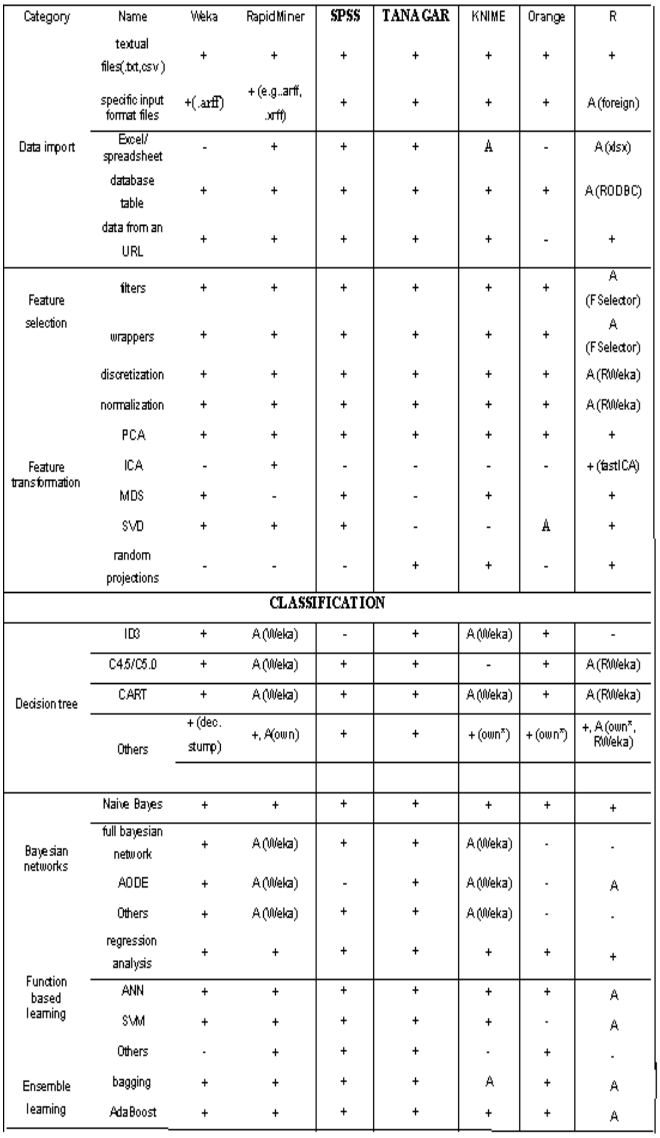


Figure 7: R Environment

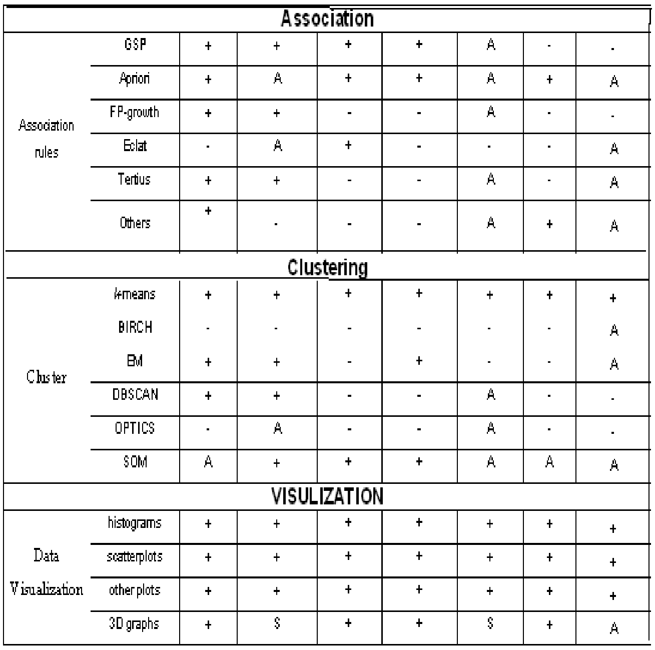
DATA MINING ALGORITHMS AND PROCEDURES SUPPORTED BY THE TOOLS

Based up on the matrix, WEKA is offering the most support for an open source basis; however, each software tool is having unique features and strengths. While R is a close second, R requires the more in-depth technical skills for executing basic tasks. Tools like Rapid Miner, KNIME, Orange, and Tanagra are providing more visual approaches; however, there is an associated cost. KNIME will be requiring a complicated installation process. Along those lines, Tanagra was developed for teaching and research; therefore, its capabilities may be outside the reach of the lay-person. Rapid Miner is only a simple installation; however, much functionality is to be removed from the open source version. Similar to Rapid Miner, Orange's visual approach and widget functionality introduces a simplified approach to creating data science tasks. One advantage to Rapid Miner is the availability of commercial support. The tools either implement an algorithm (+), use an external add-on (A) to support it, show some degree of support for the procedure (S), or do not implement it (-) at all. It must be noted here that the data in Table III should be considered temporary.

TABLE III. DATA MINING ALGORITHMS AND PROCEDURES SUPPORTED BY THE TOOLS



Category	Name	Weka	RapidMiner	SPSS	TANAGAR	KNIME	Orange	R
Data import	textual files(.txt,csv)	+	+	+	+	+	+	+
	specific input format files	+(.arff)	+ (e.g..arff, .xrff)	+	+	+	+	A (foreign)
	Excel/ spreadsheet	-	+	+	+	A	-	A (xlsx)
	database table	+	+	+	+	+	+	A (RODBC)
	data from an URL	+	+	+	+	+	-	+
Feature selection	filters	+	+	+	+	+	+	A (F Selector)
	wrappers	+	+	+	+	+	+	A (F Selector)
Feature transformation	discretization	+	+	+	+	+	+	A (RWeka)
	normalization	+	+	+	+	+	+	A (RWeka)
	PCA	+	+	+	+	+	+	+
	ICA	-	+	-	-	-	-	+ (fastICA)
	MDS	+	-	+	-	+		+
	SVD	+	+	+	-	-	A	+
	random projections	-	-	-	+	+	-	+
	CLASSIFICATION							
Decision tree	ID3	+	A (Weka)	-	+	A (Weka)	+	-
	C4.5/C5.0	+	A (Weka)	+	+	-	+	A (RWeka)
	CART	+	A (Weka)	+	+	A (Weka)	+	A (RWeka)
	Others	+ (dec. stump)	+, A(own)	+	+	+ (own*)	+ (own*)	+, A (own*, RWeka)
Bayesian networks	Naive Bayes	+	+	+	+	+	+	+
	full bayesian network	+	A (Weka)	+	+	A (Weka)	-	-
	AODE	+	A (Weka)	-	+	A (Weka)	-	A
	Others	+	A (Weka)	+	+	A (Weka)	-	-
Function based learning	regression analysis	+	+	+	+	+	+	+
	ANN	+	+	+	+	+	+	A
	SVM	+	+	+	+	+	-	A
	Others	-	+	+	+	-	+	-
Ensemble learning	bagging	+	+	+	+	A	+	A
	AdaBoost	+	+	+	+	+	+	A



Category	Name	Weka	RapidMiner	SPSS	TANAGAR	KNIME	Orange	R
	Association							
Association rules	GSP	+	+	+	+	A	-	-
	Apriori	+	A	+	+	A	+	A
	FP-growth	+	+	-	-	A	-	-
	Eclat	-	A	+	-	-	-	A
	Tertius	+	+	-	-	A	-	A
	Others	+	-	-	-	A	+	A
	Clustering							
Cluster	k-means	+	+	+	+	+	+	+
	BIRCH	-	-	-	-	-	-	A
	EM	+	+	-	+	-	-	A
	DBSCAN	+	+	-	-	A	-	-
	OPTICS	-	A	-	-	A	-	-
	SOM	A	+	+	+	A	A	A
	VISULIZATION							
Data Visualization	histograms	+	+	+	+	+	+	+
	scatterplots	+	+	+	+	+	+	+
	other plots	+	+	+	+	+	+	+
	3D graphs	+	S	+	+	S	+	A

V .DATA MINING TECHNIQUES APPLIED DOMAINS

Data mining is said to be an interdisciplinary field and with wide diverse applications. There are some nontrivial gaps between the data mining principles and the domain-specific applications, few application domains of the Data Mining are listed below.

1. Healthcare
2. Finance
3. Retail industry
4. Telecommunication
5. Text Mining
6. Web Mining
7. Higher Education

VI .CONCLUSION

We briefly reviewed the various data mining tools and their applications from its inception to the future. This review puts focus evolution and trends of data mining. Each tool has its strong points and weaknesses. Nevertheless, RapidMiner, R, Weka, and KNIME have most of the desired characteristics for a fully-functional DM platform and therefore their use can be recommended for most of the DM tasks.

VII. REFERENCES

1. Data Mining: Concepts and Techniques, Jiawei Han and Micheline Kamber,

2. P. Adriaans and D. Zantinge. Data Mining. Addison-Wesley: Harlow, England, 1996.

3. J. Bertin. Graphics and Graphic Information Processing. Berlin, 1981.

4. Insight into Data Mining Theory and Practice,K.P.Soman, Shyam diwakar, V.Ajay.

5. M. Hofmann and R. Klinkenberg, RapidMiner: Data Mining Use Cases and Business Analytics Applications, Boca Raton: CRC Press, 2013.

6. IBM SPSS Modeler 16 Applications Guide.

7. Ricco RAKOTOMALALA, "TANAGRA: a free software for research and academic purposes", in Proceedings of EGC'2005, RNTI-E-3, vol. 2, pp.697-702, 2005. (in French).

8. 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.

9. J. Demšar, T. Curk, and A. Erjavec, “Orange: Data Mining Toolbox in Python,” Journal of Machine Learning Research, vol. 14, pp. 2349–2353, 2013.
Y. Zhao, R and Data Mining: Examples and Case Studies, San Diego: Academic Press, 2012.