Comparative Analysis of Data Analysis Tools

Talha Hanif Butt, L18-1864, MS Data Science

National University of Computer and Emerging Sciences, Lahore, Pakistan

Abstract: -Today the rapid development of information technology and adoption of its several applications has created the revolution in business and various fields significantly. The growing interest in business using electronics and technology has brought vital improvement in data mining field also, since it's an important part of data accessibility. Data mining and it's applications can be viewed as one of the emerging and promising technological developments that provide efficient means to access various types of data and information available worldwide. Not only this, these applications also aids in decision making. A better understanding of these applications helps in aking choice among all available application and tools. The paper gives the comprehensive and theoretical analysis of six open source data mining tools. The study describes the technical specification, features, and specialization for each selected tool along with its applications. By employing the study the choice and selection of tools can be made easy.

I. Introduction

There has been a dramatic increase in amount of information and data which is stored in electronic format since last few decades. The size of data base has been in the process of continuous increment and has reached up to terabytes. This explosive rate of data increment is growing day by day and estimations tell that the amount of information in world doubles every 20 months. Thus the most important question concerned with data is its retrieval which finds the most suitable answer in data mining. Data mining is the process of extraction of predictive information from large data masses. It can also be described as a process of analyzing data from different perspectives and summarizing it into useful information. With a vast history deeply rooted in machine learning, artificial intelligence, database along with statistics data mining was coined very early. Data mining is strongly associated with data science which involves manipulation and classification of data by applying statistical and mathematical concepts. Data mining is an important phase in knowledge discovery and includes application of discovery and analytical methods on data to produce specific models across data. Data are available everywhere. It can be used to predict the future. Usually the statistical approach is used. Data mining is an extension of traditional data analysis and statistical approaches in that it incorporates analytical techniques drawn from a range of disciplines. Due to the widespread availability of huge, complex, information-rich data sets, the ability to extract useful knowledge hidden in these data and to act on that knowledge has become increasingly important in today's competitive world .Thus data mining is analysis of large observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to data owner. Briefly, data mining is an approach to research and analysis. It is exploration and analysis of large quantities of data in order to discover meaningful patterns and rules. Sometime, data may be in different formats as it comes from different sources, irrelevant attributes and missing data. Therefore, data needs to be prepared before applying any kind of data mining. Data mining is also known under many other names, including knowledge extraction, information discovery, information harvesting, data archeology, and data pattern processing. Many researchers and practitioners use data mining as a synonym for knowledge discovery but data mining is also just one step of the knowledge discovery process. All the techniques follow an automated process of knowledge discovery (KDD) i.e., data cleaning, data integration, data selection, data transformation, data mining and knowledge representation.

Types of data that can be mined

Flat files: Flat files are actually the most common data source for data mining algorithms, especially at the research level. Flat files are simple data files in text or binary format with a

structure known by the data mining algorithm to be applied. The data in these files can be transactions, time-series data, scientific measurements, etc.

Relational Databases: Briefly, a relational database consists of a set of tables containing either values of entity attributes, or values of attributes from entity relationships. Tables have columns and rows, where columns represent attributes and rows represent tuples. A tuple in a relational table corresponds to either an object or a relationship between objects and is identified by a set of attribute values representing a unique key.

Data Warehouses: A data warehouse as a store house, is a repository of data collected from multiple data sources (often heterogeneous) and is intended to be used as a whole under the same unified schema. A data warehouse gives the option to analyze data from different sources under the same roof.

Transaction Databases: A transaction database is a set of records representing transactions, each with a time stamp, an identifier and a set of items. Associated with the transaction files could also be descriptive data for the items. For example, in the case of the video store, the rentals table.

Multimedia Databases: Multimedia databases include video, images, audio and text media. They can be stored on extended object-relational or object-oriented databases, or simply on a file system. Multimedia is characterized by its high dimensionality, which makes data mining even more challenging. Data mining from multimedia repositories may require computer vision, computer graphics, image interpretation, and natural language processing methodologies.

Spatial Databases: Spatial databases are databases that, in addition to usual data, store geographical information like maps, and global or regional positioning. Such spatial databases present new challenges to data mining algorithms.

World Wide Web: The World Wide Web is the most heterogeneous and dynamic repository available. A very large number of authors and publishers are continuously contributing to its growth and metamorphosis, and a massive number of users are accessing its resources daily. Data in the World Wide Web is organized in inter-connected documents. These documents can be text, audio, video, raw data, and even applications. Conceptually, the World Wide Web is comprised of three major components: The content of the Web, which encompasses documents available; the structure of the Web, which covers the hyperlinks and the relationships between documents; and the usage of the web, describing how and when the resources are accessed.

Time-Series Databases: Time-series databases contain time related data such stock market data or logged activities. These databases usually have a continuous flow of new data coming in, which sometimes causes the need for a challenging real time analysis. Data mining in such databases commonly includes the study of trends and correlations between evolutions of different variables, as well as the prediction of trends and movements of the variables in time.

II. Data Mining Methods

• Classification: Supervised Learning. The classes are known.
• Clustering: Unsupervised Learning. The classes are unknown.
• Association Rule Mining: Identifying the hidden, previously unknown relation between the entities.
• Temporal mining: Use with temporal data, modeling temporal events, time series, pattern detection, sequences and temporal association rules are some tasks.
• Time Series Analysis: Describe the trend, nature and behavior of time series data. Predict the future trend and behavior of the data.
• Web Mining: Mining web data; Web content mining, Web structure mining and Web usage mining.
• Spatial Mining: Use with GIS for mining knowledge from spatial database. Spatial classification and clustering and rule generation are some task under this mining.

III. A Brief Over view of data mining tools

Data mining has a wide number of applications ranging from marketing and advertising of goods, services or products, artificial intelligence research, biological sciences, crime investigations to high-level government intelligence. Due to its widespread use and complexity involved in building

data mining applications, a large number of Data mining tools have been developed over decades. Every tool has its own advantages and disadvantages. Within data mining, there is a group of tools that have been developed by a research community and data analysis enthusiasts; they are offered free of charge using one of the existing open-source licenses. An open-source development model usually means that the tool is a result of a community effort, not necessary supported by a single institution but instead the result of contributions from an international and informal development team. This development style offers a means of incorporating the diverse experiences Data mining provides many mining techniques to extract data from databases. Data mining tools predict future trends, behaviors, allowing business to make proactive, knowledge driven decisions. The development and application of data mining algorithms requires use of very powerful software tools. As the number of available tools continues to grow the choice of most suitable tool becomes increasingly difficult. The top six open source tools available for data mining are briefed as below.

A . Weka

Waikato Environment for Knowledge Analysis. Weka is a collection of machine learning algorithms for data mining tasks. These algorithms can either be applied directly to a data set or can be called from your own Java code. The Weka (pronounced Weh-Kuh) workbench contains a collection of several tools for visualization and algorithms for analytics of data and predictive modeling, together with graphical user interfaces for easy access to this functionality.

B. KEEL

Knowledge Extraction based on Evolutionary Learning is an application package of machine learning software tools. KEEL is designed for providing solution to data mining problems and assessing evolutionary algorithms. It has a collection of libraries for preprocessing and post-processing techniques for data manipulating, soft-computing methods in knowledge of extracting and learning, and providing scientific and research methods.

C. R

Revolution is a free software programming language and software environment for statistical computing and graphics. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed.

D. KNIME

Konstanz Information Miner, is an open source data analytics, reporting and integration platform. It has been used in pharmaceutical research, but is also used in other areas like CRM customer data analysis, business intelligence and financial data analysis. It is based on the Eclipse platform and, through its modular API, and is easily extensible. Custom nodes and types can be implemented in KNIME within hours thus extending KNIME to comprehend and provide firsttier support for highly domain-specific data format.

E. RAPIDMINER

is a software platform developed by the company of the same name that provides an integrated environment for machine learning, data mining, text mining, predictive analytics and business analytics. It is used for business and industrial applications as well as for research, education, training, rapid prototyping, and application development and supports all steps of the data mining process. Rapid Miner uses a client/server model with the server offered as Software as a Service or on cloud infrastructures.

F. ORANGE

Orange is a component-based data mining and machine learning software suite, featuring a visual programming frontend for explorative data analysis and visualization, and Python bindings and libraries for scripting. It includes a set of components for data preprocessing, feature scoring and filtering, modeling, model evaluation, and exploration techniques. It is implemented in C++ and Python. Its graphical user interface builds upon the cross-platform framework.

G. SAS

SAS(Statistical Analysis System) is a product of the SAS Institute, one of the world's largest privately-owned software companies. SAS is the leading data mining tool for business analysis and is also the most expensive of the programs listed here. However, it is the one that is best suited for use in large companies. SAS is particularly good when it comes to the prognostic sector and interactive data visualization, which is ideal for large presentations. In principle, this data mining software provides a comprehensive all-round solution for successful data mining. The tool is characterized by very high scalability, so it's possible to increase the performance proportionally by adding additional hardware or other resources. This also makes it a powerful tool for high-quality business solutions. For technically less experienced users, it has a graphical user interface. However, this software can only be used free of charge if you get a corresponding license from a public institution. SAS is usually subject to a fee. The costs are decided upon request and depend on special conditions i.e. it's cheaper for authorities or educational institutions. SAS is one of the more expensive alternatives among commercial tools. However, it is possible to customize the range of functions and therefore influence the price. SAS is mainly used in pharmaceutical companies where it has established itself as standard. It is also frequently used in the banking sector and offers optimal solutions for BI and web mining. Among other things, it has its own business intelligence software for this purpose. This makes it one of the most powerful data mining tools on the market.

Name	Programming language	Operating system	Price/license
RapidMiner	Java	Windows, macOS, Linux	Freeware, Various fee-based versions
WEKA	Java	Windows, macOS, Linux	Free software (GPL)
Orange	C++, Python	Windows, macOS, Linux	Free software (GPL)
KNIME	Java	Windows, macOS, Linux	Free software (GPL)
SAS	SAS language	Windows, macOS, Linux	Limited freeware available through educational institutions
KEEL	Java	Cross Platform	GNU GPL v3
R	C, Fortran and R	Cross Platform	GNU General Public License

Name	Mode	Data Type	Data Source	Operating System
PENTAHO	Open Source	Structured and Semi-structured Data	Hadoop, NoSQL and analytic database	Windows, Linux, OSX
JASPERSOFT	Commercial, Open Source	Structured and Unstructured data	JDBC, Delimited text, Positional text, LD IF, XML	OS Independent
SPLUNK	Commercial	Unstructured data Time-series textual Machine data	Files, the network scripted outputs	Windows XP, Vista, 7 and 8

TABLEAU	Commercial	Structured and Unstructured data	Database, Cubes, Hadoop Cloud	MS Windows 8.1, Vista or Server 2012 R2, 2012, 2008 or 2003

IV. CONCLUSION

Data mining is the extraction of useful patterns and relationships from data sources, such as databases, texts, the web etc. Various free toolkits are available to understand and extrapolate data and information. This research has conducted a comparison between different data mining toolkits and web mining. The complete analysis of these data mining and web mining software tools focuses the usefulness and importance of these tools by considering various aspects. Analysis presents various benefits of these data mining tools with respect to functionalities, advantages and disadvantages, and compared them accordingly. The analysis took into account support of APIs, various database systems, PMML support, statistical analysis capabilities and visualization specific to the respective software packages. Acording to study the functionality built into to Weka and available through add-ons makes the software highly robust for a variety of users. The RapidMiner are for those users who with the skills to write code or to seek out add-ons, the software can perform many high-level functions related to the process of data mining. The description of the functions of KNIME might make it seem to be an application that is intended for those who either do not have coding and programming skills or who want something that is easy to use. NetTools Spider mining tool is basically use for the web mining purpose. Open-source data mining suites of today have come a long way from where they were only a decade ago. They offer nice graphical interfaces, focus on the usability and interactivity, support extensibility through augmentation of the source code or (better) through the use of interfaces for add-on modules. They provide flexibility either through visual programming within graphical user interfaces or prototyping by way of scripting languages. The study presented the specific details along with description of various open source data mining tools enlisting the area of specialization. With the recent endeavors of various developers concerning the use of tools in various fields one can expect a more enhanced environment along with more technical improvements. The work can be a helping hand to provide an insight in future to develop an application with more efficiency and availability i.e. a tool can be designed which instead of supporting a specific area can be extended to more fields. The effort may be increased and the development may be a complex procedure but indeed it can result in an efficient product.

REFERENCES

[1] Hand David, Mannila Heikki, Smyth Padhraic.: “Principles of data mining”, Prentice hall India, pp.1, 2004.

[2]. Sethi I. K., “Layered Neural Net Design Through Decision Trees, Circuits,and Systems”, IEEE International Symposium,1990.

[3]. Meheta M., Aggarwall R., Rissamen I. : “SLIQ:A fast Scalable Classifier for Data Mining”, In Proc. International Conference Extending data base Technology(EDBI), Avignon, France, March 1996.

[4]. Fayyad, U., Piatesky-Shapiro, G., Smyth, P., and Uthurusamy, R. (Eds.),. Advances in Knowledge Discovery and Data Mining, AAAI Press, Cambridge, 1996..

[5]. Kittipol Wisaeng . “An Empirical Comparison of Data Mining Techniques in Medical Databases”, International Journal of Computer Applications (0975 – 8887), Volume 77– No.7, September 2013.

[6]. S.R.Mulik, 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

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

[8]. Witten, I.H., Frank, E.: “Data Mining: Practical machine Learning tools and techniques”, 2nd addition,Morgan Kaufmann, San Francisco(2005).
[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(2009).
[10]. Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., Euler,T. “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] http://orange.biolab.si/features/
[12] https://github.com/Dans-labs/recommender-systems/blob/.../datamining.r
[13]. http://www.r-project.org/
[14] http://www.knime.org/
[15] http://rapidminer.com/