International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES)
Impact Factor: 3.45 (SJIF-2015), e-ISSN: 2455-2585
Volume 4, Issue 5, May-2018

A comprehensive study on Open Source Tools and Techniques of Data Mining

Juhi Singh
Amity University Haryana

Abstract: The amount of information and data flow are increasing day by day due to tremendous use of digitization. Now there challenge is to extract the meaningful information from the large stored databases and further the extracted information will help in decision making process by analyzing them. The paper gives the idea of the data mining tools and techniques to extract and analyses the information and also gives you the idea of different kind of data and their mining process .This paper also gives the comprehensive and theoretical description of data mining tools and three major classifications of tools. By applying the study of data mining tools, the selection of tools can be easy.

Keywords: Data Mining, Data Mining Tools, Open Source Tools

1. Introduction

Data mining is the process of extraction of predictive information from large database. It can also be described as a process of analyzing data from different perspectives which summarizes data into useful information. Data mining is emphatically connected with information science which includes control and characterization of information by applying measurable and scientific ideas. Data mining is an imperative stage in learning revelation and incorporates utilization of disclosure and diagnostic techniques on information to create particular models crosswise over information. Generally the factual approach is utilized. Data mining is an augmentation of conventional information examination and factual methodologies in that it consolidates scientific systems drawn from a scope of orders. Because of the across the board accessibility of enormous, complex, data rich informational collections, the capacity to remove valuable learning covered up in these information and to follow up on that information has turned out to be progressively critical now days. Information mining is a way to deal with research and examination.. [1] It is exploration and analysis of large quantities of data in order to discover meaningful patterns and rules. [2] to discover a meaningful information from the large available datasets the KDD process is being used, KDD refers to Knowledge Discovery in Databases is the process of finding knowledge in data, and emphasizes the "high-level" application of particular data mining methods.

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 [3]

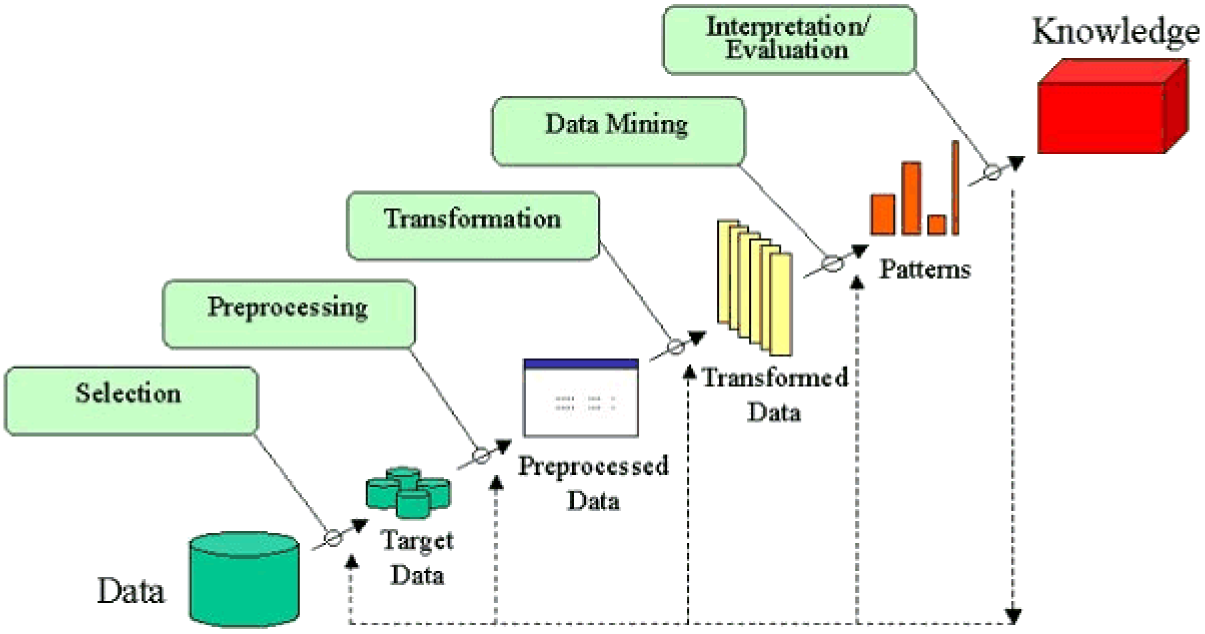

Figure1: data mining as a step in the process of knowledge discovery [2]

2. Types of data that can be mined [1]

- **Flat files**: Flat files are simple data files in text or binary format with a structure known by the data mining algorithm to be applied.
- **Relational Databases**: 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.
- **Data Warehouses**: A data warehouse as a store house is a repository of data collected from multiple data sources (often heterogeneous) and it 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. For example, in the case of the video store, the rentals table.
- **Multimedia Databases**: Multimedia databases include video, images, audio and text media and it is more challenging due to its high dimensionality, which makes data mining even more challenging.
- **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**: Data in the World Wide Web is organized in inter-connected documents. These documents can be text, audio, video, raw data, and even applications.
- **Time-Series Databases**: Time-series databases contain continuous flow of new data coming in time related data such stock market data or logged activities, which sometimes causes the need for a challenging real time analysis.

3. Data Mining Trends [4]

Here is the list of trends in data mining that reflects pursuit of the challenges such as construction of integrated and interactive data mining environments, design of data mining languages:

- Application Exploration
- Scalable and Interactive data mining methods
- Integration of data mining with database systems, data warehouse systems and web database systems
- Standardization of data mining query language
- New methods for mining complex types of data
- Data mining and software engineering
- Distributed Data mining
- Multi Database data mining
- Visual Data Mining
- Biological data mining
- Web mining
- Real time data mining
- privacy protection and Information Security in

4. Tools of data mining

Data mining tools are segment and strategy that permit end-user to mine valuable data from unstructured information. Some basic uses of data mining are instruction, design discovering, research, and promoting and misrepresentation recognition. Data mining tools use scientific and factual systems to investigate information to uncover the hidden patterns. Information Mining is a community oriented instrument which contains database frameworks, machine learning, measurement, perception data science and other train. The Data Mining devices incorporate a nonpartisan system segment that encourages order, expectation and profiling. Data mining can regularly help in making expectation about future occasions. [4]Most of the Data Mining tools can be classified into one of three major categories:

i) Traditional Data Mining Tool: Traditional data mining tools are installed on the desktop to monitor the data and highlight trends and others capture information residing outside a database.

ii)Dashboards: dashboards reflect data changes in and updates onscreen often in the form of a chart or table enabling the user to see how the business is performing.

iii)Text-mining Tools: The third type of data mining tool sometimes is called a text-mining tool because of its ability to mine data from different kinds of text from Microsoft Word and Acrobat PDF documents to simple text files, for example. These tools scan content and convert the selected data into a format that is compatible with the tool's database Scanned content can be unstructured or structured.

S.No.	Tool Name	Description
1	Knowledge Extraction based on Evolutionary Learning	KEEL [4]is an open source Java software tool that can be used for a large number of different knowledge data discovery tasks. KEEL provides a simple GUI based on data flow to design experiments with different datasets and computational intelligence algorithms (paying special attention to evolutionary algorithms) in order to assess the behavior of the algorithms. It contains a wide variety of classical knowledge extraction algorithms, preprocessing techniques (training set selection, feature selection, discretization, and imputation methods for missing values, among others), computational intelligence based learning algorithms, hybrid models, statistical methodologies for contrasting experiments and so forth.
2	RapidMiner	Rapid Miner[5] as a powerful engine for analytical ETL, data analysis, and predictive reporting, the new business analytics server. Rapid Analytics is the key product for all business critical data analysis tasks and a milestone for business analytics.
3	Weka	Weka[6] is a collection of machine learning algorithms for data mining task. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes
4	Konstanz Information Miner	KNIME[7] is a user-friendly graphical workbench for the entire analysis process: data access, data transformation, initial investigation, powerful predictive analytics, visualization and reporting. The open integration platform provides over 1000 modules (nodes)
5	Orange	Orange[8] is an Open source data visualization and analysis for novice and experts. Data mining through visual programming or Python scripting. Components for machine learning. Add-ons for bioinformatics and text mining. Packed with features for data analytics.
6	Rattle	Rattle[9] (the R Analytical Tool To Learn Easily) presents statistical and visual summaries of data, transforms data into forms that can be readily modelled, builds both unsupervised and supervised models from the data, presents the performance of models graphically, and scores new datasets.

Table 1: major of data mining tools

As per the mentioned data mining tool in table 1, the comparison is being made for the six main freely available open source data mining tools on the basis of their advantages and limitations in the following given table:

S.No.	TOOL NAME	ADVANTAGES	LIMITATIONS
1	RAPID MINER	Visualization, Statistical,Attribute Selection, Outlier detection,parameter optimization	Requires prominent knowledge of database handling
2	ORANGE	Better debugger, Shortest scripts,poor statistics,suitable for novoice Experts	Big installation, Limited reporting capabilities
3	KNIME	Molecular analysis, Mass spectrometry. Chemistry Development kit	Limited error measurements, no wrapper methods for descriptor selection,poor parameter optimazation
4	WEKA	Ease of use,can be extended in RM	Poor documentation,weak classical statistics,poor parameter optimization,weak csv reader
5	KEEL	Evolutionary algorithms,fuzzy systems	Limited algorithms
6	Rattle	Purely statistical	Less specialized for data mining, requires knowledge of array language

Table 2: advantages and limitation of six major open source tools of data mining

Of the six information mining tools that have been inspected, KNIME is the tool that would be prescribed for individuals who are amateurs to such programming to the individuals who are profoundly talented. The product is essentially exceptionally powerful with worked in highlights and with extra usefulness that can be acquired from outsider libraries. In light of the investigation, Weka would be viewed as a nearby second to KNIME due to its numerous implicit highlights that require no programming or coding information. In correlation, Rapid Miner and Orange would be viewed as proper for advanced users, especially those in the hard sciences, due to the extra programming abilities that are required, and the constrained perception bolster that is given. It can be observed from above tables that however information mining is the fundamental idea to all device yet, Rapid miner is the main tool which is free of dialect restriction and has measurable and prescient examination capacities, So it can be effectively utilized and executed on any framework, in addition it coordinates most extreme calculations of other tools capabilities, So it can be easily used and implemented on any system, moreover it integrates maximum algorithms of other mentioned tools

5. Conclusion

Open-source information mining suites of today have made some amazing progress from over the years. They offer good graphical interfaces, center around the ease of use and intuitiveness, bolster extensibility through enlargement of the source code or better using interfaces for add-on modules. They give adaptability either through visual programming inside GUI or prototyping by method for scripting dialects. The examination displayed the particular points of interest alongside depiction of different open source information mining apparatuses enrolling the region of specialization. With the ongoing undertakings of different engineers concerning the utilization of instruments in different fields one can expect a more upgraded condition alongside more specialized enhancements. The work can be some assistance to give knowledge in future to build up an application with more effectiveness and accessibility.

References:

1.Rangra et al., International Journal of Advanced Research in Computer Science and Software Engineering 4(6), June - 2014, pp. 216-223

2. 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.

3. 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.

4IJERTV3IS10024

4. I. Triguero, S. González, J. M. Moyano, S. García, J. Alcalá-Fdez, J. Luengo, A. Fernández, M. J. del Jesus, L. Sánchez, F. Herrera. KEEL 3.0: An Open Source Software for Multi-Stage Analysis in Data Mining International Journal of Computational Intelligence Systems 10 (2017) 1238-1249

5. https://rapidminer.com/data-mining-tools-try-rapidminer/

6. https://www.cs.waikato.ac.nz/~ml/weka/

7. http://www.knime.org/

8. http://orange.biolab.si/features/

9. http://www.r-project.org/