www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3 Issue 12 December 2014, Page No. 9450-9452

A Study on Free Open Source Data mining Tools

K. Saravanapriya[1]

[1] PG Department of Computer Applications,
Sacred Heart College (Autonomous)
Tirupattur, Vellore District, Tamilnadu, India
rajpriya2109@gmail.com

Abstract: ***Today in this real world, where we deal with vast amounts of data, we are in a situation to extract the best among them. Data mining is a new and imminent trend to discover the information and knowledge from the existing data. The gained knowledge can be used for applications ranging from market analysis, fraud detection, and customer retention, to production control and science exploration. It uses machine learning, statistical and visualization techniques to discover the present knowledge in the form which is easily understandable by the users. The data mining tools are used to predict the future trends and behaviours of various business organisations and aid them to make good decisions for their growth and development. They search the databases for hidden patterns, and to make predictions beyond the mind-set of the experts. This paper explains about various open source data mining tools through which efficient predictive analysis can be done.***

Keywords: Data mining, Nearest Neighbour, Orange, Rapid Miner, Weka, Jhepwork, Knime .

1. Introduction

Nowadays the data sources such as database, data warehouse, flat files and other data repositories contain a vast amount of data and information which is impossible to justify as a valuable one for good decision making process. Data mining is the process of discovering interesting knowledge from large amounts of data stored in databases, data warehouses, or other information repositories [1]. Data mining involves an integration of techniques from multiple disciplines such as database and data warehousing technology, statistics, machine learning, high-performance computing, pattern recognition, neural networks, data visualization, information retrieval, image and signal processing, and spatial or temporal data analysis [2]. Data mining has many application fields such as marketing, business, science and engineering, economics, games and bioinformatics. Such kind of urge in knowledge discovery led to the development of data mining tools which will assist us in transforming those vast amounts of data into useful information and knowledge. Different kinds of data mining tools are now available in market for commercial purpose, one has to find which best suits his organization for a successful decision making process. This paper reviews various free open source tools available at present.

2. Data Mining Techniques and their Applications

In addition to using a particular data mining tool, internal auditors can choose from a variety of data mining techniques. The most commonly used techniques include artificial neural networks, decision trees, and the nearest-neighbor method. Each of these techniques analyses data in different ways [3].

2.1 Artificial neural networks

Artificial neural network is a non-linear, predictive model that learns through training. Apart from their complex structure, long training time, and uneasily understandable representation of results, they have high acceptance ability for noisy data and high accuracy that are preferred in data mining [4]. It is particularly used in the areas of fraud detection.

2.2 Decision trees

A decision tree is a flow-chart-like tree structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes or class distributions [2]. It is a predictive model that maps observations about an item to conclusions about the item's target value. They can be easily converted to classification rules. It can be used to visually and explicitly represent decisions and decision making. In data mining, a decision tree describes data but not decisions.

2.3 The Nearest - Neighbor method

Nearest-Neighbor method is based on learning by analogy, that is, it compares a given test tuple with training tuples that are similar to it. The training tuples are described by *n* attributes. Each tuple represents a point in an *n*-dimensional space so that, all the training tuples are stored in an *n*-dimensional pattern space. When given with an unknown tuple, it searches the pattern space for the *k* training tuples that are closest to the unknown tuple. These *k* training tuples are the "*k* nearest neighbors" of the unknown tuple [2]. It is used to define a

document that is fascinating to the user and ask the system to search for similar items.

These techniques need a good and flexible data mining tool to be applied in. Though there are many tools available with efficient features, it is not economically feasible for the organisations to switch over various tools for proper decision making and research scholars to buy new tools often for their research work. So here are some free open source tools which helps them to make good analysis and proper decisions.

3. Open source Tools

3.1 Orange

Orange is a component - based data mining and machine learning software suite, which features visual programming as a front-end for an explorative data analysis. It remembers the frequently used items sets and selects the best communication channel for the widget to be used. It is packed with various visualization methods such as bar charts, scatter plots to networks, dendrograms, python bindings and libraries for scripting. It includes a set of components for data pre-processing, feature scoring and filtering, modelling, model evaluation, and exploration techniques. It is implemented in C++ and Python. Its graphical user interface is built upon the cross-platform Qt framework. Orange is distributed free under the GPL [3].

3.2 Rapid Miner

RapidMiner is a software platform developed by the rapid miner company. It provides a unified background for machine learning, data mining, text mining, predictive analytics and business analytics. It is mainly used for business and industrial applications as well as for research, education, training, rapid prototyping, and application development and supports all the steps of the KDD including results such as visualization, validation and optimization.

Figure 1: Rapid Miner

It is formerly known as YALE (Yet Another Learning Environment). It uses a client/server model with the server as the Software as a Service (SAAS) or on a cloud infrastructure.

FIGURE 2: Rapid Miner

3.3 Weka

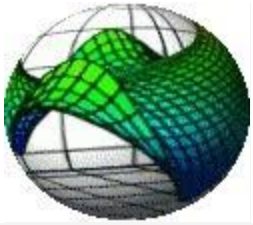
Weka is the Waikato Environment for Knowledge Analysis. It is developed by the Machine Learning Group at the University of Waikato available under the GNU public license. It holds the name of a bird seen on the islands of New Zealand [6]. Its oldest version tool was basically designed to analyse the data for agricultural domain, but its new version was fully Java-based and is now used in various application areas, especially for educational purposes and research. Its algorithm can be directly applied to a data set or as a code written in java. It supports several data mining tasks, such as data preprocessing, clustering, classification, regression, visualization, and feature selection. It is specially designed for machine learning systems.

Figure 3: Weka GUI Chooser

The techniques of Weka depends on the types of the data that are available as a flat file or a relative data that points to a fixed number of attributes. [6]. The user can interact through the Weka explorer or use command line with the same functionality or with the component-based knowledge flow face.

Figure 4: Weka

3.4 JHepWork



jHepWork is a framework developed as a trial to make a data-analysis environment with a clear user interface to compete the viable suites. It is primarily designed for scientists, engineers and students. It focusses on interactive scientific plots in 2D and 3D and contains numerical scientific libraries implemented in Java for mathematical functions, random numbers, and other data mining algorithms [7]. jHepWork is based on a high-level programming language Jython, but also suits programs such as Java to call numerical and graphical libraries.

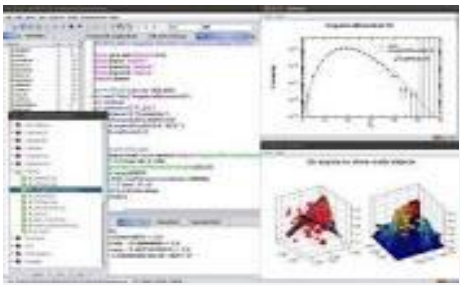


Figure 5: JHepWork

3.5 Knime

KNIME (Konstanz Information Miner) is a user friendly, knowledgeable and comprehensive data integration, processing, analysis, and exploration platform. It provides the users to create data flows or pipelines visually, users can selectively execute some or all analysis steps, study the results, prototypes, and collaborative interpretations. KNIME is written in Java, and based on Eclipse.

Figure 4: Knime

4. Conclusion

Though there are many open source tools available for data mining applications, these tools have profoundly led their way for the in-depth observations of various analysis and prediction. They also had a major impact on excellent decision making in some marketing firms.

References

[1] Goebel, M., Gruenwald, L., A survey of data mining and knowledge discovery software tools, ACM SIGKDD Explorations Newsletter, v.1 n.1, p.20-33,June 1999 [doi>10.1145/846170.846172].

[2] Han, J., Kamber, M., Jian P., Data Mining Concepts and Techniques. San Francisco, CA: Morgan Kaufmann Publishers, 2011.

[3] http://orange.biolab.si/

[4] David Norris, "RapidMiner - a potential game changer," Bloor Research, November 13, 2013

[5] Ajay Ohri, "Interview with Rapid-I Ingo Mierswa and Simon Fischer," *KDnuggets*, August 2011

[6] G. Holmes; A. Donkin and I.H. Witten (1994). "Weka: A machine learning workbench". Proc Second Australia and New Zealand Conference on Intelligent Information Systems, Brisbane, Australia. Retrieved 2007-06-25.

[7] S.Chekanov: *jHepWork*. HEP choices for data analysis. Jan 21, 2008.

[8] www.knime.org