Tools for Predictive Analytics: An Overview

Razeef Mohmmad[1], Muheet Ahmed Butt[2], Majid Zaman[3]
[1](Ph.D Scholar, PG Department of Computer Sciences, University of Kashmir, Srinagar, J&K, India)
[2] (Scientist, PG Department of Computer Sciences, University of Kashmir, Srinagar, J&K, India)
[3](Scientist, Directorate of Information Technology and Support Systems, University of Kashmir, Srinagar, J&K, India)

ABSTRACT
Predictive analytics deals with analysis of historical data to uncover hidden pattern and provide new insights using techniques from statistics, modeling, machine learning and artificial intelligence. Predictive analytic tools help in the process of extraction and prediction which help the users in better decision making. Predictive Analytic tools provide business users with simple self explanatory charts, graphs and scores which help them to understand the likelihood of possible outcomes. In this work we have presented few popular state-of-the-art predictive analytic tools with their most important feature and approach and also the programming language in which they are implemented.

Keywords *- Predictive Analytics, Predictive Analytic Tools, WEKA, RAPIDMINER, Orange.*

1. INTRODUCTION

Predictive Analytics is branch of data mining which involves the analysis and extraction of useful hidden information from large data sets to determine patterns and predict future trends and probabilities of occurrences of events. Predictive Analytics encompasses a variety of statistical techniques from predictive modeling, machine learning, and data mining. In Predictive Analytics, Predictive model is generated to analyze current and historical facts to make predictions about future or otherwise unknown events [1].
There are various tools available in the market that helps in the process of predictive analytics. Few decades ago using Predictive Analytic tools as well as the results they generated require advance skill and knowledge [9][10][11]. Most of them were tough to use and provide command line interface. The history of software packages for data mining is short but eventful. Although the term data mining was coined in the mid-1990s [2], statistics, machine learning, data visualization, and knowledge engineering–research fields that contribute their methods to data mining were at that time already well developed and used for data exploration and model inference [3]. A modern predictive analytics tool does not require any data specialists to operate and analyze the results [12][13][14][15]. Today predictive analytic tools are interactive and attractive with user friendly interface and approach.

2. CATEGORIES OF PREDICTIVE ANALYTIC TOOLS

Predictive analytic tools are broadly divided into two main categories:

a) Open Source and freeware Predictive Analytic Tools.
b) Proprietary commercially available Predictive Analytic Tools

2.1 Open Source and freeware Predictive Analytic Tools:

Apache Mahout, GNU Octave, KNIME, OpenNN, Orange, R, Scikit-learn, WEKA, RapidMiner, Tanagra, Data Science Studio (DSS), H2O, Lavastorm Public Edition, LIBLINER are some of the predictive open source and freeware software.

2.2 Proprietary commercially available Predictive Analytic Tools

Alpine Data Labs, Alteryx, Angoss Knowledge STUDIO, BIRT Analytics, MATLAB, IBM SPSS Statistics and IBM SPSS Modeler, KXEN Modeler, Mathematica, Minitab, Neural Designer, Oracle Data Mining (ODM), Pervasive, Predixion Software, RCASE, Revolution Analytics, SAS and SAS Enterprise Miner, STATA, Statgraphics, STATISTICA, TeleRetail, TIBCO are some of the Proprietary commercially available Predictive Analytic Tools.

3. POPULAR PREDICTIVE ANALYTIC TOOLS COMMONLY USED:

Although there is plethora of predictive analytic tools available in market but few Predictive Analytic tools are worth mentioning and commonly used for predictive analytics.

3.1 R

R (http://www.r-project.org) is an open source software tool for statistical computing and graphic. It is widely used by statisticians and data miners for predictive analytics and modeling and also in developing statistical software and data analysis. It provides an extensive methods and techniques for statistical means, predictive modeling, data analytics and visualization. It has become a de facto standard open source library for statistics and modeling. Most of its computationally intensive methods are efficiently implemented in C, C++, and FORTRAN, and then interfaced to R, a scripting language similar to the S language originally developed at Bell Laboratories [4]. It is free software licensed under the GNU General Public License In bioinformatics for genomic data analysis, there is an R library and software development project called Bioconductor (www.bioconductor.org). In general, R is a well-supported, open source, command line driven, statistics package used by statistician and predictive analysts.

3.2 WEKA

Waikato Environment for Knowledge Analysis (Weka) is a popular data mining machine learning software, developed at the University of Waikato New Zealand. It is free software licensed under the GNU General Public License available on www.cs.waikato.ac.nz/ml/weka/ . Weka is a Java based open source predictive analytic tool consisting of many data mining and machine learning algorithms, including pre-processing on data, classification, clustering, and association rule extraction [5]. Weka provides three graphical user interfaces:

1. Explorer for exploratory data analysis to support preprocessing, attribute selection, learning, visualization
2. Experimenter for experimental environment for testing and evaluating machine learning algorithms, and
3. Knowledge Flow for new process.

Weka works with data file with formats of ARFF, CSV, and C4.5, binary.

3.3 RAPIDMINER

Rapidminer (http://www.rapidminer.com) is an open source GUI based data mining tool which provides an environment for machine learning and data mining processes. It provides an integrated environment for data mining, text mining, predictive analytics, business analytics and machine learning. Rapidminer is platform independent (Cross platform) software i.e. can be installed on any operating system. It is licensed by AGPL Proprietary and can be downloaded from www.rapidminer.com. Rapidminer supports about twenty two file formats [6]. It implements many procedures for model creation and evaluation. Rapidminer contains more than 100 learning schemes for regression classification and clustering analysis [7]. This data mining software works well with different database files and can read/write excel files also. This software is popular among data analysts, data miner and statisticians.

3.4 KNIME

KNIME (http://www.knime.org), Konstanz Information Miner is an open source data mining tool written in java and based on Eclipse platform. It is licensed by GNU General Public License and is compatible with Linux, OS X, and Windows operating systems. KNIME is specialized software used for, pharmaceutical research, Enterprise reporting, Business Intelligence, data mining and predictive analytics. It has modular data exploration platform that enables the user to visually create data flows (often referred to as pipelines), selectively execute some or all analysis steps, and later investigate the results through interactive views on data and models. The one aspect of KNIME that truly sets it apart from other data mining packages is its ability to interface with programs that allow for the visualization and analysis of molecular data [8].

3.5 ORANGE

Orange (http://www.orange.biolab.si) is an open source, component-based data mining and analytic machine learning tool for novice and experts. Orange is cross platform GUI based software, specialized for data visualization and data mining. It is licensed by GNU General Public License, implemented in C/C++ and Python. Orange data mining software provides data processing components, filtering and scoring, model creation and evaluation with exploration techniques. Data mining in Orange is done through visual programming or Python scripting. It allows visual programming front-end for explorative data analysis and visualization, and Python bindings and libraries for scripting in the back-end. It has shortest script for doing training, cross validation, algorithms comparison and prediction [5].

4. CONCLUSION

There are various predictive analytic tools available in the market that the business and research organizations can integrate into their computer system to optimize the activities and decision making. In this work, we have mentioned two main categories of the predictive analytic tools. Five commonly used predictive analytic tools are discussed mentioning their benefits and language used. We have broadly stated the features and significance of the tools so that one can use them according to the needs and available budget. All the predictive analytic tools mentioned were found to be very user-friendly and quite easy to adopt and implement.

REFERENCES

[1] https://en.wikipedia.org/wiki/Predictive_analytis#cite_note-buettner2016h-1

Thesis:

[2] Fayyad U, Piatetsky-Shapiro G, Smyth P, et al, editors. Advances in knowledge discovery and data mining. Menlo Park (CA): AAAI Press; 1996.

[3] Zupan Blaz, Demsar Janez; "Open-Source Tools for Data Mining", University of Ljubljana, Trzaska 25, SI-1000 Ljubljana, Slovenia, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA.

[4] Becker RA, Chambers JM. S: an interactive environment for data analysis and graphics. Pacific Grove (CA): Wadsworth & Brooks/Cole; 1984.

Journal Papers

[5] Rangra, Kalpana, and K. L. Bansal. "Comparative study of data mining tools," *International Journal of Advanced Research in Computer Science and Software Engineering* 4.6 (2014): 216-223.

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

[7] 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.

[8] Rangra Kalpana, Bansal K. L," Comparative Study of Data Mining Tools", *International Journal of Advanced Research in Computer Science and Software Engineering*, Volume 4, Issue 6, June 2014.

[9] Butt, Muheet Ahmed, and Majid Zaman. "Assessment Model based Data Warehouse: A Qualitative Approach." *International Journal of Computer Applications 62.10 (2013).*

[10] Zaman, Majid, and Muheet Ahmed Butt. "Enterprise Data Backup & Recovery: A Generic Approach." *International Organization of Scientific Research Journal of Engineering (IOSRJEN) (2013): 2278-4721.*

[11] Butt, Er Muheet Ahmed, S. M. K. Quadri, and Er Majid Zaman. "Star Schema Implementation for Automation of Examination Records." *Proceedings of the International Conference on Frontiers in Education: Computer Science and Computer Engineering (FECS).* The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2012.

[12] Zaman, M., S. M. K. Quadri, and Er Muheet Ahmed Butt. *"Information Integration for Heterogeneous Data Sources." IOSR Journal of Engineering 2.4 (2012): 640-643.*

[13] Butt, M. A., and M. Zaman. "Data quality tools for data warehousing: enterprise case study." *IOSR Journal of Engineering 3.1 (2013): 75-76.*

[14] Zaman, Majid, and Muheet Ahmed Butt. "Enterprise Management Information System: Design & Architecture." *International Journal of Computational Engineering Research (IJCER), ISSN 2250 (2013): 3005.*

[15] Butt, Muheet Ahmed. *"Information extraction from pre-preprinted documents."* Energy 20.8 (2012): 729-743.