View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by International Journal of Innovative Technology and Research (IJITR)

Ramachandra Rao Meka* et al.
(IJITR) INTERNATIONAL JOURNAL OF INNOVATIVE TECHNOLOGY AND RESEARCH
Volume No.3, Issue No.4, June - July 2015, 2229 – 2233.

# Performing Predictive Data Analytics in Data Mining Using Various Tools

**RAMA CHANDRA RAO MEKA**
CSE Department
AGMR College of Engineering & Tech Varur,
Hubli-581207

**Dr.NOORULLAH SHARIFF C**
ISE Department
Ballari Institute of Technology and Management
Alipur, Bellary-583104

**AMARESH PATIL**
CSE Department
AGMR College of Engineering & Tech Varur,
Hubli-581207

*Abstract*— **Predictive Data Analytics is a branch of Data Mining. Performing Predictive Data Analytics on huge data sets will help us in quick Decision Making forecasting on the results obtained on live or sample data. Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies or individuals to focus on the most important information in their data warehouses. This paper helps us in specifying how to do Predictive Data Analytics in data mining using various tools. There are various Open Source Tools which help us in performing Predictive Analytics such as R Studio, Weka, KNIME etc. This paper also lists various predictive analytic tools and specify there features and usage. A comparison also can be made or decision can be taken by the reader to use a specific tool based on the requirement. The main scope is to enhance the study of predictive data analysis and provide the necessary help in quick decision making in any of the important area. Predictive data analytics can be performed in various areas such as medical, agriculture, behavior prediction of kids, behavior of a customer in a particular business etc. In this aspect the paper elaborates on the tools available to do perform predictive data analytics and also introduce the importance of data mining. Predictive data analysis is done using variables as attributes known as predictors. This paper also includes information about Knowledge Discovery Process which is a part of Data Mining.**

*Keywords*— **Data Mining, Predictive Analytics, Data Sets, Decision Making, Attributes, Predictor, Knowledge Discovery.**

## I. INTRODUCTION

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies or individuals to focus on the most important information in their data warehouses. We can predict future trends and behaviors using various Data mining tools, allowing businesses to make proactive, knowledge-driven decisions. Mainly Predictive Analytics results help in decision making which is an important aspect in any business or individuals. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. We can make quick decisions and save lot amount of time and apply immediately on the prediction results.

Data collection is the major task in performing predictive data analytics. Most companies already collect and refine massive quantities of data. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems as they are brought on-line. When implemented on high performance client/server or parallel processing computers, data mining tools can analyze massive databases to deliver answers to questions such as, "Which clients are most likely to respond to my next promotional mailing, and why?"

Predictive analytics is the branch of data mining concerned with the prediction of future probabilities and trends. The central element of predictive analytics is the predictor, a variable that can be measured for an individual or other entity to predict future behavior. For example, an insurance company is likely to take into account potential driving safety predictors such as age, gender, and driving record when issuing car insurance policies.

Multiple predictors are combined into a predictive model, which, when subjected to analysis, can be used to forecast future probabilities with an acceptable level of reliability. In predictive modeling, data is collected, a statistical model is formulated, predictions are made and the model is

 @ 2013 http://www.ijitr.com All rights Reserved. 

validated (or revised) as additional data becomes available. Predictive analytics are applied to many research areas, including meteorology, security, genetics, economics, and marketing [1].

## II. SCOPE OF DATA MINING

Data mining derives its name from the similarities between searching for valuable business information in a large database — for example, finding linked products in gigabytes of store scanner data — and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find exactly where the value resides. Given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing these capabilities:

### *A. Automated prediction of trends and behaviors.*

Data mining automates the process of finding predictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly from the data — quickly. A typical example of a predictive problem is targeted marketing. Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events [2].

### *B. Automated discovery of previously unknown patterns.*

Data mining tools sweep through databases and identify previously hidden patterns in one step. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying error.

## III. KNOWLEDGE DISCOVERY PROCESS

Knowledge Discovery in Databases (KDD) is an automatic, exploratory analysis and modeling of large data repositories.KDD is the organized process of identifying valid, novel, useful, and understandable patterns from large and complex data sets. The knowledge discovery process consists of six stages:

- Data Selection
- Cleaning
- Enrichment
- Coding
- Data mining
- Reporting

### *A. Data Selection*

Data Selection can be explained with an example containing records of subscription data for the magazines. It is a selection of operational data from the publisher's invoicing system and contains information about people who have subscribed to a magazine. The records consist of: client number, name, address, date of subscription, and type of magazine. In KDD process it initiates a copy of this operational data is drawn and stored in a separate database. Contents are illustrated in the below table

### *B. Cleaning*

There are several types of cleaning process, some of which can be executed in advance while others are invoked only after pollution is detected at the coding or the discovery stage. There is an important element in a cleaning operation i.e. de-duplication of records. The reason of de-duplication happens due to many cases such as people making typing errors, or of clients moving from one place to another without notifying the change of address. There are also cases in which people deliberately spell their names incorrectly or give incorrect information about themselves, especially in situations where individuals have been refused some type of insurance etc. These cases are with respect to the example cited. Data mining and data cleaning are two different concepts; pattern recognition algorithm can be applied in cleaning the data.

### *C. Enrichment*

In the example cited we will suppose that we have purchased extra information about our clients consisting of date of birth, income, amount of credit, and whether or not an individual owns a car or a house as shown below. For this example it is how ever not necessary to appreciate that the new information can easily be joined to the existing client records.

### *D. Coding*

The data in the example can undergo a number of transformations. First the extra information that was purchased to enrich the database is added to the records describing the individuals. In the next stage, we select only those records that have enough information to be of value as shown. A general rule states that any deletion of data must be a conscious decision, after a thorough analysis of the people consequences.

### *E. Data Mining*

The discovery stage of the KDD process is fascinating. Data mining is not so much a single technique as the idea that there is more knowledge hidden in the data than shows itself on the surface.

 @ 2013 http://www.ijitr.com All rights Reserved. 

From this point of view, data mining is really an "anything goes" affair. Any technique that helps extract more out of your data is useful, so data mining techniques from quite a heterogeneous group.

### *F. Reporting*

This stage is helpful in viewing the results using various reports. Reports could be view in pictorially using various graphs.

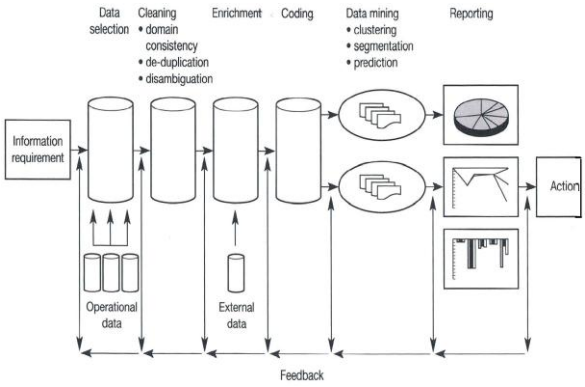


***Figure 3.1: Knowledge Discovery Process***

## IV. TOOLS FOR PREDICTIVE DATA ANALYSIS

### *A. Data Mining Tools*

There are various effective software tools for Data Mining that can help to find the relationships, clusters, patterns, categorizing, summarizing, etc. from the huge data sets. Such data mining tools can help one to take most accurate decisions which come out profitable for their business.

### *B. Categories of Data Mining Tools*

There are many tools used for Data Mining. They are broadly classified into three categories Traditional data mining tools, Dashboards and text-mining tools.

### *C. Traditional Data Mining Tools*

Traditional mining programs help the companies in establishing data patterns and trends by using various complex algorithms and techniques. Some of these tools are installed on the desktop computers to monitor the data and emphasize trends and others capture information residing outside a data base. Majority of these programs are supported by windows and UNIX versions. However, some software specializes in one operating system only. In addition to that some may work in only one database type. But, Most of the software will be able to handle any data using online analytical processing or a similar technology.

### *D. Dashboards*

Dashboards reflect data changed and update on screen. Dashboards are normally installed in computers to monitor information in a database and it reflects data changes and updates the data in the form of a chart or table on the screen. It enables the user to see how the business is performing. Historical data can be referenced and checks against the current status in order to see the changes in the business. By this way, dashboards is very easy to use and helps the manager a lot with great appeal to have an overview of the company's performance.

### *E. Text-Mining Tools*

The third type of data mining tools is called as a text-mining tool because of its ability to mine data from different kind of text starting from Microsoft Word, Acrobat PDF documents to simple text files. This provides facility of scanning the content and converts the selected into a format that is compatible with the tools database without opening different applications.

## V. OPEN SOURCE TOOLS FOR DATA MINING

### *A. R*

R is an open source programming language and environment for statistical computing and graphics. R provides a wide variety of graphical and statistical techniques such as linear and non-linear modeling, classical statistical tests, series analysis, classification clustering and is highly extensible. Researchers in various fields of applied statistics have adopted R for statistical software development and data analysis. Extensibility and superb data visualization are the two main reasons for the success of R [7].

R, a GNU project, is written in R itself? It's primarily written in C and FORTRAN. And a lot of its modules are written in R itself. It's a free software programming language and software environment for statistical computing and graphics. The R language is widely used among data miners for developing statistical software and data analysis. Ease of use and extensibility has raised R's popularity substantially in recent years.

Besides data mining it provides statistical and graphical techniques, including linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, and others.

### *B. Weka*

Weka is a collection of machine learning algorithms for data mining tasks and well suited for

 @ 2013 http://www.ijitr.com All rights Reserved. 

developing new machine learning schemes. Weka is a java based software capability of working under various operating systems and contains tools for data pre-processing, classification, regression, clustering, association rules and visualization. The algorithms can either be applied directly to a dataset or called from a user's java code. Weka is probably the most successful open source data mining software which has inspired by the development of other programs with more sophisticated graphical user interface and better visualization methods [6].

The original non-Java version of WEKA primarily was developed for analyzing data from the agricultural domain. With the Java-based version, the tool is very sophisticated and used in many different applications including visualization and algorithms for data analysis and predictive modeling. Its free under the GNU General Public License, which is a big plus compared to Rapid Miner, because users can customize it however they please.

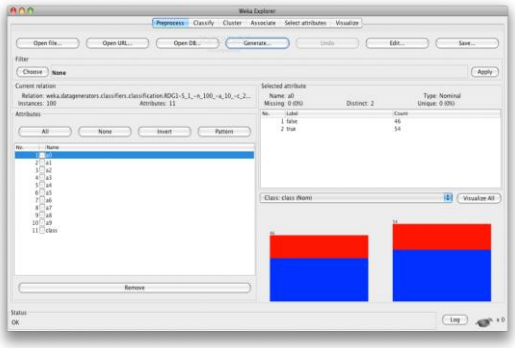

***Figure 5.1: Sample result with Weka tool [7].***

Weka supports several standard data mining tasks, including data preprocessing, clustering, classification, regression, visualization and feature selection. Weka would be more powerful with the addition of sequence modeling, which currently is not included.

### *C. Orange*

Orange is an open source data mining and visualization software with active community and which helps novice and experts for their analysis. It has the ability to work under various platforms like windows, Mac Os C and GNU/Linux operating systems and it's packed with data analytics features. It enables design of data analysis process through user friendly visual programming or python scripting. Hence, this can be used as a scripting language for respective tasks of data mining. It represents most major algorithms for data mining and contains different visualization, from scatter plots, bar charts, trees to dendrograms, networks and heat maps. It remembers user's choices, suggests most frequently used combinations, and intelligently chooses which communication channels to use. It has specialized add-ons like Bioorange for bio informatics [5].

Python is picking up in popularity because it's simple and easy to learn yet powerful. Hence, when it comes to looking for a tool for your work and you are a Python developer, look no further than Orange, a Python-based, powerful and open source tool for both novices and experts.

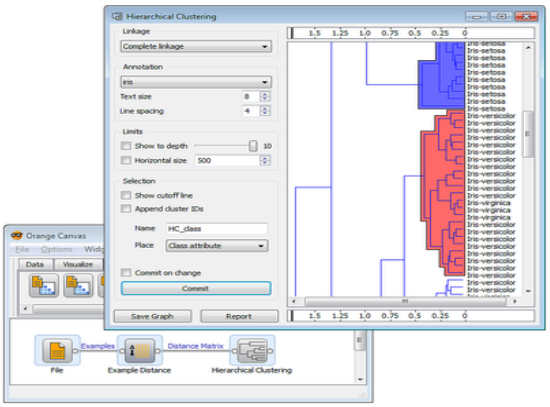

***Figure 5.1: Hirarchical clustering sample result with Orange tool [7].***

### *D. KNIME*

Data preprocessing has three main components: extraction, transformation and loading. KNIME does all three. It gives you a graphical user interface to allow for the assembly of nodes for data processing. It is an open source data analytics, reporting and integration platform. KNIME also integrates various components for machine learning and data mining through its modular data pipelining concept and has caught the eye of business intelligence and financial data analysis.

Written in Java and based on Eclipse, KNIME is easy to extend and to add plug-in. Additional functionalities can be added on the go. Plenty of data integration modules are already included in the core version.

## VI. REFERENCES

[1] Philip K. Chan, Florida Institute of Technology Wei Fan, Andreas L. Prodromidis, and Salvatore J. Stolfo, Columbia University "Distributed Data Mining in Credit Card Fraud Detection" IEEE intelligent systems.

[2] P. Chan and S. Stolfo, "Metalearning for Multistrategy and Parallel Learning," *Proc. Second Int'l Workshop Multistrategy Learning*, Center for Artificial Intelligence, George Mason Univ., Fairfax, Va., 1993, pp. 150–165

[3] P. Chan and S. Stolfo, "Toward Scalable Learning with Nonuniform Class and Cost Distributions: A Case Study in Credit Card Fraud Detection," *Proc. Fourth Int'l Conf.*

 @ 2013 http://www.ijitr.com All rights Reserved. 

*Knowledge Discovery and Data Mining*, AAAI Press, Menlo Park, Calif., 1998, pp. 164–168

[4] http://www.cid.harvard.edu/ciddata/ciddata.html

[5] http://thenewstack.io/six-of-the-best-open-source-data-mining-tools/

[6] S. Stolfo et al., "*JAM*: Java Agents for Metalearning over Distributed Databases," *Proc. Third Int'l Conf. Knowledge Discovery and Data Mining*, AAAI Press, Menlo Park, Calif., 1997, pp. 74–81.

[7] http://www.techgyd.com/10-best-data-mining-software-better-analysis/14362/

[8] Paško Konjevoda and Nikola Štambuk, "Open-Source Tools for Data Mining in Social Science ," *Theoretical and Methodological Approaches to Social Sciences and Knowledge Management*, pp.163-176

 @ 2013 http://www.ijitr.com All rights Reserved. 