International Journal of Soft Computing and Engineering (IJSCE)
ISSN: 2231-2307, Volume-2, Issue-3, July 2012

A Study of Data Mining Tools in Knowledge Discovery Process

Y. Ramamohan, K. Vasantharao, C. Kalyana Chakravarti, A.S.K.Ratnam

Abstract*— Data mining, the extraction of hidden predictive information from large databases, *is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. It uses machine learning, statistical and visualization techniques to discovery and present knowledge in a form which is easily comprehensible to humans. Various popular data mining tools are available today. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. 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. This paper presents an overview of the data mining tools like Weka, Tanagra, Rapid Miner, Orange.

***Index Terms*—Data mining, Rapid Miner, Tool, WEKA.**

I. INTRODUCTION

Data Mining [1][2], also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. The following figure (Figure 1) shows data mining as a step in an iterative knowledge discovery process.

In its simplest form, data mining automates the detection of relevant patterns in a database, using defined approaches and algorithms to look into current and historical data that can then be analyzed to predict future trends. Because data mining tools predict future trends and behaviors by reading through databases for hidden patterns, they allow organizations to make proactive, knowledge-driven decisions and answer questions that were previously too time-consuming to resolve.

Manuscript received on July, 2012.

Y.Ramamohan Persuing M.Tech(CSE) From Vignan's Lara Instiurte of Technology & Sceince. Vadlamudi, Guntur. AP.. India. My research interests are Datamining and Computer networks.

K.Vasantharao Persuing M.Tech(CSE) From Vignan's Lara Instiurte of Technology & Sceince. Vadlamudi, Guntur. AP.. India. My research interests are Datamining and Computer networks.

C.Kalyana Chakravarti Persuing M.Tech(CSE) From Vignan's Lara Instiurte of Technology & Sceince. Vadlamudi, Guntur. AP.. India. My research interests are Datamining and Computer networks.

A.S.K.Ratnam ,Assoc.Professor & Head.Department of Computer science engineering at Vignan's Lara Institute of Technology & Science. Vadlamudi. Guntur. AP., India.My research interests are Image Processing.

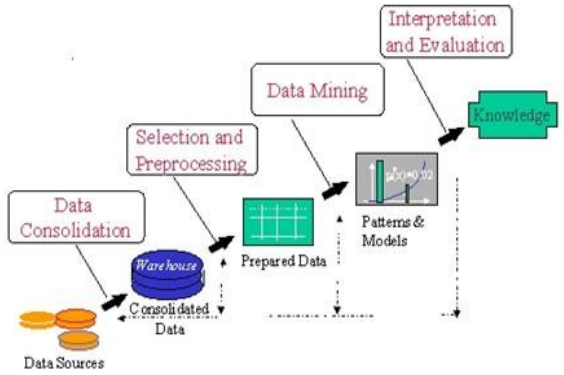

Figure 1: Data mining is the core of Knowledge discovery process.

Organizations that wish to use data mining tools can purchase mining programs designed for existing software and hardware platforms, which can be integrated into new products and systems as they are brought online, or they can build their own custom mining solution. For instance, feeding the output of a data mining exercise into another computer system, such as a neural network, is quite common and can give the mined data more value. This is because the data mining tool gathers the data, while the second program (e.g., the neural network) makes decisions based on the data collected.

Different types of data mining tools are available in the marketplace, each with their own strengths and weaknesses. Internal auditors need to be aware of the different kinds of data mining tools available and recommend the purchase of a tool that matches the organization's current detective needs. This paper presents an overview of the data mining tools available today. For example- weak, Tangara, RapidMiner, Orange.

II. CATEGORIES OF DATA MINING TOOLS

Most data mining tools can be classified into one of three categories: traditional data mining tools, dashboards, and text-mining tools. Below is a description of each.

A. Traditional Data Mining Tools

Traditional data mining programs help companies establish data patterns and trends by using a number of complex algorithms and techniques. Some of these tools are installed on the desktop to monitor the data and highlight trends and others capture information residing outside a database. The majority are available in both Windows and UNIX versions, although some specialize in one operating system only. In addition, while some may concentrate on one database type, most will be able to handle any data using online analytical processing or a similar technology.

*B. **Dashboards***

Installed in computers to monitor information in a database, dashboards reflect data changes and updates onscreen — often in the form of a chart or table — enabling the user to see how the business is performing. Historical data also can be referenced, enabling the user to see where things have changed (e.g., increase in sales from the same period last year). This functionality makes dashboards easy to use and particularly appealing to managers who wish to have an overview of the company's performance.

C. 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, thus providing users with an easy and convenient way of accessing data without the need to open different applications. Scanned content can be unstructured (i.e., information is scattered almost randomly across the document, including e-mails, Internet pages, audio and video data) or structured (i.e., the data's form and purpose is known, such as content found in a database). Capturing these inputs can provide organizations with a wealth of information that can be mined to discover trends, concepts, and attitudes.

When evaluating data mining strategies, companies may decide to acquire several tools for specific purposes, rather than purchasing one tool that meets all needs. Although acquiring several tools is not a mainstream approach, a company may choose to do so if, for example, it installs a dashboard to keep managers informed on business matters, a full data-mining suite to capture and build data for its marketing and sales arms, and an interrogation tool so auditors can identify fraud activity.

III. WEKA TOOL

WEKA[3], formally called Waikato Environment for Knowledge Learning, is a computer program that was developed at the University of Waikato in New Zealand for the purpose of identifying information from raw data gathered from agricultural domains. WEKA supports many different standard data mining tasks such as data preprocessing, classification, clustering, regression, visualization and feature selection. The basic premise of the application is to utilize a computer application that can be trained to perform machine learning capabilities and derive useful information in the form of trends and patterns.

WEKA is an open source application that is freely available under the GNU general public license agreement. Originally written in C the WEKA application has been completely rewritten in Java and is compatible with almost every computing platform. It is user friendly with a graphical interface that allows for quick set up and operation. WEKA operates on the predication that the user data is available as a flat file or relation, this means that each data object is described by a fixed number of attributes that usually are of a specific type, normal alpha-numeric or numeric values. The WEKA application allows novice users a tool to identify hidden information from database and file systems with simple to use options and visual interfaces. The following figure 2 presents the WEKA GUI chooser.

Figure 2: WEKA GUI chooser

IV. RAPIDMINER TOOL

RapidMiner[4], formerly YALE (Yet Another Learning Environment), is an environment for providing data mining and machine learning procedures including: data loading and transformation (ETL), data preprocessing and visualization, modelling, evaluation, and deployment. The data mining processes can be made up of arbitrarily nestable operators, described in XML files and created in RapidMiner's graphical user interface (GUI). RapidMiner is written in the Java programming language. It also integrates learning schemes and attribute evaluators of the Weka machine learning environment and statistical modelling schemes of the R-Project.

RapidMiner can be used for text mining, multimedia mining, feature engineering, data stream mining and tracking drifting concepts, development of ensemble methods, and distributed data mining. RapidMiner is found in the: electronics industry, energy industry, automobile industry, commerce, aviation, telecommunications, banking and insurance, production, IT industry, market research, pharmaceutical industry and other fields. The following figure 3 shows the GUI for RapidMiner.

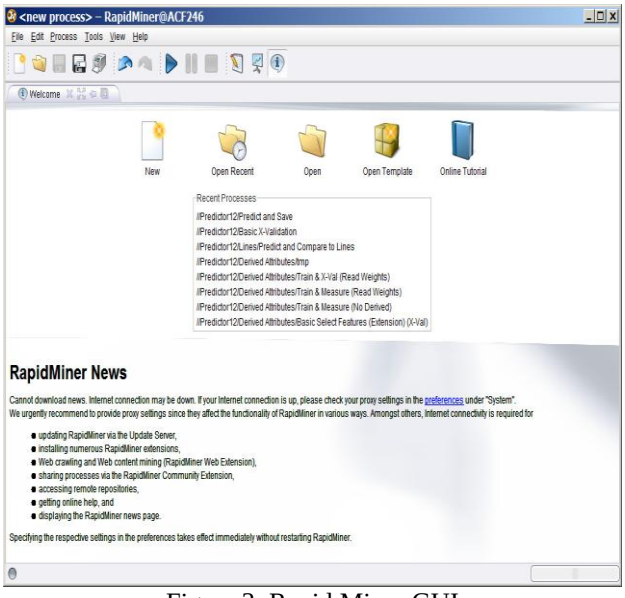

Figure 3: Rapid Miner GUI

V. TANAGRA TOOL

TANAGRA [5] is a free DATA MINING software for academic and research purposes. It proposes several data mining methods from exploratory data analysis, statistical learning, machine learning and databases area. This project is the successor of SIPINA which implements various supervised learning algorithms, especially an interactive and visual construction of decision trees. TANAGRA is more powerful, it contains some supervised learning but also other paradigms such as clustering, factorial analysis, parametric and nonparametric statistics, association rule, feature selection and construction algorithms. TANAGRA is an "open source project" as every researcher can access to the source code, and add his own algorithms, as far as he agrees and conforms to the software distribution license.

The main purpose of Tanagra project is to give researchers and students an easy-to-use data mining software, conforming to the present norms of the software development in this domain (especially in the design of its GUI and the way to use it), and allowing to analyze either real or synthetic data. The second purpose of TANAGRA is to propose to researchers an architecture allowing them to easily add their own data mining methods, to compare their performances. TANAGRA acts more as an experimental platform in order to let them go to the essential of their work, dispensing them to deal with the unpleasant part in the programmation of this kind of tools: the data management. The third and last purpose, in direction of novice developers, consists in diffusing a possible methodology for building this kind of software. They should take advantage of free access to source code, to look how this sort of software is built, the problems to avoid, the main steps of the project, and which tools and code libraries to use for. In this way, Tanagra can be considered as a pedagogical tool for learning programming techniques. The following figure 4 shows the GUI for Tanagra.

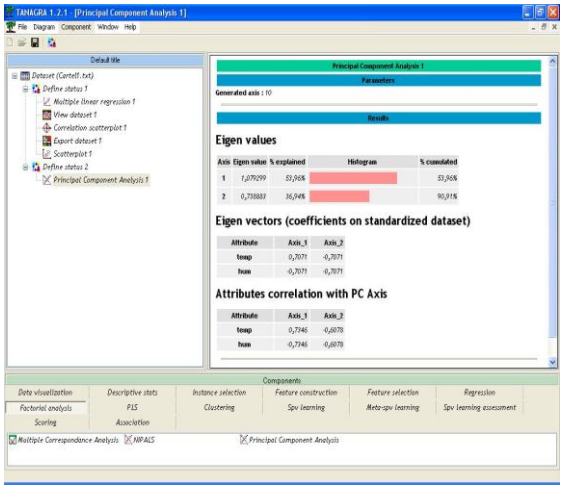

Figure 4: Tanagra GUI

VI. DBMINER TOOL

DBMiner[6], a data mining system for interactive mining of multiple-level knowledge in large relational databases, has been developed based on our years-of-research. The system implements a wide spectrum of data mining functions, including generalization, characterization, discrimination, association, classification, and prediction. By incorporation of several interesting data mining techniques, including attribute-oriented induction, progressive deepening for mining multiple-level rules, and meta-rule guided knowledge mining, the system provides a user-friendly, interactive data mining environment with good performance.

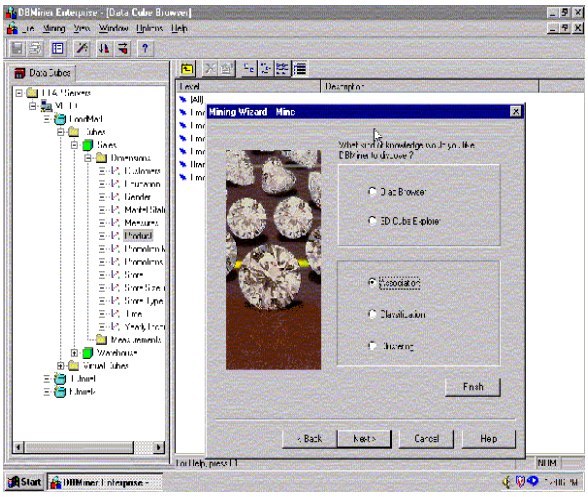

Figure 5: DBminer GUI

DBminer performs interactive data mining at multiple concept levels on any user-specified set of data in a database using an SQL-like Data Mining Query Language, DMQL, or a graphical user interface. Users may interactively set and adjust various thresholds, control a data mining process, perform roll-up or drill-down at multiple concept levels, and generate different forms of outputs, including generalized relations, generalized feature tables, multiple forms of generalized rules, visual presentation of rules, charts, curves, etc. The figure 5 shows GUI for DBminer tool.

VII. WITNESS MINER TOOL

WITNESS Miner[7] is a graphical data mining tool comprising a collection of data structures and algorithms written specifically for the tasks required in knowledge discovery. Designed to be easy to use, it provides a visual method of constructing streams, containing data preparation and data mining tasks that form the knowledge discovery process. The key features of this tool are: decision trees, clustering , discretization , rule induction using modern heuristic techniques, the ability to handle missing values, host of standard data processing tools, HTML output and in the case of the decision tree, XML output options, feature subset selection. Today's organizations collect a large amount of operational data relating to all kinds of activities. If properly analyzed, this data can have a significant effect on a company's performance and profitability. WITNESS Miner provides both a useful tool and the project framework for such investigations.

WITNESS Miner offers a way of making sense of data in Manufacturing, Finance, Health, Retail and Government. It provides knowledge from raw data through, data analysis, easy data modeling, powerful rule evaluation and high quality reporting. Most importantly, it allows an exploration of data to determine fundamental relationships that affect business. The WITNESS Miner module offers easy to understand rules generated directly from the data. Rules

determined are expressed in simple terms and enable simple decision rules to be implemented at many stages of key processes in order to affect service levels, costs and other major performance indicators. Figure 6 shows witness miner GUI

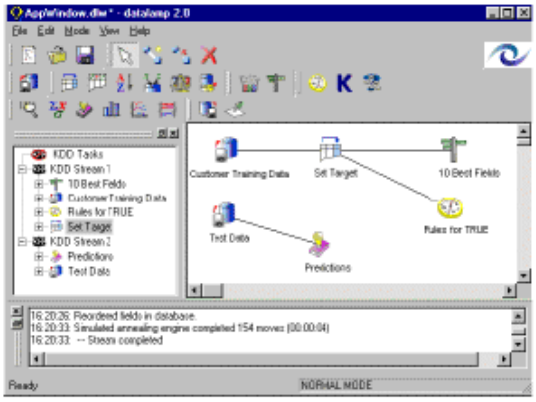

Figure 6 : Witness Miner GUI

VIII. ORANGE TOOL

Orange[8] is a powerful free and open source component-based data mining and machine learning software suite.It contains complete set of components for data preprocessing, feature scoring and filtering, modeling, model evaluation, and exploration techniques. It is based on C++ components, that are accessed either directly (not very common), through Python scripts (easier and better), or through GUI objects called Orange Widgets.Orange is distributed free under GPL and can be downloaded from the download page. Orange is a component-based framework, which means you can use existing components and build your own ones. You can even prototype your own components in Python, and use it in place of some standard C-based Orange component.Orange is supported on various versions of Linux,Apple's ,Mac OS X and Microsoft Windows.

The features of orange are: Preprocessing: feature subset selection, discretization, feature utility estimation for predictive tasks; Predictive modelling: classification trees, naive bayesian classifer, k-NN, majority classifier, support vector machines, logistic regression, rule-based classifiers (e.g., CN2); Ensemble methods, including boosting, bagging, and forest trees. Data description methods: various visulizations (in widgets), self-organizing maps, hierarchical clustering, k-means clustering, multi-dimensional scaling, and other; Figure 7 shows the GUI for orange.

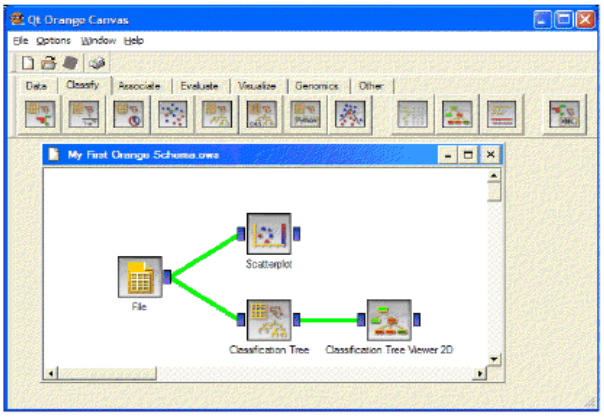

Figure 7 : Orange GUI

IX. CONCLUSION

Data mining is the extraction of useful patterns and relationships from data sources, such as databases, texts, the web…Using data mining to understand and extrapolate data and information can reduce the chances of fraud, improve audit reactions to potential business changes, and ensure that risks are managed in a more timely and proactive fashion. Auditors also can use data mining tools to model "what-if" situations and demonstrate real and probable effects to management, such as combining real-world and business information to show the effects of a security breach and the impact of losing a key customer.

REFERENCES

[1] J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2000.

[2] G. Piatetsky-Shapiro, U. M. Fayyad, and P. Smyth. From data mining to knowledge discovery: An overview. In U.M. Fayyad, et al. (eds.), Advances in Knowledge Discovery and Data Mining, 1-35. AAAI/MIT Press, 1996.

[3] The WEKA data mining sodtware: An update, Mark Hall, Eibe Frank, G. Holmes, B. Pfahringer, P. Reutemann, IH Witten, ACM SIGKDD Explorations, Newsletter, Pages 10-18, volume 11 issue 1, june 2009.

[4] http://rapid-i.com/

[5] http://eric.univ-lyon2.fr/~ricco/tanagra/

[6] DBMiner: A System for Data Mining in Relational Databases and Data Warehouses, Data Mining Research Group, Intelligent Database Systems Research Laboratory School of Computing Science, Simon Fraser University, British Columbia, Canada, http: db.cs.sfu.ca DBMiner.

[7] www.uea.ac.uk/polopoly_fs/1.3589!introductionkdd.pdf

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