A STUDY ON VARIOUS OPEN SOURCE TOOLS FOR DATA MINING

[1]Sathish G, [2]Nandhini Devi S
[1]Assitant Professor, [2]II MCA student
[1 2] Department of Computer Applications,
[1 2]Priyadarshini Engineering College, Vaniyambadi, Vellore, Tamilnadu, India.

Abstract: Today Information Technology plays a vital role in every aspects of the human life. It is very essential to gather data from different sources. This data can be stored and maintained to generate information and knowledge. This information and knowledge has to be disseminated to every stake holders for the effective decision making process. Due to the increase in the data, it is important to extract knowledge/information from the large data repositories. Hence, Data mining has become an essential factor in various fields including business, education, health care, finance, scientific etc. To analyses this vast amount of data and depict the fruitful conclusions and inferences, it needs specific data mining tools such as R, Weka, Orange. This paper discusses the knowledge discovery process, data mining, and various open source tools.

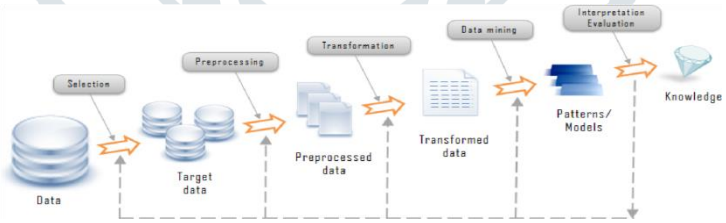
IndexTerms: Data Mining, KDD, Open Source Tool

INTRODUCTION

The development of Information technology has paved way to generate large amount of databases and huge data in various areas. The research in databases and information technology has given rise to approach to store and manipulate precious data for further decision making. Data mining is a process to extract the implicit information and knowledge by extracting from the mass, incomplete, noisy, fuzzy and random data with knowing the data well in advance and which is potentially useful to various fields. This paper is organized as follows open source tools for data mining like, Tanagra**,** Weka, Orange, KNIM and G Gobi

KNOWLEDGE DISCOVERY PROCESS

The process of discovering useful knowledge from a huge data is called as Knowledge Discovery in Database (KDD) and which is often referred to as Data mining. While data mining and knowledge discovery in databases are normally treated as synonyms, but, in fact data mining is a part of knowledge discovery process. The KDD process comprises of few steps as shown

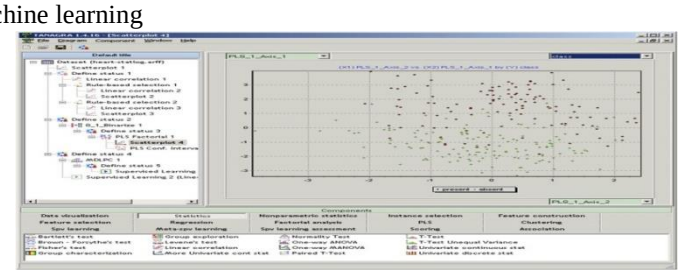


OPEN SOURCE TOOLS

TANAGRA

Tanagra is a data mining suite built around a graphical user interface where in data processing and analysis components are organized in a tree-like structure in which the parent component passes the data to its children .For example, to score a prediction modeling Tanagra the model is used to augment the data table with a column encoding the predictions, which is then passed to the component for evaluation.

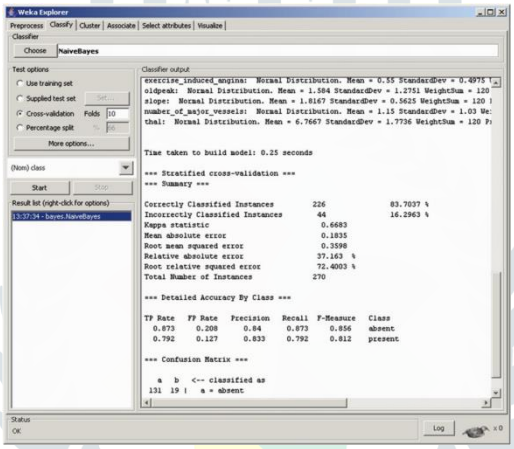
Although lacking more advanced visualizations, Tanagra is particularly strong in statistics, offering a wide range of un i - and multi varient parametric and nonparametric tests. Equally impressive is its list of feature selection techniques. Together with a compilation of standard machine learning



The component soft the data processing tree are dragged from the list the bottom (Components); the snapshot show sonly those related to statistics. The scatterplot on the right side shows these parathion of the instances based on the first woxes as found by the partial least squares analysis, where each symbol represents a patient, with the symbol's shape corresponding to a diagnosis. Techniques, it also includes correspondence analysis, principal component analysis, and the partial least squares methods. Presentation of machine learning models is most often not graphical, but instead machine learning suites include several statistical measures. The difference in approaches is best illustrated by the naive Bayesian classifier, whereby, unlike Weka and Orange, Tanagra reports the conditional probabilities and various statistical assessments of importance of the attributes (e.g., c^2, Cramer's V,and Tschuprow' set) Tanagra data analysis components report their results in a nicely formatted

WEKA

WEAK (Waikato Environment for Knowledge Analysis, is perhaps the best-known open-source machine learning and data mining environment. Advanced users can accessits components through Java programming or through a command-line interface. For others, Weka provides a graphical user interface in an application called the Weka Knowledge Flow Environment featuring visual programming and providing a less flexible interface that is perhaps seasier to use. Both environments include Weka impressive array of machine learning and data mining algorithms. They both offer some functionality for data and model visualization, although not as elaborate as in the other suites reviewed here. Compared with R, Weka is much weaker in classical statistics but stronger in machine learning techniques. Weka community has also developed a set of extensions covering diverse areas, such as text mining, visualization, bioinformatics, and grid computing. Like R in statistics,

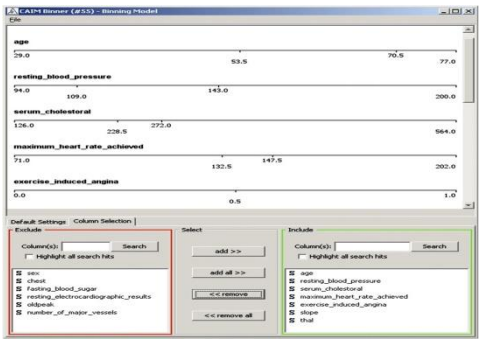


Weka Explorer with which we loaded the heart disease data set and induced a naive Bayesian classifier. On the right side of the window are the results of evaluation of the model using 10-fold cross-validation.

Weka became a reference package in the machine learning community, attracting a number of users and developers. Medical practitioners would get the easiest start by using Weka Explorer, and combining it with extensions for more advanced data and model visualizations.

.KNIME:

KNIME (Konstanz Information Miner, is a nicely designed data mining tool that runs inside the IBM's Eclipse development environment. The application is easy to try out because it requires no installation besides downloading and un archiving. Like YALE, KNIME is written in Java and can extend its library of built-in supervised and



`A dialog of the node ‘‘CAIM Banner ‘that transforms continuous features into discrete features (discretization).Featurestobediscretizedareselectedinthebottompart of the window, with the top part of the window displaying the corresponding split points.

Unsupervised data mining algorithms with those provided by Weka. But unlike that of Yale, KNIME’s visual programming is organized like a data flow. The user ‘‘programs’’ by dragging nodes from the node repository to the central part of the bench mark. Each node performs a certain function. such as reading the data, filtering ,modeling, visualization, or similar functions. Nod have input and output ports; most ports send and receive data, whereas some handle data models, such as classification trees. Unlike nodes in Weka Knowledge Flow, different types of ports are clearly marked, relieving the beginner of the guesswork of what connects where.

Typical nodes in KNIME’s Knowledge Flow have two dialog boxes one for configuring the algorithm or a visualization and the other for showing its results .Each node can be in one of the three states, depicted with a traffic-light display they can be disconnected, not properly configured, or lack the input data (red); be ready for execution (amber); or have finished the processing (green).A nice feature called HiLite (allows the user to select a set of

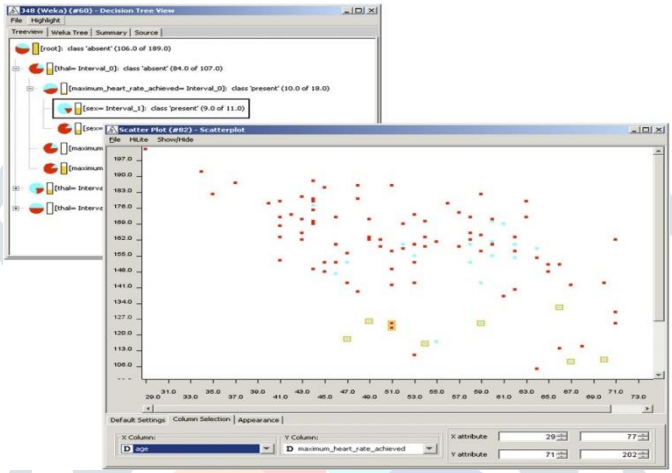
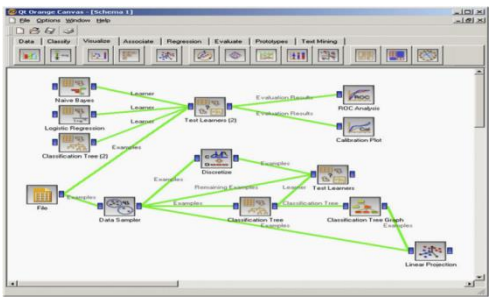


Fig.4: KNIME HiLiteing

Where the instances from the selected classification tree node are Halite and marked in the scat plot .instances in one node and have very other visualization in the current application, in this way further supporting a data analysis.

ORANGE

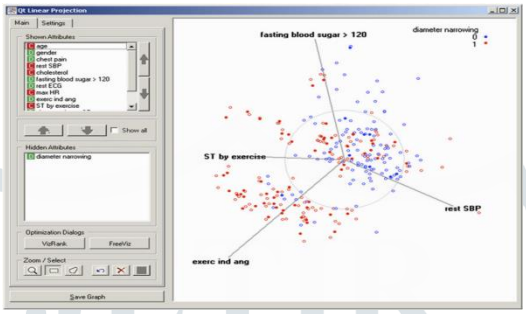
Orange Is a data mining suite built using the same principles as KNIME and Weka Knowledge Flow. In its graphical environment called Orange Canvas the use places widget son canvas and connects the min to a schema. Each widget performs some basic function. Snapshot of the Orange can vas. The upper part of the schema center around ‘‘Test Learners’’ uses cross-validation to compare the performance of three classifiers: naive Bayes, logistic, and a classification tree. Numerical scores are display ‘‘Test Learner the valuation results also passed onto ‘‘ROC Analysis and ‘‘Calibration Plot’’ that provide means to graphically analyses the predictive performance. The bottom part contains a setup similar to that in KNIME the data instances are split into training and test sets. Both parts are fed into ‘Test Learners,’’ which ,in this case ,requires as a classification tree built on the training set that is also visualized in ‘‘Classification Tree Graph.’’ ‘‘Linear Projection’’ visualizes the training instances, separately marking the subset selected in the ‘‘Classification Tree Graph’’ widget.



But unlike in KNIME with two data types models and set so instances the signals passed around Orange's schema may be of different types ,and may includeobjectssuchaslearners,classifiers,evaluationresults,distance matrices, dendrograms, and so forth. Orange's widgets are also coarser then KNIME's nodes so typically a smaller number of widgets is needed to accomplish the same task. The difference is most striking in setting up a cross validation experiment, which is much more complicated in KNIME, but with the benefit of giving the user more control in setting up the details of the experiment, such as separate preprocessing of training and testing example sets.

Besides friendliness and simplicity of use, Orange's strong points are a large number of different visualizations of data and models, including intelligent search for good visualizations, and support of exploratory data analysis through interaction. In a concept similar to KNIME's HiLiteing (yet subtly different from it), the user can select a subset of examples in a visualization, in a model, or with an explicit filter, and pass them to, for instance, a model inducer or another visualization widget that can show them as a marked subset of the data .

Orange is weak in classical statistics; although it can compute basic statistical properties of the data, it provides no widgets for statistical testing. Its



The linear projection widget from Orange displaying a two-dimensional projection of data, where the and y-axis are a liner combination of feature values whose components are delineated with feature vectors.ComingfromtheschemashowninFig.8,the points corresponding to instances selected in the classification tree are filled and those not in the selection are open.

Reporting capabilities are limited to exporting visual representations of data and models .Similar to R ,the computationally in pensive part so Orange are written in CPP where as the upper layers are in developed in the scripting language Python, allowing advanced users to supplement the existing suite with their own algorithms or with out in from Python .

G GOBI

Data visualization was always considered one of the key tools for successful data mining. Particularly suited for data mining and explorative data analysis, G Gobi is an open-source visualization program featuring interactive visualizations through, for instance, brushing (Fig.10),where by a user's selection is marked in all other opened visualizations, and grand tour which uses two-dimensional visualizations and in a movie-like fashion shifts between two different projections. G Gobi can also plot networks, a potentially useful feature for analysis of larger volumes of data, such as those from biomedicine. By itself G Gobi is only intended for visualization-based data mining ,but can be nicely integrated with other statistical and data mining approaches when used as a plug-in for R or used through interfaces for the scripting languages Perl and Python.

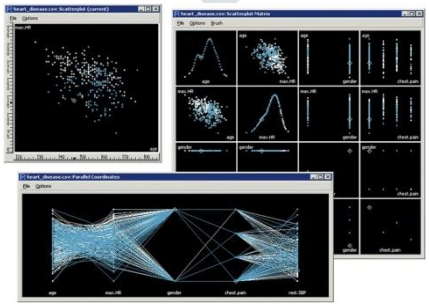
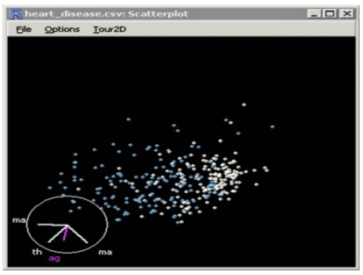


Fig.7. Scatterplot, a matrix of scatterplots and parallel coordinates as displayed by G Gobi. The instances selected in one visualization (scatterplot, in this case) are marked in the others.



GGobi'sGrandtourshowsaprojectionsimilartotheLinearProjectioninOrangebut animates it by smoothly switching between different interesting projections, which gives a good impression of positions of the instances in the multidimensional space.

CONCLUSION

Data mining will be considered one of the most important frontiers and one of the most promising interdisciplinary developments in Information technology. In this paper, we try to briefly review the knowledge Discovery Process, various open source tools and trends from its beginning to the future. This review would help the researchers to focus on the various issues of data mining. Data mining is useful for both public and private sectors for finding patterns, forecasting, discovering knowledge in different domains such as finance, marketing, banking, insurance, health care and retailing. Data mining is commonly used in these domains to increase the sales, to reduce the cost and enhance research to reduce costs, enhance research.

REFERENCES

[1] Fayyad, Piatetsky-ShapiroG, SmythP, etal,editors.Advances in knowledge discovery and data mining. Menlo Park (CA): AAAI Press; 1996.
[2] Quinlan JR. C4.5: programs for machine learning. San Mateo (CA): Morgan Kaufmann Publishers ;1993.
[3] MichalskiRS, KaufmanK.Learningpatternsinnoisydata:theAQapproach.In:Paliouras G, Karkaletsis V, Spyropoulos C, editors. Machine learning and its applications. Berlin: Springer-Verlag;2001.p.22–38.
[4] ClarkP, NiblettT the CN2inductionalgorithm.MachineLearning1989; 3:261–83.
[5] AsuncionA, NewmanDJ. UCIMachine Learning Repository. Availableat:http://www.ics. uci.edu/wmlearn/MLRepository.html.AccessedApril15,2007.Irvine,CA:Universityof California,DepartmentofInformationandComputerScience;2007.
[6] WallL,ChristiansenT,OrwantJ.ProgrammingPerl.3rdedition.Sebastopol, CA:O'Reilly Media,Inc.;2000.
[7] Khaki R,Sommer field Dougherty Data mining using:MLCPPamachine learning librarian's .InternationalJournalonArtificialIntelligenceTools1997;6:537–66.
[8] BrunkC, KellyJ,KohaviR.MineSet: anintegratedsystefor datamining.InProc.3rdIntl. Con .on Knowledge Discovery and Data Mining ,Menlo Park (CA).p.135–8.
[9] Witten IH, FrankE .Data mining: practical machine learning tools and techniques with Java implementations.2ndedition.SanFrancisco (CA): MorganKaufmann; 2005.
[10] ZupanB, Holmes JH, BellazziR.Knowledge-baseddataanalysisandinterpretation.Artif IntellMed2006; 37: 163–5.
[11] BellazziR, ZupanB .Predictive data mining in clinical medicine: current issues and guidelines.IntJMedInform2006; in press.