

A Survey on Software Suites for Data Mining, Analytics and Knowledge Discovery

R.Kiruba Kumari¹, V.Vidya², M.Pushpalatha³

¹ HoD, Department of Computer Science, Padmavani Arts and Science College for Women, Salem

² Asst.Prof. In Computer Science, Padmavani Arts and Science College for Women, Salem

³ Asst.Prof. In Computer Science, Padmavani Arts and Science College for Women, Salem

Abstract

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. Data mining software 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. This paper presents some of the most common commercially available data mining tools, with their most important features, side by side, and some considerations regarding the evaluation of data mining tools by companies that want to acquire such a system.

Keywords: data mining, data mining tools, features, software

I. INTRODUCTION

Today, a large number of standard data mining methods are available. The typical life cycle of new data mining methods begins with theoretical papers based on in house software prototypes, followed by public or on-demand software distribution of successful algorithms as research prototypes [1]. Then, either special commercial or open source packages containing a family of similar algorithms are developed or the algorithms are integrated into existing open source or commercial packages. This paper confers some of the commercial available software suits for data mining.

II. THE ADVANCEDMINER PROFESSIONAL SYSTEM

The Advanced Miner Professional System developed by Stat Consulting is modern and advanced analytical software. It provides a wide range of tools for data transformations, construction of Data Mining models, advanced data analysis and reporting. Advanced Miner Professional [2] is an integrated environment dedicated to the development of analytical projects. The system offers various tools supporting the work of analysts and programmers. The software not only includes data processing tools, but also provides a wide range of statistical algorithms and Data Mining methods, which can be used to construct effective analytical models. It provides tools for performing various tasks, such as classification, approximation, clustering, association rules, and survival analysis.

The features of Advanced Miner Professional includes

- extracting and saving data from/to different database systems and files,
- performing a wide range of operations on data, such as sampling, joining datasets, dividing into testing/training/validating sets, assigning roles to attributes,
- graphical and interactive data exploration,
- outlier filtering, supplying missing values, PCA, various data transformations, etc.,
- building association models, clustering analyses, variable importance analyses, etc.,
- constructing various analytical models with the use of diverse Data Mining and statistical algorithms (such as classification trees, neuron networks, linear and logistic regression, K-means, association rules),
- creation of scoring code so that the models can be integrated with other IT applications (scoring code may include the models as well as data transformations),
- model quality evaluation and comparison of Data Mining models (LIFT, ROK, K-S, Confusion Matrix), generation of model quality reports (MS Office, OpenOffice).

AdvancedMiner Professional is based on well-tested Java technologies, providing platform independence. The system operates in MS Windows as well as in operating systems from the Unix family (including Linux). AdvancedMiner Professional is compatible with relational database management systems which provide the JDBC/ODBC interface (e.g. MySQL, MS SQL, Oracle, Sybase SQL Anywhere Studio).

The below figure 1 shows the working environment of AdvancedMiner Professional System

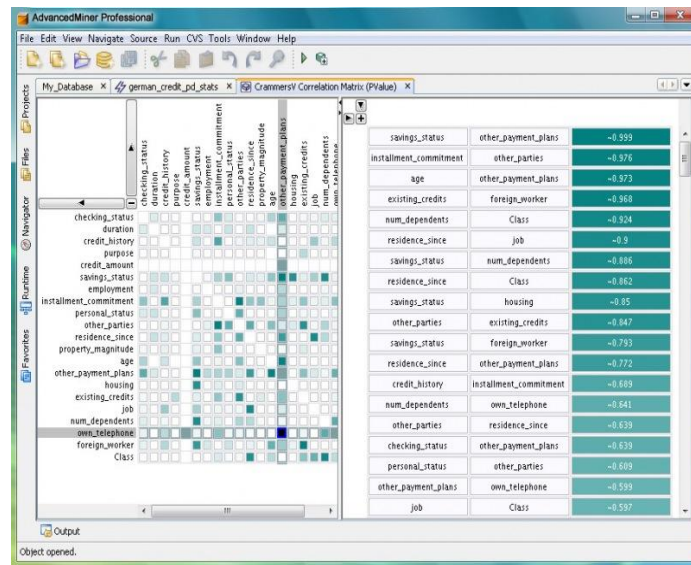


Fig 1. Working environment of AdvancedMiner Professional System

An internal database – GDBase - is also available for the users. AdvancedMiner Professional can import data from practically any relational database supporting the ODBC/JDBC standard (including MS SQL, MySQL, Oracle, Sybase), data warehouses, CSV files, spreadsheets. The user may not only explore but also edit databases, as well as create new tables using the built-in scripting language. It is possible to use SQL queries and special language constructions for effective data transformations.

III. BAYESIALAB

BayesiaLab is a bayesian network publishing and automatic learning program which represents expert knowledge and allows one to find it among a mass of data. [3] The handling of graphs is very intuitive, making it a wonderful communication aid. There's no need to be an expert to quickly grasp the represented knowledge and put it to good use for decision making. BayesiaLab enables to automatically find unknown relations, full of information (data mining) in the data. In this way, it is possible to transform tables that are difficult to decipher into a very meaningful graph.

BayesiaLab[4] offers an array of analysis tools, which can provide the analyst with a deeper understanding of the domain's underlying dynamics. The Bayesian network also provides the basis for a wide range of scenario simulation and optimization algorithms implemented in BayesiaLab. Beyond mere one-time predictions, BayesiaLab allows dealing with evidence interactively and incrementally, which makes it a highly adaptive tool for real-time inference. BayesiaLab is a powerful desktop application (Windows/Mac/Unix) with a highly-sophisticated graphical user interface, which provides scientists a comprehensive “lab” environment for machine learning, knowledge modeling, diagnosis, analysis, simulation, and optimization.

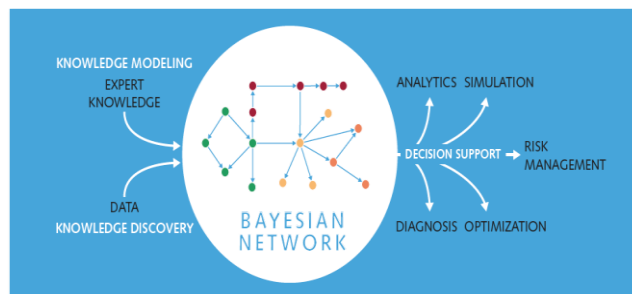


Fig 2: BayesiaLab Workflow with Bayesian Networks at its core.

With BayesiaLab, Bayesian networks have become a powerful and practical tool to gain deep understanding of high-dimensional domains. It leverages the inherently graphical structure of Bayesian networks for exploring and explaining complex problems. Java based program, BayesiaLab is works under Windows, OS X, Linux/Unix. It is available in English, Japanese, Chinese, Spanish and French. BayesiaLab is the first commercial tool that comes with Causal inference (Pearl's Do-Calculus).

The Causal Intervention is compatible with all the analysis tools that are available in BayesiaLab, and with all the kinds of evidence (hard, soft, probability distribution, mean value). The type of inference, observation or intervention, can be set for each variable. BayesiaLab is designed around the Bayesian network paradigm, as is illustrated in Figure 2. It covers the entire research workflow from model generation to analysis, simulation, and optimization. The entire research workflow is fully contained in a single "lab" environment, which provides analysts the ultimate flexibility in moving back and forth between different elements of the research task.

In addition to the adaptation of BayesiaLab to Java 7, here is a small selection of features:

- An entirely redesigned Target Optimization Tool, which uses a Genetic Algorithm for comprehensive optimizations.
- Automatic Computation of Contributions for each network generated through Multi-Quadrant Analysis.
- Disjunctive Inference and Negation of the Evidence Set for scenario analysis.
- Workspace to start BayesiaLab with a set of previously opened networks.
- A "Token Borrowing" functions for floating licenses, which allows users to work offline, e.g. while traveling.

The figure 3 depicts the interface of bayesia lab for clustering.

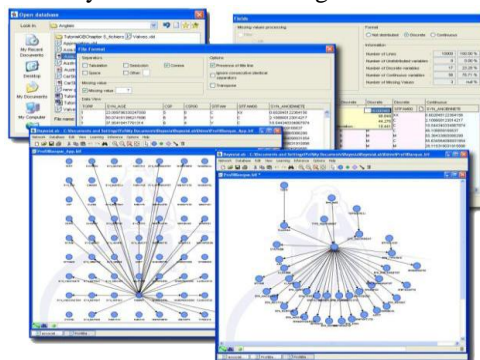


Fig 3: Interface of bayesia lab for implementing clustering

IV. SAS ENTERPRISE MINER

SAS Enterprise Miner [5] streamlines the data mining process to create highly accurate predictive and descriptive models based on large volumes of data from across the enterprise. It offers a rich, easy-to-use set of integrated capabilities for creating and sharing insights that can be used to drive better decisions.

Features

- Powerful, easy-to-use GUI, as well as batch processing for large jobs
- Interactive GUI for building process flow diagrams.
- Batch processing code for scheduling large modeling and scoring jobs.
- Data preparation, summarization and exploration
- Access and integrate structured and unstructured data sources.
- Outlier filtering.
- Data sampling.
- Data partitioning.
- File import.
- Merge and append tools.
- Univariate statistics and plots.
- Bivariate statistics and plots.
- Batch and interactive plots.
- Segment profile plots.
- Easy-to-use Graphics Explorer wizard and Graph Explore node.
- Interactively linked plots and tables.
- Data transformations.
- Time series data preparation and analysis.
- Interactive variable binning.
- Rules Builder node for creating ad hoc data-driven rules and policies.
- Data replacement.
- Enterprise Miner 12.3 includes new pre-built Enterprise Miner data mining process flow diagram templates. The data mining process flow diagram templates serve as examples of analytic data mining modeling approaches for some of the more common specific business problems.

- The Enterprise Miner 12.3 client can now open to directly load a specific data mining project or diagram. The Project Navigator tree now displays the most recently opened projects and process flow diagrams at the top.
- The Enterprise Miner 12.3 Link Analysis node converts relational and transactional data into a data model that can be visualized as a network of effects. The node can detect the linkages between any two variables' levels for relational data and between two items' co-occurrence in transactional data. Multiple centrality measures and community information are provided for users to better understand linkage graphs. The Link Analysis node generates cluster scores from raw data that can be used for data reduction and segmentation. The Link Analysis node also uses weighted confidence statistics to provide "next-best-offer" information to customers.
- The Enterprise Miner 12.3 [6] Reporter node improves image and table displays for upgraded PDF and RTF output. The Reporter node report files are smaller in size, but contain improved graphic displays and provide new graphic output font scaling options.
- The Enterprise Miner 12.3 Impute node now supports imputation of special missing values.
- The Enterprise Miner 12.3 Survival node now supports time-varying covariates, as well as user-specified censor and truncation dates.

The following figure 4 shows the layout of SAS Enterprise Miner

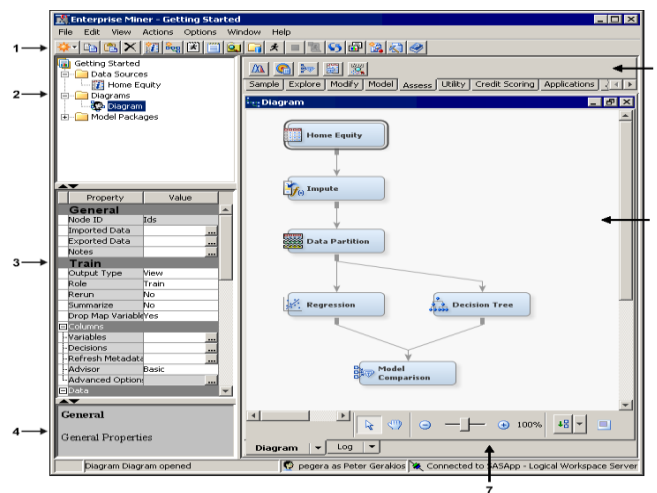


Fig 4: Layout of SAS Enterprise Miner

System Requirements:

Host Platforms/Server Tier

HP/UX on Itanium: 11iv3 (11.31)

IBM AIX R64 on POWER architecture 7.1

IBM z/OS: V1R11 and higher

Linux x64 (64-bit): Novell SuSE 11 SP1; Red Hat Enterprise Linux 6.1;

Oracle Linux 6.1

Server: Windows Server 2008 x64 SP2 Family; Windows Server 2008 R2 SP1 Family;

Windows Server 2012 Family

Solaris on SPARC: Version 10 Update 9

Solaris on x64 (x64-86): Version 10 Update 9; Version 11

Client Tier

Microsoft Windows (64-bit): Windows 7* x64 SP1; Windows 8** x64

Required software

Base SAS®

SAS/STAT®

SAS Rapid Predictive Modeler requires SAS Enterprise Miner to produce predictive models. The SAS Rapid Predictive Modeler task is available from either SAS Enterprise Guide or SAS Add-In for Microsoft Office (Microsoft Excel only).

With SAS Rapid Predictive Modeler, business analysts and subject-matter experts can rapidly explore and analyze their data using either the familiar, visual interfaces available in Microsoft Excel or the guided analysis capabilities of SAS Enterprise Guide. In addition, data mining specialists and statisticians can generate quick, baseline models when they are short on time and resources.

V. ESTARD DATA MINER

ESTARD Data Miner (EDM) [7] is a data mining tool, able to discover most unexpected hidden information in the data. Most databases contain data that is accumulated for many years. These databases (also called data warehouses) can become a valuable source of new knowledge for analysis. The newest business intelligence techniques were incorporated into ESTARD Data Miner for carrying out automated data analysis. User-friendly interface and wizards allow to start working with the tool in a few clicks.

In comparison to common business intelligence tools, ESTARD Data Miner is able to provide with something more than just operating statistics - it gives power to work with predictive analysis.

Predictive analysis is a business intelligence method for creating decision models. ESTARD Data Miner creates predictive models in "if-then" form. Such models can be implemented in any field of business or science. For example, EDM can create models describing customers with high risk of bad debt. These models can be used for detecting what a company can expect from a new client. The unique feature of ESTARD Data Miner is the analysis flexibility. ESTARD Data Miner is based on unique data mining algorithms.

Some of tool functions are: importing data from various databases, statistical analysis, decision trees creation and revealing all if-then rules describing hidden correlations in data. The data mining tool allows to create reports on discovered knowledge and to export discovered data patterns to various files.

It is possible to create decision rules, revealing all the if-then rules in the data and also build decision trees.

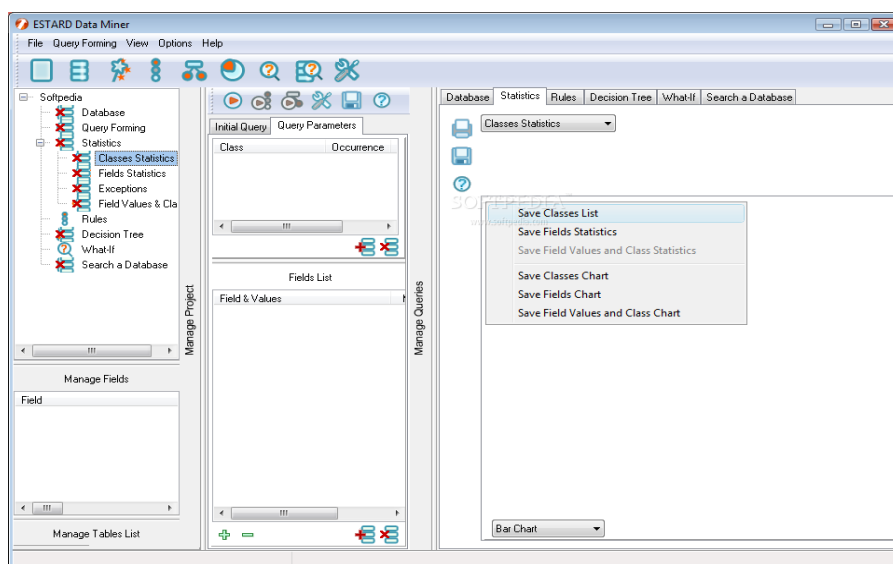


Fig 5: Layout of ESTARD Data Miner

The obtained decision rules and decision trees are represented in a user-friendly, intuitive form.

Statistical module contains charts and reports that are easy to understand, print and save.

Reports on decision rules, decision trees and statistical analysis are provided.

Rules can be edited or deleted in case if users want to combine their own knowledge with discovered one.

Wizards for data mining and data base loading will ease the process of data mining.

Different analysis settings for expert data mining customization are available.

Save the discovered rules, trees and statistics for further exploration and usage.

Perform WHAT-IF analysis within few clicks

Discover data patterns in databases

Use previously saved, uploaded or just obtained rules and decision trees to analyze databases, discovering classes within them.

REQUIREMENTS

Operating system: Microsoft Windows 98, Microsoft Windows NT, Microsoft Windows 2000, Microsoft Windows XP Professional- or Home Edition, Windows Vista, Windows 7

Processor: 1 GHz or better

Memory: 256 MB RAM (512 MB recommended)

Disk space: The installation footprint is approximately 9 MB

Other: Data handling infrastructure

VI. CONCLUSION

The different methods of data mining are used to extract the patterns and thus the knowledge from this variety databases. Selection of data and methods for data mining is an important task in this process and needs the knowledge of the domain. Several attempts have been made to design and develop the generic data mining system but no system found completely generic. Thus, for every domain the domain expert's assistant is mandatory. The domain experts shall be guided by the system to effectively apply their knowledge for the use of data mining systems to generate required knowledge. The domain experts are required to determine the variety of data that should be collected in the specific problem domain, selection of specific data for data mining, cleaning and transformation of data, extracting patterns for knowledge generation and finally interpretation of the patterns and knowledge generation. Thus this paper has focused a variety of techniques, approaches and different areas of the research which are helpful and marked as the important field of data mining technologies.

REFERENCES

- [1] Ralf Mikut and Markus Reischl "Data mining tools" John Wiley & Sons, Inc. Volume 00, January/February 2011
- [2] http://www.statconsulting.eu/assets/files/AdvancedMinerProfessional_EN.pdf
- [3] BayesiaLab Professional Edition 5.0 www.bayesia.com
- [4] Stefan Conrady, Dr. Lionel Jouffe "Introduction to Bayesian Networks & BayesiaLab", September 3, 2013, www.bayesia.sg
- [5] <http://www.sas.com/technologies/analytics/datamining/miner/>
- [6] <http://www.sas.com/resources/factsheet/sas-enterprise-miner-factsheet.pdf>
- [7] <http://www.estard.com/>