CMPUT695: Assignment#2

Evaluation Report on PolyAnalyst 4.6

Hongqin Fan and Yunping Wang

1. INTRODUCTION

PolyAnalyst 4.6 professional edition (PA) is a commercial data mining tool developed by Megaputer Intelligence Inc. With over 300 users world-wide including several Fortune 100 companies, PA is one of the well-recognized and comprehensive software designed for business data analysis.

The intended users of PA are the domain experts who are not computer specialists but have basic understanding of statistics and data mining technology. As a result, this product comes with an intuitive and easy-to-learn user interface, and various visualization features. It can produce easy-to-understand report in both textual and graphical format,import data from various data structures such as spreadsheets, relational databases, and text files.

With the detailed documentation, hands-on tutorials and relevant data samples included in the product, PA can be learned and used with minimum efforts of training. The menu-driven feature together with toolbars, object browser and logs get people started towards a solution within a fairly short amount of time. The product offers a comprehensive set of data analysis engines for data association analysis, clustering, classification, predication and text mining. Some concepts and algorithms such as Symbolic Rule Language for expression of mining rules, Localization of Anomalies (LA) for data clustering are direct research results from Megaputer research team; therefore they are specific to PA product.

This report will evaluate the product on its functionalities, the positive as well as the negative aspects of the tool, its user interface, flexibility of the API, etc.

2. TOOL EVALUATION

2.1 System Requirements and Platforms Supported

PolyAnalyst works with Microsoft Windows 98/2000 and Windows NT but unfortunately does not offer either Unix or Linux support. Windows 98 will work with all algorithms except *Find Laws*. This is reasonable if we think of the fact that Windows is the most popular operating system for individual and corporate users.

2.2 Intended End Users

The intended end users are business people who want to utilize the power of data mining to make informed decisions, they understand the data mining rationale but they are not computer technicians. Megaputer provides a COM version of PolyAnalyst for system developers, so that the exploration engines can be seamlessly integrated with their current

decision support system, or simply incorporated into the Microsoft Excel for data analysis.

2.3 Major Functionality and Algorithms

PA design focuses on domain experts, which partially determines the features of the software as demonstrated through our trial usage: the system is well designed for ease-of–use and high compatibility in data source connection, communicating with other software. Problems may arise for advanced users or in some special data analysis scenarios when the control of manipulation on some algorithms is desired.

There are 19 exploration engines (11 algorithms) included in PA, as listed below:

Table 1: List of Exploration Engines included in PA system

DM Tasks	Number of Exploration Engines	Exploration Engines
General	3	• Discriminate • Find dependencies • Summary statistics
Association	2	• Link analysis • Market basket analysis
Clustering	1	• Cluster
Classification	3	• Classify • Decision forest • Decision tree
Predicative analysis	4	• Find laws • Linear regression • Memory based reasoning • PolyNet Predictor
Text Analysis	6	• Link terms • Taxonomy Categorizer • Text Analysis • Text Categorization • Text De-repeater • Text OLAP

Because we are especially interested in association analysis, clustering, classification and predication, our review is mainly limited to these algorithms under each category in the following paragraphs.

2.3.1 Association Analysis

PA system offers two ways to do association analysis: Market Basket Analysis and Transaction Basket Analysis. The algorithms behind these two methods are almost identical; the only difference between them is the format of input data .

Association analysis can be used to successfully retrieve frequent item set and group of association rules with given support and confidence. The speed of analysis is quite fast, it is within 1 second to retrieve result for 1000 transactions with 500 distinct items. (We only have evaluation version on hand, the maximum testing case should be less than 1000 transactions). According to PA manual, the testing cases should be 500 to 3,000,000 transactions, which is good enough to analyze large databases.

However, the input data for Association analysis (both Market Basket Analysis and Transaction Basket Analysis) should be binary data (0 or 1 to represent buy or not buy). It is very memory-consuming and it is very inconvenient for user to analyze data in real world. Most data in real world do not follow this kind of input format; the user needs to take some time to pre-process raw data.

2.3.2 Clustering

The algorithm for clustering in PA is Localization of Anomalies (LA), which is proposed by Megaputer research team and summarized in the paper " LA - a Clustering Algorithm with an Automated Selection of Attributes" [1]. LA is "an algorithm based on the comparison of the n-dimensional density of the data points in various regions of the space of attributes with an expected homogeneous density obtained as a simple product of the corresponding one-dimensional densities". The LA algorithm consists of two logical components. The first component selects the best combination of attributes which provides the most significant and contrast clustering. The second component finds clusters in the space of a fixed set of attributes. The algorithm searches for combinations of different attributes and divides the dataset space into different regions using hyper planes. The algorithms have the following advantages over other clustering algorithms as claimed by the authors:

- Using functional derivatives of the attributes in replacement of attributes will not change the clustering results
- Computational time only depends on the number of data records very weakly
- LA works best in the case of a great number of records
- LA algorithm is noise tolerant.

LA algorithm can be used for the detection of outliers, unidirectional data clustering or preprocessing for data mining. Attributes of records can be of any data type. There is no need to input parameters as is similar to TURN* [2]. The algorithm will also identify the most influential attributes for clustering results.

The limitation of this algorithm is that the number of records should be large. Usually more than 3 to the power of the number of attributes are required by the algorithm, otherwise the result will be doubtful.

2.3.3 Classification

Data classification is used to predict categorical labels. First, a model is built and trained using known dataset and tested, and then this model can be used for labeling the new data records. There are three exploration engines for data classification: Classify, Decision Tree and Decision Forest.

Classify Engine in PA is used to split a dataset into two groups, or predict a binary discrete attribute value of a record. Several features of this Classify engine (1) Use fuzzy logic to classify, the label is based on the probability of one record falling into a class, i.e. whether the probably exceeds one threshold value set by the user (2) the Classify is in fact implemented through one of the three algorithms: Find laws, PolyNet Predicator (PA neural network solution), or Linear Regression.

Decision Tree (DT) is used to classify cases into multiple categories. The target attributes should be categorical or Boolean values. Underlying algorithms of DT are (1) Information Gain splitting criteria, and (2) Shannon information theory and statistical significance tests. In many cases this is the fastest and most easy-to-interpret algorithm in PA system.

Decision Forest (DF) provides an efficient technique for solving the task of categorizing data records into multiple classes. Its difference from DT is that multiple trees are built for different categories, each with a yes/no target attribute. DF is more accurate and efficient when classifying cases into multiple categories.

2.3.4 Predication

Data prediction is used to predict a continuous-valued attribute if knowing other attributes. There are four exploration engines for data predication: Find laws, Linear Regression, Memory Based Reasoning and PolyNet Predicator.

Find Laws is to find non-linear dependencies between attributes; the target attribute must of numerical type. *Find Laws* is based on Symbolic Knowledge Acquisition Technology (SKAT), an algorithm developed by Megaputer.

Linear Regression (LR) finds the linear relationship between the target numerical attribute and other attributes, based on stepwise linear regression.

Memory Based Reasoning (MBR) assigns values to data points based on their “proximity” to other data points. It is the only algorithm in PA system to predict any data types, either a numerical value, categorical value or Boolean. The user can choose to keep the current attributes or normalize the attributes so that they have different or comparable weights. MBR uses Genetic Algorithm (GA) to find the best size of neighbors and the best method of calculating the distance (weighted or not weighted).

PolyNet Predicator (PN) is a neural network algorithm built in PA system. PN can be used for prediction of a continuous value when there are large number of records and few attributes.

2.3.5 Text Mining

PA offers a convenient way to do text analysis; it can extract key concepts from text saved in a database, categorize individual database records, and retain useful information from natural language documents. The input text can be in the form of memorandums, contracts, regulations, manuals, e-mail messages, or natural language fields in a database

The basic idea of text analyses and test categorization is to deliver a categorization tree, the categorization tree allows for the creation of rules, based on the nodes in the tree, to be applied to datasets allowing textual categorization.

2.3.6 Others

Megaputer also provides a separate product called "WebAnalyst" for web mining, WebAnalyst combines data mining and text mining within websites so that some intelligent features can be integrated into a website, such as product recommendations, web page navigation, customer care, etc.

2.3.7 Accuracy

All the algorithms, if applied according to their intended purposes, should give satisfactory results. However, to achieve good results, there are some requirements for data sources. For example, the minimum number of data records, varying with algorithms, should be at least 100~1,000. Optimum number of records also depends on different algorithms. Even though some algorithms are noise tolerant, mining on dataset that contains noises above an acceptable level will generate a doubtful model and give erroneous result.

2.4 Client-Server Process

Data mining applications often use very large data sets which need to be stored in physical RAM. Algorithms always run far slower when hundreds of candidate inputs are considered in models. Client-server-processing model provides great advantage; it can use a single high-powered workstation for processing, but let multiple analysts access the tools from PCs on their desks

PolyAnalyst Knowledge Server is a separate server product which can be used with PolyAnalyst Professional 4.6 together. PolyAnalyst Professional 4.6 has a special design for distributed Client/Server architecture, and serves as client-side in the case of client-server process. To achieve a good performance, most computationally intensive algorithms of PolyAnalyst are multithreaded.

2.5 Automation and Project Documentation

Data mining process is always iterative; it is a good idea to automate analyses data so that analyst can be freed from some mundane and error-prone tasks of linking and documenting research findings.

PA does not provide iterative data mining automatically, but developers can extent the application to accomplish some automation functionality. For example, the “Scheduler” in PA allows user to schedule automated DM processes using scripting languages. There is related documentation about automation design in PA and its COM version.

There are also some wizards in PA system to guide the user through data pre-processing, or data visualization steps.

2.6 Ease of Use

2.6.1 Data Sources

The following data sources are supported:

- Comma Separated Value (SCV) file
- External data through ODBC
- External data through OLE DB
- Microsoft Excel
- Folder with text documents
- SAS data file

PA system is fully compatible with Microsoft Data communication standards: Object Linking and Embedding in Database (OLE DB) and Open Database Connectivity (ODBC), therefore most data sources via their OLE DB or ODBC drivers can be connected for data analysis, for examples, Microsoft SQL server, Access, Oracle, DB2 etc. In addition, data can also be read directly from Comma Separated Value (CSV) file, excel text files and SAS data files.

In one word, PA system can work with any structured data including Boolean data, time series data and categorical data while other data mining tools are limited to work with numerical data. However, the limitation of PA is the format of data sources, i.e. how data records and attributes are presented, and some algorithms have specific requirements for data layout. For examples, the data source for Market Basket Analysis must be in a format where all the transactional items listed in the columns as a Boolean value and records in data rows. This format is regarded as extremely inefficient in the case of great number of items, and no mechanism is provided in the system for such transformation.

2.6.2 Data Preprocessing

PA system provides a number of methods for data pre-processing including (1) handle with missing attributes; the dataset can be purified by removing the data tuples with

number of missing attributes exceeding a percentile value. (2)split, the dataset can be split to equal intervals or equal parts based on selected attribute(3) random sampling, new sub-dataset can be generated by random sampling, this technique is used for creating test data on mining models (4) create rules, new attributes can be generated from the existing ones based on user-defined dependent rules. (5) sub-datasets can be used to generate new dataset through operations such as union, intersection, supplement etc.

2.6.3 Results Presentation and Visualization

A DM project in PA includes attributes, datasets, graphs and reports. Graphs and reports are major deliverables of DM results. Most graphs are statistical charts, include histogram, 2D and 3D charts, snake charts, link charts etc, while reports are mixed format of text, tables, and visualization of DM models.

Some well-designed visualization features in the system include visualization of decision trees for classification, gain chart, lift chart for marketing etc. However we noticed most visualization functions deal with statistical results, visualization of DM models is regarded as a limitation in this product if we compare with other DM tools.

2.6.4 GUI Design

The GUI design is reasonable, it is easy for use to explore the tool, comparing with other tools such as weak, the GUI of PA is quite user friendly, it is easy for new user to practice on it.

2.6.5 Examples and Tutorials

There are eleven sample projects in PA system, designed to demonstrate the use of major exploration engines and data mining procedures. In many cases, one problem is tackled with different algorithms for the purpose of comparison. These sample projects coupled with tutorials provide a mechanism for self-paced learning of PA system, from data source connection, dataset creation, model selection and validation, to visualization and reporting.

2.6.6 System Compatibility

Even though the system is a stand-alone application, it is highly compatible with many other systems. For example, some exploration engines in PA can be incorporated into Microsoft Excel spreadsheet for data analysis, the data mining models in the system can be exported to PMML format, and used by other DM tools.

2.6.7 Manual Information

PA comes with provides a completed manual; it is easy for new user to learn, it is good for developer to extend the application.

2.7 Flexibility of API

To provide the flexibility of Application Programming Interface (API), Megaputer released a COM version of PolyAnalyst, with target user group of developers. The client chooses only the needed algorithms from the COM version and pays for what he needs. These modules can be easily integrated into any existing information system with programming languages like VBA, or C++ . So the users can still use the current familiar interface while exploiting the power of data mining techniques.

3. COMPARISON WITH OTHER DATA MINING TOOLS

PA system is compared with another data mining tool, Weka 3 developed by the University of Waikato from perspectives of platform supported, algorithms included, ease-of-use, visualization, etc.

Our comparison results are summarized in Table 2.

Table 2: Comparison between PA and WEKA

Items		**PolyAnalyst 4.6**	**WEKA 3.0**
Platform supported		Windows only	Windows, Mac OS, Linux
Algorithms included	Association	Not Known	Apriori
	Clustering	LA	K-means, Density-based, EM etc.
	Classification	Classify, Decision Tree, Decision Forest	Bayesian classifier, decision tree, etc, up to 7 categories of approaches, including dozens of algorithms
	Predication	Find laws, Linear regression, Memory based reasoning, PolyNet Predictor	Included in the classifiers, see above
	Others	Text mining, web mining	Not available
Data input and Model output options		Various data sources, some restrictions to data format. Model output expressed in various formats.	Limited to arff, csv format etc. good data filtering functions, not much control on model output.
Usability ratings		User-friendly, good system compatibility, easy to learn and use.	Some user-friendly features, lack of help, documentation and tutorials, user interface not rich, good user control on algorithm usage.
Visualization capabilities		Visualization focuses on	Some visualization

	statistical results and accuracy of mining results. Good visualization for decision tree, some degree of visualization for association rules and clustering	capabilities for classifier errors, cost curves, tree structures etc. Rated as average, a little bit poor than PA.
Modeling automation methods	To some degree	No

4. IDEAL FEATURES FOR DATA MINING TOOL

A good data mining tool should, ideally, have the following features and characteristics:

- *Wide platform support*: compatible with different operating systems.
- *Client-server process*: client-server implementation help improve the efficiency of carrying out some computation-intensive data mining tasks.
- *Intuitive and easy to use GUI:* most users are domain experts, well-designed GUI facilitates their learning and usage of the system
- *Sufficient number of algorithms*: no algorithm is a single winner for a data mining task, there should be various algorithms available for different scenarios.
- *Well-designed report and visualization*: help people understand the results, evaluate the results and communicate the results.
- *Extensibility*: can be extended or being used as plug-ins for current enterprise DSS.
- *Compatibility:* support to various data sources, provision of data transformation functions if special data layout or format is needed, capability of exporting results to other systems**.**
- *Complete User Manual and Documentation***:** help user to become familiar with the tool within a short time.

5.CONCLUSION

PolyAnalyst Professional 4.6 is generally evaluated as an appropriate data mining tool for our application scenario because of its ease-of-use, complete suite of algorithms, well-designed reports and extensibility. PA can help us fulfill our routine tasks of association analysis, clustering, classification etc.

Besides its unfortunate Windows-centricity, the other shortcomings we find with PA are its stringent requirements for data format, limited capability of visualization, this can be

inconvenient for some users. But generally, PolyAnalyst's flexibility, ease of use, complete suit of data mining algorithms, and affordable price make it worth to buy. If you want to perform some data mining on your own customers' data, we strongly suggest you using PA data mining system.

References

[1] KISELEV, M.V., ANANYAN, S.M. & ARSENIEV, S.B. "LA - a Clustering Algorithm with an Automated Selection of Attributes", http:// www.megaputer.com/tech/wp/cluster.pdf , (1998)

[2] A. Foss and O. Zaiane, "A Parameterless Method for Efficiently Discovering Clusters of arbitrary Shape in Large Datasets", 2002 IEEE International Conference on Data Mining (ICDM'02) December 09 - 12, Maebashi City, Japan , 2002