

Decision Tree Induction & Clustering Techniques In SAS Enterprise Miner, SPSS Clementine, And IBM Intelligent Miner – A Comparative Analysis

Abdullah M. Al Ghoson, Virginia Commonwealth University, USA

ABSTRACT

Decision tree induction and Clustering are two of the most prevalent data mining techniques used separately or together in many business applications. Most commercial data mining software tools provide these two techniques but few of them satisfy business needs. There are many criteria and factors to choose the most appropriate software for a particular organization. This paper aims to provide a comparative analysis for three popular data mining software tools, which are SAS® Enterprise Miner, SPSS Clementine, and IBM DB2® Intelligent Miner based on four main criteria, which are performance, functionality, usability, and auxiliary Task Support.

Keywords: Data mining, classification, decision tree, clustering, software evaluation, SAS Enterprise Miner, SPSS Clementine, IBM Intelligent miner, Comparative Analysis, evaluation criteria.

1. INTRODUCTION

Businesses face challenges such as growth, regulations, globalization, mergers and acquisitions, competition, and economic changes, which require fast and good decisions rather than guess work. Taking good decisions requires accurate and clear analysis such as prediction, estimation, classification, or segmentation using data mining techniques. Decision tree induction and Clustering are two of the most important data mining techniques that find interesting patterns. There are many commercial data mining software in the market, and most of them provide decision trees induction and clustering data mining techniques. There is no doubt that commercial data mining software are expensive and costly, and choosing one of them is crucial and difficult decision. Therefore, this paper objective is to help organizations to make the decision of choosing one of three pre-selected famous and giant commercial data mining software by providing comparative analysis among them based on selected criteria. These software tools are: SAS® Enterprise Miner, SPSS Clementine, and IBM DB2® Intelligent Miner. The analysis is based on four criteria, which are performance, functionality, usability, and auxiliary Task Support. Performance criterion focused on hosting variety, architecture, and connectivity. Functionality criterion focused on algorithm variety, and prescribed methodology criterion. Usability Criterion focused on user interface, and visualization. Auxiliary task support criterion focused on data cleansing, and binning. However, there are many commercial data mining software in the market. Our choice for SAS® Enterprise Miner, SPSS Clementine, and IBM DB2® Intelligent Miner doesn't mean that they are the best. In addition, the chosen criteria for the comparative analysis are not sufficient to decide which of these tools is the best where there are other criteria not covered such as security, price, flexibility and reusability. Also, this paper has covered only two data mining techniques, which are decision tree induction and clustering whereas there are many other important techniques that are not covered such as Neural Network, association rules, Logistics Regression. Of course, the more techniques the tool has, is better. In short, the choice of certain commercial data mining software and the choice of certain evaluation criteria depend more on the business objectives and goals.

2. DECISION TREE INDUCTION OVERVIEW

Decision trees are class of data mining techniques that break up a collection of heterogeneous records into smaller groups of homogeneous records using a directed knowledge discovery. Directed knowledge discovery is "goal-oriented" where it explains the target fields in terms of the rest of the input fields to find meaningful patterns in order to predict the future events using a chain of decision rules^[1]. In this way, decision trees provide accuracy and explanatory models where the decision tree model is able to explain the reason of certain decisions using these decision rules. Decision trees could be used in classification applications that target discrete value outcomes by classifying unclassified data based on a pre-classified dataset, for example, classifying credit card applicants into three classes of risk, which are low, medium or high. Also, decision trees could be used in estimation applications that have continuous outcomes by estimating value based on pre-classified datasets, and in this case the tree is called a regression tree, for example, estimating household income. Moreover, decision trees could be used in prediction applications that have discrete or continuous outcomes by predicting future value same as classification or estimation, for example, predicting credit card loan as good or bad.

2.1 Decision Tree Models

Decision tree models are explanatory models, which are English rules so they are easy to evaluate and understand by people. The decision tree model is considered as a chain of rules that classify records in different bins or classes called nodes [1]. Based on the model's algorithm, every node may have two or more children or have no child, which is called in this case leaf node [1]. Building decision tree models requires partitioning the pre-classified dataset into three parts, which are training, test, and evaluation sets. The training set teaches the model by generating explanation rules of the target variables in terms of the input variables until it has been constructed. The test set makes the model more general by validating and refining it where the validation process avoids the over-fitting problem by validating the model each time by a different set of training set and pruning the tree branches. The evaluation set measures and assesses the model performance and reliability for applying the model in the future on unseen data [1]. Based on the decision tree algorithm, models could generate decision trees. Not all decision tree algorithms are the same and usable in all cases. Each decision tree has its own decision tree algorithms' features, and some features are better than others based on the case.

2.2 Decision tree algorithms

There are many algorithms for generating decision trees where the selecting particular algorithm or splitting criteria depends on many factors such as number of splits, input variables type, and target variables type. Decision trees have two types of splits, which are binary splits, and multi splits, and using a combination of input variables in binary split is more complex, slower, and increases the tree depth. In the case of using a categorical target variable type, the decision tree model classifies records into categorical classes, and the generated tree is called a classification tree. In other hand, if the target is a continuous variable type, the decision tree model estimates the value, and the generated tree is called regression tree. Decision-tree algorithms start building a tree by finding the best split for each node among target values using the input variable that does best split results. Based on the target type, there are two types of best split measures algorithms which are increasing purity measures for classification algorithms and reducing the variance measures for regression tree algorithms.

2.2.1 Classification algorithms

Classification algorithms are used as splitting criteria in classification trees by increasing the purity of a categorical variable in generated child nodes. There are several splitting measures for categorical variables including: Gini (population diversity), entropy (information gain), information gain ratio, Chi-square test. Corrado Gini, Italian statistician and economist, has invented a measure for level of economy called Gini which calculates the probability of selecting two random items from population being in the same class by calculating the Gini's score. Gini's score is summation of the squares of the classes' proportions where 1 probability indicates to pure population. Similarly, Gini measure is used to measure the best split of a decision tree node by calculating Gini's score for the node population. For example, Gini's score for a decision tree node contains 4 items from Class A, and

6 items from Class B is $(4/10)^2 + (6/10)^2 = 0.52$. Entropy measures the decision tree node impurity by finding the number of conditions or rules that determine system states. In contrast the information gain is reducing these rules (Entropy) by adding additional information [3]. ID3 stands for “Iterative Dichotomiser 3”, and it is a decision tree tool developed by J. Ross Quinlan to solve the entropy split measure problem, which was creating a bushy tree when it handles categorical input variables. The bushy tree is caused by creating splits for every value (intrinsic information of a split.), which decreases the entropy value because of reducing the number of values in each node. ID3 uses the ratio of information gain to intrinsic information of a split to measure the best split. C4.5 is the later version of ID3, and it uses the total ratio of information gain to intrinsic information of a split to measure the best split. CHAID stands for Chi-square Automatic Interaction Detector Chi-square test, and it was developed by Karl Pearson in 1900. It measures decision tree node split by the higher value of the Chi-square variation which obtained by “the sum of squares of the standardized differences between the expected and observed frequencies of some occurrence between multiple disjoint samples [3.]” CHAID algorithm requires all input variables be discrete values and target variables be binary, and that includes binning interval input variables into categorical classes.

2.2.2 Regression tree algorithms

The second type of decision tree algorithms is regression tree algorithms. The splitting criteria in the regression tree are decreasing the variance in the target variable’s values that have continuous or numeric values in generated child nodes. There are many splitting measures for numeric variables including: Reduction in Variance, F-Test, C5, and AID and SEARCH algorithms. The Reduction in Variance splitting criterion measures the values variance from the mean by calculating the sum of square of the deviation. The lowest values of variance are the closest to the mean, and the opposite is correct [3]. Ronald A. Fisher has developed the F-Test splitting criterion that “provides a measure of the probability that samples with different means and variances are actually drawn from the same population.” Therefore, F score is the ratio of the combined sample estimates to the population estimate. The highest results are the best splits [3]. C5 is also developed by J. Ross Quinlan but it handles categorical input variables on interval target variables using binary split. The best split is determined by maximizing the gain ratio

2.2.3 Classification and Regression Tree algorithms

One of the most famous algorithms that could be used for both Classification and Regression Trees is CART algorithm. CART stands for Classification and Regression Trees. It creates binary splits for both categorical and continuous target variables by performing multiple validations to improve the accuracy.

3. CLUSTERING OVERVIEW

Clustering is an exploratory data mining technique that finds interesting patterns of data by segmenting a collection of heterogeneous records into “natural groups” of homogeneous records called clusters, using “undirected knowledge discovery” or “unsupervised learning” based on similarity measures. “Unsupervised learning” is grouping the population based on the relationships in the data where there is no target variable to supervise the result of clusters based on the target variable domain as in decision tree inductions [3]. This segmentation process is based on two aspects. First, objects in each cluster should be alike, and that is called “high cohesion”. Second, objects in each cluster should be dissimilar to other clusters object, and that is called “low coupling [4].” There are many potential applications for clustering data mining technique. Clustering data mining technique could be applied in marketing segmentation by determining similar behavior customers. Marketers could match their promotions to their potential customers using historical sales transactions. Also, it could be applied for fraud detection in insurance companies by finding fraud patterns in customers’ data. There are many approaches for clustering interpretation. One approach is using decision trees modules supervised by cluster label and using the decision tree rules to assign records to each cluster. Another approach is using visualization to explain clusters based on input variables. Also, clusters could be explained by examining the input variables distributions [3.]

3.1 Clustering algorithms

There are two main methods to segment data in different clusters, which are hierarchical and partitional. Hierarchical clustering forms a tree that fits data in a sequence using one of two approaches. First, the bottom-up approach uses an agglomerative algorithm. The second top-down approach uses a Divisive algorithm. The partitional clustering method segments data using non-hierarchical techniques [2.] It includes k-Means, Expectation Maximization, Fuzzy-C-Mean, and Artificial Neural Network algorithms. In the partitional clustering method, a certain number of clusters is required to be specified before starting the process. In other hand, it is not required to specify the number of clusters in hierarchical clustering method, which may cause failing to clustering process [5]. This paper will focus on the algorithms used in the selected data mining tools software, which are SAS® Enterprise Miner, SPSS Clementine, and IBM DB2® Intelligent Miner.

3.1.1 K-Means algorithm

K-Means is one of the most common and popular algorithms published first by J. B. Macqueen in 1967. From the algorithm's name, it's required to specify a K number of desired clusters. Then, the algorithm randomly selects K data records as initial seeds for clustering. Next, the algorithm assigns the rest of the records to the closest seeds. Next, the algorithm calculates the new cluster centroids by taking the average value for every dimension. That changes the clusters' boundaries. The algorithm repeats the process of calculating new cluster centroids until clusters' boundaries become stable. There are two major downsides of K-Means algorithm. It is non-overlapping algorithm where records can not belong to more than one cluster. Also, it is sensitive to outliers [1.]

3.1.2 Agglomerative algorithm

Agglomerative algorithms start on each single data element in dataset as a cluster in order to merge them gradually until reaching one large cluster, and this is called bottom up hierarchical clustering approach. The merging process is iterated process based on distance measure between clusters using one of three common approaches: single linkage, complete linkage, or centroid distance. Single linkage approach measures distance between two clusters by measuring distance between the closest members in these two clusters. Complete linkage approach measures distance between two clusters by measuring distance between the most far-away members in these two clusters. Centroid distance approach measures distance between two clusters by measuring distance between the centroids of these two clusters [1.]

3.1.3 Divisive algorithm

In contrast to agglomerative algorithm, Divisive algorithm start on the whole dataset in order to divide it into two clusters, and keep dividing sub clusters reaching smaller clusters, and this is called top-down hierarchical clustering approach. This algorithm use purity function to partition segment of data into clusters as decision trees. This function increases impurity by decreasing inter-cluster distance average value and decreasing intra-cluster distance average value [1.]

3.1.4 Self-Organizing Maps algorithm (Kohonen Networks)

Self-organizing maps algorithm is a class of artificial neural networks that known as undirected learning neural networks or Kohonen Networks because learning process is unsupervised by target variable, and it has invented by the Finnish researcher Dr. Tuevo Kohonen. Self-organizing maps algorithm is not only used for business data mining applications such as marketing applications, fraud detection applications etc., but also it is used originally for graphical application such as two-dimensional images detection and sounds application such as sounds detection. Self-organizing maps neural network consists of two major fully connected layers, which are input layer and output layer where each layer contains units of neurons. Input layer's neurons are connected to the input variables where each neuron corresponds to one input vector. Each output layer's neuron is connected to every neuron unit in the input layer, and each input record is assigned to the output unit that has the closest weight to that record, which called "best matching neuron" process. Self-Organizing Maps algorithm does not require specifying the number of outcome clusters in order to segment data [1] [6.]

3.1.5 Demographic algorithm

Demographic algorithm segments dataset based on comparison of pairs of records by comparing individual fields' values, and it measures the distance between those records using voting techniques called Condorcet in order to assign objects to specific clusters. This technique judges the objects to be similar according to the degree of number of field's similarity. Scores of similar pair of record are calculated by getting +1 vote for every identical value in same vector, and it gets -1 vote for every dissimilar value in same vector. Based on the overall score, algorithm decides to which cluster will assign records. Demographic algorithm has two outstanding advantages. First, it could handle both categorical and numerical input variables while numerical input variables needs to be partitioned into categorical segments, which is called "predetermined tolerance." Second advantage is that it does not require specifying the number of clusters prior clustering process [7].

4. EVALUATION CRITERIA

Most organizations have huge data assets that are dispersed across the organization systems such as database servers, data warehouse systems, legacy systems, excel files on client machines etc. Therefore, the following evaluation criteria are selected to accommodate these various systems and data sources by figuring if data mining software is able to provide a complete end-to-end data mining solution for the organization needs.

4.1 Performance Criteria

Performance criterion evaluates the efficiency of the data mining tool where organizations should consider data mining software's performance ability to provide host variety, architecture, and connectivity [8.] Host variety evaluates the ability of the software to be hosted in various platforms such as Windows Linux, UNIX, etc. Architecture evaluates the software architecture flexibility what if is client-server architecture or a stand-alone architecture? Connectivity evaluates the ability of the data-mining tool to connect to various data sources such as ORACLE, SQL Server, Excel sheets, text files, etc.

4.2 Functionality Criteria

Functionality is an assortment of "capabilities, techniques, and methodologies" that facilitate solving a variety of problems [8.] Two important functionalities are covered in this paper, which are algorithmic variety and pre-described methodology. Algorithmic variety evaluates the data mining software ability to provide various algorithms for decision tree induction and clustering. Algorithmic variety will increase the data mining software functionality to provide solutions for various real world problems. This paper will focus on algorithms that are available on at least in one of the preselected data mining software. In the case of decision tree induction, there are three types of algorithms. Algorithms can handle categorical target variables, binary target variables, continuous target variables, or both categorical and continuous target variables. In addition, algorithmic variety includes what if the data mining software is able to provide a variety of splitting methods for both binary and multi splits. Finally, algorithmic variety criteria include what if data mining software is able to handle categorical and continuous input variables. In the case of clustering, there are several algorithms such as K-Means algorithm, Agglomerative Algorithm, Divisive Algorithm, Self-Organizing Maps algorithm (Kohonen Networks), and Demographic algorithm. K-Mean algorithm requires specifying the number of clusters prior clustering process, and the rest of them are not. All these algorithms are suitable for numeric input variables except demographic algorithm is suitable for both categorical input variables and partitioned numeric input variables. Also, some algorithms provide hierarchical clustering such as Agglomerative and Divisive algorithms. The second functionality criterion covers prescribed methodology that evaluates the data mining software to follow data mining methodology such as CRISP-DM.

4.3 Usability Criteria

Usability criteria evaluate data mining software simplicity of learning curve and usage. One criterion is user interface criterion that evaluates how much the data mining software is user-friendly? The other criterion is visualization criterion that evaluates data mining software ability to visualize results in clear charts, and graphs.

4.4 Auxiliary Task Support Criteria

Data mining process requires some auxiliary tasks before, during, and after conducting data mining process. One of the important auxiliary tasks is data preparation. This criteria focus on some important tasks that prepare the dataset for modeling, which are data cleansing and binning. Data cleansing criterion evaluates data mining software ability to handle defective data that make noise such as missing values and outliers. Binning criterion evaluates data mining software ability to partition numeric values into categorical values.

5. SAS® ENTERPRISE MINER OVERVIEW

SAS® Enterprise Miner is one of the end-to-end enterprise analytic solutions for data mining, introduced by SAS Corporation. This software is one part of SAS Analytics, which is “an integrated environment for predictive and descriptive modeling, text mining, forecasting, optimization, simulation, and experimental design.” Also, SAS Analytics supports decision makers by leveraging existing data and infrastructures into business intelligence environments [9.]

SAS Enterprise miner software evaluation

5.1 Performance Criteria:

- Hosting variety criterion:

SAS Enterprise Miner is platform independent software where it could be run on any platform.

- Architecture criterion:

SAS® Enterprise Miner is stand-alone architecture.

- Connectivity criterion:

SAS® Enterprise Miner uses Input Data Source node that can extracts data from wide variety of data sources: relational database, legacy mainframe systems, and ODBC, with the ability of scheduling, filtering, sorting, and data format conversions [10.] Input Data Source node reads the data source records from SAS data set or from import Wizard in order to create a dataset metadata automatically. Metadata identify variable attributes by assigning a level of measurement and role for each variable. In addition, Input Data Source node provides statistics summary for both interval-valued and categorical valued.

5.2 Functionality Criteria

- Algorithmic variety for decision tree induction:

Algorithmic Variety: SAS® Enterprise Miner provides a variety of decision tree algorithms which are CART, CHAID, and C4.5, and it provides three multi split classification algorithms that split for categorical target variables. Two of these algorithms split categorical and continuous input variables, which are Gini and entropy, and the last one splits only categorical input variable, which is Chi square test. In addition, SAS Enterprise Miner provides two other multi-split algorithms for regression tree that split for numeric target variables, which are Reduction in Variance and F-Test.

- Algorithmic variety for clustering:

SAS® Enterprise Miner provides a variety of clustering algorithms. It provides the three hierarchical agglomerative algorithms. Also, it provides K-Means algorithm, and Self-Organizing Maps algorithm (Kohonen Networks).

- Prescribed Methodology criterion:

Enterprise Miner uses SEMMA data mining methodology, which stands for Sampling, Exploring, Modifying, Modeling, and Assessing. See table 1.

**Table 1:
Relationship between SEMMA and the
Enterprise Miner Nodes**

| SEMMA | Enterprise Miner Nodes |
|------------|--|
| 1. Sample | Input Data Source, Data Partition, Sampling |
| 2. Explore | Distribution Explorer, Multiplot, Insight, Association, Variable Selection, Link Analysis |
| 3. Modify | Data Set Attributes, Transform Variables, Filter Outliers, Replacement, Clustering, SOM/Kohonen, Time Series |
| 4. Model | Regression, Tree, Neural Network, Princomp/Dmneural, User Defined, Ensemble, Memory-Based Reasoning, Two Stage Model |
| 5. Assess | Assessment, Score, Reporter |

Source: www.sas.com

5.3 Usability Criteria:

- User interface criterion:

SAS Enterprise miner provides user friendly interface where building models doesn't cost more than simple clicks and drag and drop objects into framework area.

- Visualization criterion:

SAS® Enterprise Miner provides a variety of clustering results' graphs and charts. For decision tree induction, it provides tree diagram that contains root, nodes, and leafs. The tree diagram explains decision tree's rules. Also, this software provides some important charts for lift. In clustering technique, it provides normalized means graph that ranks input variables based on their spreads on clusters where the input variable that have big spread comes first. The second graph is called distance graph that provides clusters' sizes and the relationships among them. Third graph is categorical variable profile that displays a three-dimensional profile grid for categorical variables. Fourth graph is same as third one but for Interval Variables. Fifth graph is extremely useful where it gives insight and interpretation of clusters by running a Tree node behind the scenes, and uses the cluster ID (_SEGMENT_) as a target variable. This graph shows hierarchical tree view contains a list of numbers and percentage of each cluster population. Also it contains a list of rules that assign the records to each tree node. Sixth graph is statistics graph that shows three-dimensional chart contains the input variables' statistical information for every cluster. In addition, SAS® Enterprise Miner provides Report node that consolidates the nodes' results within the process flow diagram in an HTML report, which could be displayed in a Web browser.

5.4 Auxiliary Task Support Criteria:

- Data cleansing criterion:

SAS Enterprise Miner provides many ways to solve missing values issues. It provides replacement node that fill missing values according to some accurate statistics. In decision trees, it is possible to treat missing values as an acceptable value. In clustering node, SAS Enterprise Miner provides two options for handling missing values. One option is excluding all objects that contain missing values during the clustering process, and the second option is replacing the missing values using one of the imputation methods which are Seed of Nearest Cluster, Mean of Nearest Cluster, and Conditional Mean. In addition, SAS Enterprise Miner provides Filter Outliers node to removes outliers or missing values from the current training dataset with two options: “eliminating rare values from the process flow diagram and keeping missing values in the analysis”. Also, it removes categorical valued variables situations that do not happen. Moreover, it allows removing outside various ranges of observations in interval-valued variables. Moreover, it allows removing interval-valued variable’s values by settings a variety of interval such as the standard deviation from mean, median absolute deviance, modal center, and extreme percentiles.

- Binning criterion:

SAS Enterprise Miner provides Transform Variables node for transforming the interval-valued variables in the current training dataset, and this node provides three transformation options: buckets, quantile, and Optimal Binning for Relationship to Target Transformation. Bucket binning is dividing the values into equally intervals whereas quantile binning is dividing the values into equally classes. Optimal Binning for Relationship to Target binning is splitting a variable into groups with a binary target.

6. SPSS CLEMENTINE OVERVIEW

Integral Solutions Ltd (ISL) has developed Clementine before its acquisition by SPSS. SPSS Clementine provides a wide assortment of data mining techniques associated with data preparation and visualizations tools. SPSS Clementine provides Application Templates (CATs) that encapsulated best practices and a variety of out-of-the-box functionalities, as add-on modules.

SPSS Clementine software evaluation

6.1 Performance Criteria:

- Hosting variety criterion:

SPSS Clementine could run on any platforms.

- Architecture criterion:

SPSS Clementine is a stand-alone architecture, and some versions can run on server environment and some on client machine.

- Connectivity criterion:

For data extraction, Clementine provides Front-end connectivity for databases that has kernel support such as SQL Server, DB2 and Oracle. Also, Clementine’s provides SQL pre-processing for table joins in users’ SQL queries by Clementine’s SQL optimization to improve its performance.

6.2 Functionality Criteria:

Algorithmic variety for decision tree induction:

SPSS Clementine provides a variety of decision tree algorithms, which are CART, CHAID, and C4.5. CART provides binary split classification and regression tree algorithms that split both categorical and interval target variables. CHAID provides multi split classification algorithms that split binary target variables.

Algorithmic variety for clustering:

SPSS Clementine provides a two clustering algorithms, which are K-Means algorithm, and Self-Organizing Maps algorithm (Kohonen Networks). SPSS Clementine has two major weaknesses: first it cannot cluster data hierarchically. Second, it cannot cluster dataset that has categorical input variables. In other hand, it can cluster dataset with specifying the number of clusters prior the process using K-Means algorithm. Also, it can cluster dataset without specifying the number of clusters prior the process using Kohonen Network algorithm.

- Prescribed Methodology criterion

Clementine supports CRISP-DM (Cross Industry Standard Process for Data Mining) methodology: Business understanding, Data understanding, Data preparation, Modeling, Evaluation, and Deployment. See Table 2.

Table 2: CRISP DM Methodology

| Business Understanding | Data Understanding | Data Preparation | Modelling | Evaluation | Deployment |
|--|--|--|---|--|---|
| Determine Business Objectives <i>Background</i> <i>Business Objectives</i> <i>Business Success</i> <i>Criteria</i> | Initial Data Collection <i>Initial Data Collection Report</i> Data Description <i>Data Description Report</i> | <i>Data Set</i> <i>Data Set Description</i> Selection <i>Rationale for Inclusion / Exclusion</i> | Generate Test Design <i>Test Design</i> Build Model <i>Parameter Settings</i> <i>Models</i> | Evaluate Results <i>Approved Models</i> <i>Assessment of Data Mining Results w.r.t. Business Success</i> <i>Criteria</i> | Plan Deployment <i>Deployment Plan</i> Produce Final Report <i>Final Report</i> <i>Final Presentation</i> |
| Situation Assessment <i>Inventory of Resources</i> <i>Requirements</i> <i>Assumptions</i> <i>Constraints</i> <i>Risks and Contingencies</i> <i>Terminology</i> <i>Costs and Benefits</i> | Data Quality Verification <i>Data Quality Report</i> Exploratory Analysis <i>Exploratory Analysis Report</i> | Cleaning <i>Data Cleaning Report</i> Construction <i>Derived Variables</i> <i>Generated Records</i> <i>Transformation</i> | Model Evaluation <i>Model Description</i> <i>Assessment</i> | Review Process <i>Review of Process</i> Determine Next Steps <i>List of Possible Actions</i> <i>Decision</i> | Plan Monitoring and Maintenance <i>Maintenance Plan</i> Review Project <i>Experience</i> <i>Documentation</i> |
| Determine Data Mining Goal <i>Data Mining Goals</i> <i>Data Mining Success</i> <i>Criteria</i> | | Integration <i>Merging</i> <i>Aggregation</i> | | | |
| Produce Project Plan <i>Project Plan</i> | | Formatting <i>Rearranging Attributes</i> <i>Reordering Records</i> <i>Within-Value</i> <i>Reformatting</i> | | | |

6.3 Usability Criteria:

- User Interface criterion:

SPSS Clementine provides user friendly Interface using simple clicks and drags and drop objects with changing these objects setting.

- Visualization criterion:

SPSS Clementine provides many graphical visualization tools for tables, distribution displays, plots and multi-plots, histograms, webs matrices, animation graphs. Furthermore Clementine provides evaluation visualization including Gains, Lift, Response, Profit and ROI charts. Clusters' results could be illustrated graphically using one of graphical tools such as plots.

6.4 Auxiliary Task Support Criteria:

- Data cleansing:

SPSS Clementine handles missing values by filling in missing values based on predefined intervals or class levels in the current training dataset using one of three options. One option is keeping data missing. Second option is estimating the missing data using simple method. Third option is estimating the missing data using complex method. Also, SPSS Clementine reduces 'skew' values (outliers).

- Binning criterion:

SPSS Clementine provides four binning options: Equal-Range, Equal-Sized Bins, Bins Based on Gaps, and Bins Based on Knowledge/Theory.

7. IBM DB2® INTELLIGENT MINER OVERVIEW

IBM Intelligent Miner is a set of "statistical, processing, and mining functions" to analyze data. IBM's Intelligent Miner contains three main products Intelligent Miner Modeling, Intelligent Miner Scoring, and Intelligent Miner Visualization. Intelligent Miner Modeling develops analytic models such as Associations, Clustering, Decision trees, and Transform Regression PMML models via SQL API. Intelligent Miner Scoring performs scoring operation for the models that created by Intelligent Miner Modeling. Intelligent Miner Visualization present data modeling results for analysis using one of the following: Visualizers: Associations Visualizer, Classification Visualizer, Clustering Visualizer, and Regression Visualizer. IBM's Intelligent Miner provides a variety of data mining techniques: predictive modeling, database segmentation or clustering, link analysis (associations), neural classification, neural clustering, sequential patterns, similar sequences, Radial Basis Function (RBF)-Prediction, and deviation detection (outliers).

IBM DB2 Intelligent Miner evaluation

7.1 Performance Criteria:

- Hosting variety criterion:

IBM DB2® Intelligent Miner could run on many platforms where the server can run on OS/390, OS/400, AIX, Sun/Solaris, or Windows NT, and the client can run on either of AIX, OS/2, Windows NT, or Windows95.

- Architecture criterion:

IBM Intelligent Miner consists of two parts: server and client, and it contains nine main components: user interface, environment layer, visualizer, data access, database tables and flat files, processing library, mining bases, mining kernels, mining results, result API, and export tools. User interface component allows users to define data mining functions using graphical interface. Environment layer API component is a collection of API functions, which defined and executed by the user interface component to control the mining execution and results. Visualizer component is a wide assortment of visualization tools to display the mining results. Data access component provides an access to database tables and views, or to flat files. Database tables and flat files component is the object that defined as an input or an output data, and it has logical descriptions of the physical data in order to be processed using the Intelligent Miner or other components. Processing library component provides access to database

functions. Mining bases component contains objects that build the data-mining model. Mining kernels component contains algorithms that run a data mining function. Mining results, result API, and export tools component is the output data running a mining tool or function, which could be presented by visualizer tools.

- Connectivity criterion:

Integration Server extracts data from wide variety of data sources: flat files or database tables in IBM DB2 tables, and Open Database Connectivity (ODBC) for other sources such as Oracle, Sybase, Informix, and/or SAS.

7.2 Functionality Criteria:

- Algorithmic variety for decision tree induction:

Intelligent Miner provides one algorithm for classification, which is (modified CART regression tree) algorithm.

- Algorithmic variety for clustering:

Algorithmic Variety: provides two clustering algorithms, which are Demographic algorithm, and Self-Organizing Maps algorithm (Kohonen Networks.) IBM DB2® Intelligent Miner has two major weaknesses: first, it cannot cluster dataset hierarchically. Second, it cannot cluster dataset based on predefined number of clusters. In other hand, it has a strong advantage that it can cluster dataset that has categorical input variables using Demographic algorithm.

- Prescribed Methodology criterion:

IBM DB2® Intelligent Miner does not follow standard methodology.

7.3 Usability Criteria:

- User Interface criterion:

User interface component provides a collection of graphical objects icons for creating data mining model. These objects are Data objects, mining and statistics settings objects, preprocessing settings objects, result objects, discretization objects, name mapping objects, or value mapping objects. Database objects are logical descriptions of physical data in a database or in a flat file. Mining and statistics settings objects are analytical functions, which are used to apply data mining technique after identifying the input data using data object. Result objects can present the results. Result objects present output data from a mining or statistics settings object, and an Intelligent Miner visualizer can view the result objects, or API programs can access them.

- Visualization criterion:

Two components of IBM DB2® Intelligent Miner are extremely important, which are Classification and Clustering Visualizer. Classification Visualizer provides quality view, Gains/Lift view, and the tree view. Clustering Visualizer provide three main views: Graphics, Text, and Details for models that are created by Intelligent Miner Modeling component. The graphics view illustrates clusters' size and their fields' values distributions. The graphics view gives users more control to change background colors, sort keys and orders, renaming clusters, specifying chart type (histogram, pie-chart, or tables), etc.

7.4 Auxiliary Task Support Criteria:

- Data cleansing:

Intelligent Miner handles missing values by filling in missing values based on predefined intervals or class

levels in the current training dataset. Outliers could be detected by predefining the values limits for each field, and they are treated as missing values.

- Binning criterion:

Intelligent Miner discrete objects convert range or continuous numeric field values into discrete-numeric categorical values by mapping these values to various separates intervals.

8. COMPARATIVE ANALYSIS

After analyzing each of the pre-selected data mining software according to the pre-selected criteria, I have rated each software tool in each criterion by a value from zero to three where zero is the worse, and three is the best. Then I have calculated the weighted average for each criterion where weights for performance, functionality, usability, and Auxiliary tasks support are 30%, 40%, 20%, and 10% respectively as shown in table 3. According to the comparative analysis results for SAS® Enterprise Miner, SPSS Clementine, and IBM DB2® Intelligent Miner, the Weighted Averages were 2.72, 2.49, and 1.56 respectively. Therefore, SAS® Enterprise Miner is the best data mining software among them for both decision tree induction and clustering techniques.

Table 3: Comparative Analysis' results

| Criteria | Weight | SAS EM | | SPSS Clementine | | IBM DB2 IM | |
|----------------------------------|--------|--------|-------|-----------------|-------|------------|-------|
| | | Rating | score | Rating | score | Rating | score |
| Performance (.30) | 0.3 | | | | | | |
| Hosting Variety | 0.4 | 3 | 1.2 | 3 | 1.2 | 1 | 0.4 |
| Architecture | 0.3 | 3 | 0.9 | 2 | 0.6 | 1 | 0.3 |
| Connectivity | 0.3 | 3 | 0.9 | 3 | 0.9 | 1 | 0.3 |
| Performance Scores | | 3 | | 2.7 | | 1 | |
| Functionality(.40) | 0.4 | | | | | | |
| Algorithmic Variety (DT) | 0.4 | 3 | 1.2 | 2 | 0.8 | 1 | 0.4 |
| Algorithmic Variety (Clustering) | 0.4 | 3 | 1.2 | 2 | 0.8 | 2 | 0.8 |
| Prescribed Methodology | 0.2 | 2 | 0.4 | 3 | 0.6 | 1 | 0.2 |
| Functionality Scores | | 2.8 | | 2.2 | | 1.4 | |
| Usability Criteria (.20) | 0.2 | | | | | | |
| User Interface | 0.5 | 2 | 1 | 3 | 1.5 | 3 | 1.5 |
| Visualization | 0.5 | 2 | 1 | 3 | 1.5 | 3 | 1.5 |
| Usability Criteria Scores | | 2 | | 3 | | 3 | |
| Auxiliary Task Support | 0.1 | | | | | | |
| Data Cleansing | 0.7 | 3 | 2.1 | 2 | 1.4 | 1 | 0.7 |
| Binning | 0.3 | 3 | 0.9 | 2 | 0.6 | 1 | 0.3 |
| Auxiliary Task Support Scores | | 3 | | 2 | | 1 | |
| Weighted Average | | 2.72 | | 2.49 | | 1.56 | |

9. CONCLUSION

There is no doubt that data mining software is an important factor for organization success, and it is very important to choose the appropriate data mining software that satsafies organization business needs. SAS® Enterprise Miner, SPSS Clementine, and IBM DB2® Intelligent Miner are three of giant data mining software. After evaluating these data mining software based on the chosen criteria, which are performance, functionality, usability, and auxiliary task support, I found out that the best tool is SAS® Enterprise Miner. These criteria are selected under most common considerations that most organizations have vast data assets in many data sources such as database servers, data warehouse systems, legacy systems, excel files on client machines etc, and the required data mining

software should provides solutions to perform data mining techniques using decision tree induction and clustering on the organization data.

AUTHOR INFORMATION

Abdullah M. Al-Ghpson, Virginia Commonwealth University, Richmond, Virginia, USA
alghoson@gmail.com

Education:

Information Systems, Master of Science January 2008 – December 2009
Virginia Commonwealth University, Richmond, Virginia (AACSB business accreditation)

Computer Science, Bachelor of Science August 1989 – May 1995
King Abdul-Aziz University, Jeddah, Saudi Arabia

Certificates:

Application of Multimedia Technology for E-Education January 11th 2006–April 26th 2006
Japan International Cooperation Center, Okinawa, Japan.

Experience:

Web and database developer, www.jeddahedu.gov.sa Aug 2004 – Dec 2006

Computer and Information Center, Jeddah Education, Jeddah

IT Educational Supervisor August 2001 – May 2004
Educational Development Administration, Jeddah

Computer Science Teacher – High School August 1995 – May 2001

REFERENCES

1. Berry, Michael J. A, and Gordon Linoff. “*Data Mining Techniques: for marketing, sales, and customer support*”. N.p.: John Wiley & Sons, Inc, 1997. Print.
2. Jovanovic, N.; Milutinovic, V.; Obradovic, Z.; Foundations of Predictive Data Mining. Neural Network Applications in Electrical Engineering, 2002. NEUREL '02. 2002 6th Seminar on 26-28 Sept. 2002 Page(s):53 – 58
3. Berry, Michael J. A, and Gordon Linoff. *Data Mining Techniques: for marketing, sales, and customer support*. 2nd Edition, N.p.: John Wiley & Sons, Inc, 1997. p180-183. Print.
4. Ajith Abraham, Swagatam Das., and Amit Konar. "Automatic Clustering Using an Improved Differential Evolution Algorithm." *IEEE Transactions On Systems, Man, And Cybernetics*. 38.1 (2008): 218-236. Print.
5. Castro, Vladimir Estivill. "Why so many clustering algorithms" *SIGKDD Explorations*". 4.1 (2009): 65-75. Print.
6. A. Ultsch, “Self Organizing Neural Networks perform different from statistical k-means clustering”. Retrieved December 6th, 2009, from <http://www.mathematik.uni-marburg.de/~databionics/downloads/papers/ultsch95kmeans.pdf>
7. Cabena, Peter. *Discovering data mining*. Prentice Hall, 1998. 78-79. Print.
8. Collier, Ken etl. “A Methodology for Evaluating and Selecting Data Mining Software”, 32nd Hawaii International Conference on System Sciences, 1999,
9. SAS Institute Inc. The SAS® Enterprise Intelligence Platform: SAS® Business Intelligence, 2008, retrieved in 2009 from <http://www.sas.com/apps/whitepaper/index.jsp?cid=3596>.
10. Eric Hunley, SAS, Cary, NC. SAS Data Quality – A Technology Overview, SAS Inc., <http://www2.sas.com/proceedings/sugi29/099-29.pdf>.
11. Randall Matignon, Data Mining Using SAS Enterprise Miner, retrieved in 2009from <http://www.sasenterpriseminer.com>.

12. Fast, scalable predictive analytics for the enterprise, SAS® Data Mining Solutions, retrieved in 2009 from www.sas.com.
13. SAS® Enterprise Miner™ for Desktop 6.1, retrieved in 2009 from www.sas.com.
14. Dave Norris, Clementine data mining workbench from SPSS, retrieved in 2009 from www.bloor-research.com.
15. Data Mining: Data Understanding and Data Preparation, SPSS Inc, retrieved in 2009 from www.vcu.edu.
16. Data Mining: Modeling, SPSS Inc, retrieved in 2009 from www.vcu.edu.
17. Peter Cabena, Hyun Hee Choi, Il Soo Kim, Shuichi Otsuka, Joerg Reinschmidt, Gary Saarenvirta Intelligent Miner for Data Applications Guide, retrieved in 2009 from www.ibm.com.
18. Daniel S. Tkach, Information Mining with the IBM Intelligent Miner Family, retrieved in 2009 from www.ibm.com.
19. Joerg Reinschmidt, Helena Gottschalk, Hosung Kim, Damiaan Zwietering, Intelligent Miner for Data: Enhance Your Business Intelligence. www.ibm.com.
20. IBM DB2 Intelligent Miner Modeling Administration and Programming, retrieved in 2009 from www.ibm.com.
21. IBM DB2 Intelligent Miner Modeling IBM DB2 Intelligent Miner Scoring Data Mining with Easy Mining procedures, retrieved in 2009 from www.ibm.com.
22. IBM DB2 Intelligent Miner Visualization Using the Intelligent Miner Visualizers, retrieved in 2009 from www.ibm.com.
23. Data Mining: Modeling, SPSS Inc retrieved in 2009 from www.vcu.edu. SAS Enterprise Miner Help files.
24. N. Jovanovic, V. Milutinovic, and Z. Obradovic, Member, IEEE, 'Foundations of Predictive Data Mining', 2002.
25. SAS Enterprise Miner help files. Retrieved in 2009.