Data Mining – Decision Tree Induction in SAS Enterprise Miner and SPSS Clementine – Comparative Analysis

Non-Refereed Research Abstract

Zulma Ramirez
2901 N Juan St.
Edinburg, TX 78541
(956)802-6283
ramirez.zy@hotmail.com

Diana Chairez
P.O. Box 1020
Weslaco, TX 78596
(956)472-7744
d_chairez2002@yahoo.com

Karla Martinez
901 E. Thomas Apt. 241
Pharr, TX 78577
(956)874-5525
Karla_galvan07@hotmail.com

Gaspar Diaz
5235 Mile 12.5 North
Mercedes, TX 78570
(956) 650-6389
diazgaspar@aol.com

Josue Elizondo
(956)655-9131
josue71489@gmail.com

The University of Texas Pan American

Dean of College of Business Administration
Dr. Teofilo Ozuna
Abstract

Decision tree induction for data mining is one of the most common techniques used by different organizations that appropriately fit their needs. Criteria and tools have been evaluated to develop business applications for these techniques. Our research supports two major commercial data mining software tools that are most used, SAS Enterprise Miner (EM) and SPSS Clementine. This paper is intended to create a comparative analysis on five important criteria’s such as the performance, functionality, usability, task support and security.

Introduction

Decision Tree is a tree-like graph or model type of application that is used in data mining to support and simplify strategic challenges and evaluations. When aiming for specific goals, decision tree is a great tool that will help identify the predictive utilization of a specific target outcome. It is conducted by giving a value to nodes that mainly are represented by shapes: squares (decision nodes), circles (choice nodes), and triangles (end nodes). Certain companies have had problems in gathering data to have an accurate, descriptive analysis and explanation of the results using decision rules. Also, certain companies would want to foresee a profitable outcome in many ways, for example finding out how a relationship will work out in the stock market or economically predicting upcoming events. There are different techniques in decision tree that should be considered by analysts towards decision-making.

A common practice for decision tree is to induce algorithmic functions rules. Classification algorithm has discrete allowing to predict the relationship between input data sets. Regression algorithm has continuous attributes that calculate a linear relationship between a dependent and independent variables to make predictions. For financial purposes, decision tree will balance the weight of firms’ technology goals so that they can get back to balancing their books. The Return on Investments (ROI) on the financial status of a firm is observed by rooting certain values about a subject to be concluded effectively. This conducts a method of techniques that could be considered to best approach the finance situation of the company. The calculated and targeted result values are viewed by rolling the tree backward reading from the right and moving towards the left. This will provide a value of perfect information that will help refine assumptions and possibilities of decision-making.

Commercial data mining software's are considerably expensive to purchase and the cost of training involved is high. For this, the best software that fits business needs is very important, crucial and difficult to decide. To determine which of the two data mining software's SAS and Clementine is more convenient, this paper is to provide a comparative analysis among them based on selected criteria. Decision tree software has many tools that are used to define each one of the criteria’s. However, there are many other criteria’s that may be needed to have a more precise comparative analysis. The choice of data mining software and the choice of certain evaluation criteria depend on the type of business objectives and goals.

Overview of Decision Tree Induction
Decision tree induction is an approach to data mining techniques that divide a collection of the different data into similar data information and gain knowledge on future decisions. Although it is an easy algorithmic application that is used to predict and analyze information, it has a tree structure in which a root consists of an entity that is being connected to different values that distribute to branches and leafs. A leaf represents the value of the targeted attribute that will not split (Wikipedia Inc., 2010). There are several decision tree types that will help specify various tasks for data mining and improve prediction results. A classification application is one type of decision tree that predicts and analyzes the value discretely by making a selection of unclassified data from the past (Larson, 2009). Another decision tree application used is the regression application, also known as an estimation application, that predicts value continuously that is considered as a real number from past-classified data (Al Ghoson, 2010). A combination of Classified and Regression Tree applications (CART) are applied and used in decision tree in which both have the discrete or continuous values for future predictions also known as the prediction application.

**Decision Tree Development**

A decision tree generates rules in English format. These types of rules are qualified statements that are easy to handle and be understood by humans within a database to identify a set of records (Oracle, 2005, 2008). It is a representation of knowledge information that represents a predictive result of a case. The decision tree is mainly formed by nodes as mentioned above in which each node has a bi-split, multi-split or no split. Binary split divides values into two subsets. Multiple split divides distinctive values into two or more subsets. It is not easy to develop and determine that the information inputted is the correct. For this, the decision tree has to be partitioned by pre-classified data into training, testing, and evaluation stages. The training stage is the creation of explanatory rules of the main entity that is built on. In other words, records are being collected where each contain attributes that identify a class. The testing stage constructs the decision tree model by validating and refining overlapping problems that avoid accuracy in the model. The evaluation stage measures and tests the model’s performance and reliability for future on unknown data (Al Ghoson, 2010).

**Classification Tree**

The best application for decision tree which its objective has a more valuable outcome, is the “classification”. An example of a classification application would be to target potential customers, who would more likely need a service and who does not (Larson, 2009). The goal of classification is to accurately predict the target class and its probability for each case (Oracle, 2005, 2008). Classification predicts a more discrete value that do not imply order. The simplest type of classification problem is binary which targets attributes that only have two possible values. However, there are several ways to measure splits for classification tree, in this paper we will only talk about three: Entropy, Gini, and Chi-square test. Entropy was developed by Claude Shannon, "the father of the Information Theory", is a measure from information gain that characterizes the impurity of an arbitrary collected example by searching the number of rules (Mitchel, 1992, 1997). Corrado Gini, an Italian statistician and economist, invented the Gini Index in which it measures income inequality of a distribution. A value of 0 expresses the number of a perfect equality distributed classes and the values that
approach to 1 are maximal inequality by calculating the Gin's score, an expression of percentage form and multiply by 100. It is usually defined mathematically based on the Lorenz curve, an example would be to plot the proportion of the total income of population (y-axis) that is cumulatively earned by the bottom x% of the population (Wikipedia Inc., 2010). Finally, the Chi-square test (aka "CHAID" - “Chi-square Automatic Interaction Detection”) (AI Access, 2010) was developed by Karl Pearson where he initially investigated to test two types of comparison: tests of goodness of fit (observed data) and expected data tests of independence. The test of goodness of fit refers to if there is a frequent distribution or not and if it differs from a theoretical distribution. The test of independence tests whether two variables are independent of each other (Wikipedia Inc., 2010).

**Formulae:**

Entropy = \[ \sum C \left( - \left( \frac{n_{bc}}{n_b} \right) \log_2 \left( \frac{n_{bc}}{n_b} \right) \right) \]

The entropy H of a discrete random variable X with possible values \{x_1, ..., x_n\} is

\[ H(X) = E(I(X)) \]

Here E is the expected value, and I is the information content of X.

\( I(X) \) is itself a random variable. If \( p \) denotes the probability mass function of X then the entropy can explicitly be written as

\[ H(X) = \sum_{i=1}^{n} p(x_i) I(x_i) = - \sum_{i=1}^{n} p(x_i) \log_b p(x_i). \]

where \( b \) is the base of the logarithm used. Common values of \( b \) are 2, Euler's number \( e \), and 10, and the unit of entropy is bit for \( b = 2 \), nat for \( b = e \), and dit (or digit) for \( b = 10 \).

In the case of \( p_i = 0 \) for some \( i \), the value of the corresponding summand 0 \( \log_b 0 \) is taken to be 0, which is consistent with the limit (Wikipedia Inc., 2010)

\[ \lim_{p \to 0+} p \log p = 0. \]

Gini Index = \[ 1 - \sum_j \left( p(j \mid t) \right)^2 \]

\( p(j \mid t) \) is the relative frequency of class j at node t).

Maximum (1 - 1/\( n_c \)) when records are equally distributed among all classes, implying least interesting information

Minimum (0.0) when all records belong to one class, implying most interesting information

\[ \text{Chi-square} = \chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}, \]

where

\( \chi^2 \) = Pearson's cumulative test statistic, which asymptotically approaches a \( \chi^2 \)

distribution.

\( O_i \) = an observed frequency;

\( E_i \) = an expected (theoretical) frequency, asserted by the null hypothesis;
\( n \) = the number of cells in the table.

**Regression Tree**

Regression is used to predict continuous value. An example of continuous value could be sales. Regression looks at the trends that repeat over time and predicts future outcomes. For example, sales for last year were $100 and for this year is $200. With regression it will predict that for this coming year, sales will be $300 by looking at the sales figures from the past years and continuous trends. Like in other data mining, Regression looks at relationships between the value of the prediction, and other continuous values. Because of employee being a more accurate prediction, sales will likely be affected. There are many splitting measures for continuous variables including: “Reduction in Variance, F-Test, C5 and AID and SEARCH algorithms (Al Gohosen, 2010).”

Reduction in variance criteria “measures the values variance from the mean by calculating the sum of square of the deviation (Al Gohosen, 2010).” They are located beneath each node on the decision tree. The prediction accuracy depends on what regression method is being used. One way to increase its accuracy is to use a larger tree. On the other hand, if you enlarge the tree, it will be difficult to understand. It is recommended to stay with a simple tree to be able to interpret it.

F-Test is any statistical test that has an F-Distribution (continuous probability distribution) to test if two standard deviations are equal. With F-Test you can compare one or two tail variance, depending on which variables are missing and prefer to use. In “a one-tailed test looks for an increase or decrease in the parameter whereas a two-tailed test looks for any change in the parameter.” (MathRevision Org.)

**Classification and Regression Task (CART)**

Classification and Regression Tree’s most famous algorithmic use is the CART. It is determined by maximizing both discrete and continuous outcomes that provide accuracy and a model that explain the reasons of different decisions used in the decision rules.

**Data Mining Software Evaluation Framework**

We adapt the data mining software evaluation framework to evaluate decision tree techniques in data mining tool using the following categories and associated evaluation criteria:

1. **Performance**- the ability to multi task a variety of data sources in a adequate aspect. It is focused on hosting variety, architecture, and connectivity. Criteria: (1) Software Architecture; (2) Heterogeneous Data Access.
2. **Functionality**- the inclusion of a variety of capabilities, techniques, and methodologies for data mining. Criteria: (1) Algorithmic Variety; (2) Prescribed Methodology
3. **Usability**- the ability to adjust to different levels and type of users without affecting the functionality and usefulness of the software. Criteria: (1) User Types; (2) Data Visualization
4. **Task Support** – allows the user to perform the beginning, during, and ending tasks and handles defective data and be able to cleanse, manipulate, transform, visualize the tasks that support data mining. 
Criteria: (1) Data Filtering; (2) Deriving Attributes

5. **Security** – examines whether the power system is secured under steady-state operating conditions. Using minimum number of cases from the available large number of contingencies in terms of their impact on the system security. Criteria: (1) Reliability

**SAS Enterprise Miner**

The data mining power of SAS Enterprise Miner is delivered and easy-to-use. It is used as simple as drag-and-drop interface.

**Performance**

**Software Architecture:** SAS Enterprise Miner is adjusted via a thin-client Web portal multiple users can view it. SAS Enterprise Miner can be configured on a standalone, workstation, or client/server. Enterprise Miner “supports Windows servers and UNIX platforms, making it the software of choice for organizations with large-scale data mining projects (SAS Enterprise Miner Incorporate, 2008).”

**Heterogeneous Data Access:** SAS software has the ability to access data from all areas in the connected world. It provides transparent read/write/update access to more than 60 data sources. Some of those sources include relational and no relational databases (non-SQL). Supported sources include ODBC, JDBC and OLE DB. (SAS Enterprise Miner Incorporate, 2008)

**Functionality**

**Algorithmic Variety:** SAS Enterprise Miner provides analytical depth with an unmatched suite of predictive and descriptive modeling algorithms that includes decision trees, linear and logistic regression and others. “SAS Enterprise Miner automatically generates score code in SAS, C, Java and PMML” (Bradford, 2009)

**Prescribed Methodology:** SAS provides flexible software that addresses complex problems regardless of the data mining preferences or skill level. SAS data mining process reaches five steps (sampling, exploration, modification, modeling and assessment). With this steps it can be apply advanced statistics. (Milley, Seabolt, & Williams, 1998)

**Usability**

**User Types:** SAS Enterprise Miner provides a friendly user interface designed for beginning, intermediate, and advanced users such as; “data miners, marketing analysts, database marketers, risk analysts, fraud investigators, business managers, engineers and scientists who need to make use of increasing amounts of data to make fast and accurate decisions” (www.saas.com).
Data Visualization: For decision tree induction, SAS Enterprise Miner provides a tree diagram contains root, nodes, and leafs, which explain the decision tree rules.

Task Support

Data Filtering: SAS Enterprise Miner has a goal in data mining tasks that is to apply models to training and validation of data in order to make accurate predictions of raw data. It integrates a step by step procedure for the user to learn the basic tasks Enterprise Miner offers. This procedure covers from creating a project to building a process flow diagram. Several tasks the user will be performing is accessing data, preparing the data, building multiple predictive models, comparing the models, selecting the best model, and applying the chosen model to new data (known as scoring data), filtering data, exploring data, and transforming variables (SAS Institute Inc., 2006). SAS also provides a replacement node where statistically detects a defective value for data filtering. By this, SAS will automatically eliminate unknown values from diagrams or removes categorical values that do not exist.

Deriving Attributes: SAS Enterprise Miner uses the metadata node to modify the columns information at some point in the process of the flow diagram. Modifications of attributes are such as roles, measurement levels, and order. Enterprise Miner metadata includes the data set’s name, location, library path, as well as variable role assignments measurement levels, and other attributes that guide the data mining process. Variable attributes consist of the following characteristics that are associated with a particular variable: name, label, format, informat, data type, and length.

Security

Reliability: SAS Enterprise Miner has developed in data integration, and analytics, a strong security control of the physical technical environment, application access and authentication, which is critical due to the complexity and sensitivity of the data (Meyer, 2010). SAS has created a metadata tree structure folder that provides users to have more control over the security, access privileges, and organization of models.

SPSS Clementine

Clementine is one of the most data mining solutions that use techniques to uncover patterns and trends in data, which can help improve current processes, make accurate and knowledgeable decisions for businesses. Clementine works by collecting data, routing through modeling algorithms, and using the results to create business predictions.

Performance

Software Architecture: SPSS Clementine is a stand-alone architecture. It introduce a server-based version of the product, with middle ware running on a middle-tier server taking the load off the client and using the superior performance of back-end databases to support in data
mining. Clementine uses front-end connectivity for databases that has kernel support such as SQL Server, DB2 and Oracle.

**Heterogeneous Data Access:** Clementine can read a variety of different file types including data stored in spreadsheets and databases. Data can be imported from variety of other packages including Excel, MS Access, dBase, FoxPro and Paradox using ODBC.

**Functionality**

**Algorithmic Variety:** SPSS Clementine offers a number of algorithms such as clustering, classification and others that can be used to build models to find solutions. This is very difficult for business to encounter and face research problems such as fraud detection, warrant analysis and others. It provides support for Oracle Data Mining (ODM) algorithms through its analyst-friendly user interface. (Bradford, 2009)

**Prescribed Methodology:** It’s based on the practical, real-world experience of how people do projects and there is a six stage process for data mining which are:

- **Business Understanding:** This initial phase focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition, and preliminary plan designed to achieve the objectives.
- **Data Understanding:** The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information.
- **Data Preparation:** covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data.
- **Modeling:** various modeling techniques are selected and applied, and their parameters are calibrated to optimal values.
- **Evaluation:** At this stage in the project you have built a model (or models) that appear to have high quality, from a data analysis perspective.
- **Deployment:** Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process.
### Usability

**User Types:** SPSS Clementine provides a user friendly GUI interface designed for advanced users. Building models with Clementine “doesn’t cost just simple clicks and drag and drop objects into framework area” (2010).

**Data Visualization:** SPSS Clementine provides a variety of graphical visualization tools for tables, distribution displays, plots, histograms and animation graphs. Clementine also provides evaluation visualization tools such as gains, lift, response, profit and ROI charts (2010).

### Task Support

**Data Filtering:** SPSS Clementine filters missing values in the training dataset by handling them in three ways: keeping, estimating missing data using a simple and complex method. SPSS may filter unselected data that will not be deleted. This is a safer process for end users to handle the filtering of data.
Deriving Attributes: Records are selected (select node) when there is a similarity. New attributes are obtained for the records in the data set, merge node which allows extending the attribute space. If new records are collected to the data set, append node can easily expand the data set into a large volume (W.Ji & I.Aranas., 2010). The expert user Interface enables the builder of the model or other user to connect, adapt, prepare data in terms of derived attributes, and make it available in the SPSS event builder from either the data warehouse or the operational data store. (SPSS Inc., 2004)

Security

Reliability: Manage analytical objects as organizational assets in a central repository that provides security and change management capabilities. It helps protect valuable analytic assets by providing powerful and flexible access controls. Administrators can assign the appropriate level of access to analysts and end users. Authentication can be configured using existing organizational policies and security infrastructure.

- Transit sensitive data securely between Clementine and Clementine Server through secure sockets layer encryption.

Comparative Analysis

Suppose that soon to graduate students in the University are researching a job opportunity in the finance field that utilizes decision tree tools. The students have taken data mining in computer information system and finance courses. They then compare and evaluate the mentioned data mining software, SAS Enterprise Miner and SPSS Clementine. The evaluation consists of the five criteria mentioned and one extra. Each categorical criterion will be arranged in importance of the weight that equalizes overall to 1.0 or 100%. Each software will be compared which consist of a rating. Clementine will be used as the reference of comparison to the other software where three will be directly for Clementine and the other rating will be in opposition to it.

<table>
<thead>
<tr>
<th>Relative Performance</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Much worse than Clementine</td>
<td>1</td>
</tr>
<tr>
<td>Worse than Clementine</td>
<td>2</td>
</tr>
<tr>
<td>Same as Clementine</td>
<td>3</td>
</tr>
<tr>
<td>Better than Clementine</td>
<td>4</td>
</tr>
<tr>
<td>Much better than Clementine</td>
<td>5</td>
</tr>
</tbody>
</table>

The five students compared each software tools and agreed upon the following ratings. For further evaluation other students can participate in the ratings as well to average the rating table below.

Decision Tree Tool Evaluation Scoring by Senior Students
<table>
<thead>
<tr>
<th>Criteria</th>
<th>Weight</th>
<th>SPSS Clementine</th>
<th>SAS Enterprise Miner</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rating</td>
<td>Score</td>
<td>Rating</td>
</tr>
<tr>
<td><strong>Performance (0.2)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Software Architecture</td>
<td>0.5</td>
<td>3</td>
<td>1.5</td>
</tr>
<tr>
<td>Heterogeneous Data Access</td>
<td>0.5</td>
<td>3</td>
<td>1.5</td>
</tr>
<tr>
<td><strong>Performance Scores</strong></td>
<td></td>
<td>3.0</td>
<td></td>
</tr>
<tr>
<td><strong>Functionality (0.15)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Algorithmic Variety</td>
<td>0.6</td>
<td>3</td>
<td>1.8</td>
</tr>
<tr>
<td>Prescribed Methodology</td>
<td>0.4</td>
<td>3</td>
<td>1.2</td>
</tr>
<tr>
<td><strong>Functionality Scores</strong></td>
<td></td>
<td>3.0</td>
<td></td>
</tr>
<tr>
<td><strong>Usability (0.2)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>User Types</td>
<td>0.6</td>
<td>3</td>
<td>1.8</td>
</tr>
<tr>
<td>Data Visualization</td>
<td>0.4</td>
<td>3</td>
<td>1.2</td>
</tr>
<tr>
<td><strong>Usability Scores</strong></td>
<td></td>
<td>3.0</td>
<td></td>
</tr>
<tr>
<td><strong>Task Support(0.15)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Filtering</td>
<td>0.6</td>
<td>3</td>
<td>1.8</td>
</tr>
<tr>
<td>Deriving Attributes</td>
<td>0.4</td>
<td>3</td>
<td>1.2</td>
</tr>
<tr>
<td><strong>Task Support Scores</strong></td>
<td></td>
<td>3.0</td>
<td></td>
</tr>
<tr>
<td><strong>Security(0.2)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reliability</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td><strong>Security Scores</strong></td>
<td></td>
<td>3</td>
<td></td>
</tr>
<tr>
<td><strong>Cost (0.1)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td><strong>Weighted Average</strong></td>
<td></td>
<td>3.0</td>
<td></td>
</tr>
</tbody>
</table>

**Software Architecture**: Clementine is worst than SAS and its rate with a 4 because SAS can be configured in many other workstations. SAS has standalone, workstations or client/server to be configured. Although it all depends who uses it because many people will preferred to just do analysis in standalone computer. Both, SAS and Clementine can run in Windows and UNIX platforms having the flexibility on choosing what platform to use and it’s not restricted to a specific platform.

**Heterogeneous data access**: Clementine and SAS are rate with a 3 because both extract data from a DBMS.

**Algorithmic Variety**: Clementine SPSS is better than SAS EM because the support for Oracle Data Mining (ODM) algorithms through its analyst-friendly user interface. So, I will give SAS a 2.

**Prescribed Methodology**: Enterprise Miner is the same as SPSS Clementine because EM uses steps EM’s SEMMA (sample, explore, modify, model and assets) as well as SPSS CRISP-DM (Cross-Industry Standard Process for Data Mining). SPSS gets a 3 for this criterion.
**User Types:** SAS Enterprise Miner can be used by three different levels of users (beginning, intermediate, and advanced). EM will be good from students who do not have much experience on data mining software, up to business professionals who have years of experience working with these types of software. On the other hand, Clementine is not specifically designed for beginners for this reason SAS gets a 4.

**Data Visualization:** For this criterion, Clementine gets the same rating as EM.

**Data Filtering:** Data is restored and removed by cleansing but not as accurate as the Clementine where it strategically encounters the values and replaces or fixes data.

**Deriving Attribute:** SAS contributes with Clementine is availability to rescue, remove missing values by using nodes.

**Reliability:** SAS Enterprise Miner receives a 2 because they had to upgrade its security features to be stronger in each version. The complexity in handling data in the technical environment, accesses and authentications, SPSS has a more restrictive and powerful control over security where it can be configured by the organizations security infrastructure in layer encryption.

**Cost:** SAS and SPSS cost range from 40K to 100K as well, there is not of a difference for companies that students will be looking for opportunities. However, the similarities notify students that software’s are capable of being good to consider when applying to a financial job.

**Conclusion**

“Computers and algorithms don’t mine data; people do!” (Kumar, 2004) Even with the knowledge of predictions there is always a risk of being incorrect. There is a better chance of having a more predictive and accurate information when working with the patterns provided by these decision trees. With no doubt data mining for business success has to be implemented to satisfy its needs. SAS Enterprise Miner and SPSS Clementine are very useful and great software that implement a sophisticated performance, functionality, usability, task support and security criteria. The best tool that was more effective and efficient to us was the SPSS Clementine because it considered common techniques performed by a business. Future organizations will be stacked with data assets in many data sources and this is where Clementine will come in and facilitate a decision made from predictions made.
References


