On the value of data mining tools

Wei Hong Zhao, Peter Dell*

School of Information Systems, Curtin University of Technology, GPO Box U1987, Perth, WA, 6000

Abstract

Improvements in ICTs lead to increasingly high bandwidth becoming widely available, allowing large volumes of data to be moved easily over vast distances. CEOs, CIOs, CFOs and managers in organisations can access increasingly large volumes of data to provide a knowledge basis for making important decisions. As the volume of data grows, making sense it becomes increasingly difficult. Data mining is used to extract useful knowledge from large, fuzzy datasets. There are many different data mining models, such as decision trees, neural networks, clustering, prediction, K-nearest neighbour, and association analysis.

Many software vendors have developed data mining tools, based on sophisticated algorithms. To understand how these algorithms work requires considerable technical knowledge that is beyond many IT practitioners. This paper poses the question of how much value such tools are to practitioners who do not have the technical background to fully understand the software and interpret the results.

This issue is investigated by comparing two tools based on the decision tree model. Preliminary results suggest that current data mining tools are of limited value to users without considerable knowledge of statistics and data mining.

1 Introduction

A vast amount of literature has explored data mining in great depth. Data mining is a process of extracting hidden or unknown information and knowledge from or within large, fuzzy, and random data. As information technologies develop, the size of databases has grown enormously; a major issue is how best to filter these databases to mine useful information and knowledge. Data mining has emerged as a key step in knowledge discovery in databases (Zhou *et al*, 2002).

In this paper, data mining is defined as knowledge discovery in databases and means to extract implicit, potentially useful information such as rules and constraints from the data contained in databases (Piatetsky-Shapiro and Frawley, 1991). Han (1996) notes that data mining has become a highly demanding task and it is recognized as an important research issue with broad applications.

Data mining can be conducted using either supervised or unsupervised approaches. In supervised data mining, the values of output attributes are predicted based on input

* Corresponding author. Email Peter.Dell@cbs.curtin.edu.au.

attributes, while unsupervised learning uses no output attributes because the purpose of unsupervised learning is to determine the best set of input attributes for subsequent supervised learning. Thus, unsupervised learning is used to determine if meaningful relationships can be discovered in the data and to evaluate the likely performance of a supervised learning model (Roiger and Geatz, 2002).

Supervised leaning models can be further categorized based on whether output attributes are discrete or categorical, and whether the purpose of models is to determine a current condition or to predict an outcome.

In addition to these categories, Groth (1998) defines a third group of data mining application, which he calls Visualisation Studies. These involve representing data in graphical form such as charts or maps.

To select a proper data mining strategy, it is important to understand the different data mining approaches (China Data Mining Research, 2002). A number of data mining approaches are described in the following list.

- *Classification* is a data analysis technique that classifies data using categorical labels. There are many classification approaches that have been developed by researchers in the areas of machine learning and statistics. Good classification approaches require the ability to deal with large volumes of data.

- The purpose of *estimation models* is to determine values for an unknown output attribute. For example, such a task might be to estimate the likelihood that a female customer will take a car service promotion.

- The purpose of a *predictive model* is to determine future outcomes based on knowledge contained within the database.

- *Unsupervised clustering strategies* group instances into classes without the use of a dependent variable. The goal is to identify concept structures within the data and it is often used as an evaluation tool for supervised learning.

- *Market basket analysis* is used to identify the relationships between retail goods so that it can help retailers develop marketing strategies and promotions, and arrange shelf or catalogue items.

- The purpose of *association analysis* is to find an association network hidden in the data (Rakesh *et al*, 1993). In this context, association refers to the rules or relationships that exist between two or more variables.

- *Time-series pattern analysis* refers to the process of identifying temporal patterns for characterisation and prediction of time series events (Jiong *et al*., 2000; Povinelli *et al*, 2004).

- The basic idea of *deviation analysis* is to determine the difference between observed results and outliers or deviant cases. Deviants often exist in the data in most databases and it is often useful to discover the unusual cases within data.

Supervised data mining techniques include decision trees, production rules, neural networks, and statistical regression. Association rules and clustering techniques are often favoured for marketing applications and unsupervised learning respectively. A number of data mining technologies have been developed to implement the approaches described above. Common data mining technologies are as follows:

- A *decision tree* has root, branches and leafs which are represented in a virtual model. Normally, decision trees are used in prediction models, and for finding valuable information by classifying large data sources into a tree model. IDS, the earliest decision tree algorithm, was developed by Quinlan (Groth, 1998; Roiger and Geatz, 2002; Rakesh *et al*., 1993).

- *Genetic algorithms* are methods of combinatorial optimisation based on the biological concept of evolution and use genetic methods to choose the best model based on the theory of "fittest species". Genetic algorithms are good at clustering data together and are often used in conjunction with neural networks to model data (Groth, 1998)

- *Neural networks* are based on an approach to computation that was inspired by the parallelism of the human brain (Freeman, 1994). They are a set of interconnected nodes that imitate the functions of brain. The advantage of neural networks is its successful problem solving ability in the area of data mining. However, they are a kind of "black box", and thus it is difficult for people to understand their decision making process (Roiger and Geatz, 2002).

- In the case of *agent network technology*, all data elements or categories of defined data elements are treated as interconnected agents. The concept of "impact" is central to this method of model building, which measures the influence of a given variable on another specified variable (Groth, 1998).

- *Hybrid models* are those based on combinations in this list. One approach is to develop a hybrid that makes use of several technologies. This is in contrast to hybrid systems that implement multiple approaches and let the user to choose which approach to use, such as Thinking Machines' Darwin product, which employs several different mining algorithms although the algorithms themselves are not hybrid (Groth, 1998).

- The statistical technique of *linear regression* has long been used for analysing data; the process is an aggregate method of predicting the difference between predicted and actual data sets (Groth, 1998).

Given the wide range of technologies in existence, it is impractical to investigate them all in a single study. This study focuses on the classification and regression tree (CART) algorithm, regarded by Groth (1998) as the best example of a statistical approach among the various decision tree approaches. Other approaches are beyond the scope of this study, although the authors note that further studies into the usability of software that supports other methods are required.

1.1 Decision trees

It is natural and intuitive to deal with complex problems by structuring them as a hierarchy (Watson and Buede, 1987), and deriving decision trees is a useful method for analyzing complex data sets, particularly for users with less experience in such analysis. Thus, it is perhaps especially important for tools supporting the creation of decision trees to be easily used by less experienced users.

Decision Trees are logically represented as binary trees and present the predicted results for the value of a target variable by evaluating the values of a set of predictor variables (Sherrod, 2003). A tree is composed of nodes, each of which represents a set of records from the dataset; leaf nodes the represent records that meet a certain criterion, while other nodes represent the sum of all the nodes below them. Thus, the root node represents all of the records in the dataset. The tree is constructed using a binary split technique known as recursive partitioning, which divides each node into two child nodes until leaf nodes are reached. The split is selected in such a way that the tree can be used to predict the value of the target variable at the root (Sherrod, 2003; Breiman *et al.*, 1984; Steinberg and Colla, 1995)

One of the strengths of decision trees is that interpreting their results is easier than with other tools. They can be used to facilitate understanding the big picture the model describes. If values are known for predictor variables, decision trees can also be used to predict target variables, in a process called “scoring” (Sherrod, 2003).

Despite an extensive search, very little research literature has been found in the area of usability of data mining tools. One study, conducted by the Center for Data Insight (CDI), concluded that one problem with easy-to-use mining tools is their potential misuse. Data mining tools should not only be easily learned, but should guide users through proper data mining rather than “data dredging” (Collier *et al*., 1999).

Anecdotal reports made known to the authors have been that that many data mining tools are too difficult to use for the average IT professional. Unless the user has specific data mining experience or training, the tools are reputedly difficult to learn, and produce results that are difficult to interpret. This paper focuses on decision tree data mining tools and attempts to either reject or confirm whether these tools do indeed pose such problems.

Specifically, the first objective of this research is to investigate whether decision tree data mining tools are of value to the average IT professional. By this, the authors mean whether tools produce results that can reasonably be understood by IT professionals with no specific data mining training, and whether those results can be trusted. The second objective of this research is to evaluate the user-friendliness of typical decision tree data mining tools.

2 Research method

To investigate this issue further, a case study was conducted in which an IT professional with no specific data mining experience or training tested two different data mining tools on a specific data set and compared the results. The purpose of this study is not to compare the tools themselves, but to identify issues pertaining to their use and to inform vendors and researchers involved in the development of decision tree data mining tools.

The tools used in this project are Salford Systems' CART 5.0, and Phil Sherrod's DTREG. Both systems implement decision trees based on binary splits between variables. DTREG performs V-fold cross validation to determine the optimal size of the tree (Sherrod, 2003); it is not clear from the product documentation if CART also performs V-fold cross validation.

More advanced decision tree variants such as TreeBoost and Decision Tree Forest models may be developed using these packages, this study only investigates the development of single trees.

In assessing the data mining tools, a number of measurements on the usability of data mining tools were adopted, based on the CDI study. These aspects are considered from the perspective of the average IT professional with no specific data mining training.

- Does the tool provide a user-friendly interface to the average IT professional?
- Is the tool difficult to learn? Can the user master its procedures within a short period of time?
- Does the tool give meaningful results? Does the tool present the modeling results in an easily understood format?
- Does the tool provide any wizard or other aid to help users in setting up models?
- Does the tool provide any debugging help?
- Can the tool be used without requiring any professional training?

3 Analysis of the decision trees

3.1 Sample data

The dataset used in this case study contains nominal data derived from survey responses and pertains to structured cabling decisions by Australian organizations[1]. Variables were given equal weight, and are described in Table 1.

Variable	**Values**	**Type**
Use fibre	T/F	Target
Budget	1 – 7	Predictor
Points	1 – 6	Predictor
BB No fibre because voice	T/F	Predictor
BB No fibre because ease of installation	T/F	Predictor
BB No fibre because cost	T/F	Predictor
BB No fibre because copper provides adequate bandwidth	T/F	Predictor
BB No fibre because distance not required	T/F	Predictor
BB No fibre because copper is robust	T/F	Predictor
BB No fibre because legacy	T/F	Predictor
BB fibre because bandwidth	T/F	Predictor
BB fibre because security	T/F	Predictor
BB fibre because resistance to interference	T/F	Predictor
BB fibre because electrical isolation	T/F	Predictor
BB fibre because reliability	T/F	Predictor
BB fibre because distance	T/F	Predictor
BB fibre because future expansion	T/F	Predictor
BB fibre because low cost	T/F	Predictor
Horz Fibre too expensive	T/F	Predictor
Horz Fibre too time consuming	T/F	Predictor
Horz Copper easier to install/manage	T/F	Predictor

Table 1: Summary of the sample data variables

In some cases, data were missing; that is, not all records had values for all variables. These cases were dealt with by using surrogates. Surrogate splitter variables are predictor variables that are not as good at splitting a group as the primary splitter, but which yield similar splits (Sherrod, 2003; Steinberg and Colla, 1997; Breiman *et al*., 1984). Note that splitting refers to creating branches in the decision tree by dividing a group into two sub-groups.

CART and DTREG use different techniques for dealing with missing data. This raises the issue that without prior training in data mining, the user is unlikely to understand the implications of any particular method. In the case study, the user had no better option than to use the default settings in both tools. This introduces an element of uncertainty about how the results should be interpreted.

3.2 Tree generation

Creation of a tree is central to the data-mining task in this study. Trees can be either classification trees or regression trees; when the target variable is categorical, as in the case of this study, a classification tree is generated. Otherwise, a regression tree is

[1] Structured cabling refers to the network infrastructure within a building's cable ducts, and can be made up of a number of different media. The most common media in use are several types of optical fibre cable, twisted pair cable composed of copper wire, and wireless connections.

created (Sherrod, 2003; Steinberg and Colla, 1997; Breiman *et al*., 1984).

The two tools tested during this study have some differences in the way trees are generated. DTREG has a minimum node size of 10, and a maximum number of 20 tree levels. CART provides some degree of flexibility in these settings, although the defaults are the same values as in DTREG.

Another difference between the two tools is in the rules used for generating splits. DTREG has four options pertaining to splitting rules: Gini, Entropy, Misclassification Cost, and Variance. CART has six options: Gini, Entropy, Symmetric Gini, Entropy, Class Probability, Twoing and Ordered Twoing. This alone illustrates the key point made in this paper – that without prior training or experience in data mining, it is quite likely the user will not understand the difference between these options. Hence, the user is left with some doubts about the model created by the tool. Were the appropriate options selected? How did these options affect the tree generation? In this study, Gini was chosen as the default splitting rule, and is supported by both tools.

After trees have been generated they must also be validated. As with splitting rules, there are several methods for validating trees. V-fold cross validation is the default validation method in both CART and DTREG, and hence was chosen for this study. The authors make the point that in the untrained user is unlikely to understand how V-fold cross validation works, or how its use affects the tree[2].

A third aspect of tree generation for classification trees for which the user needs prior knowledge is the method used for misclassification cost. To change the default values requires understanding of the data being mined. Essentially, these settings can be fine-tuned if it is particularly important not to misclassify a particular variable in comparison to other variables. Again, inexperienced users will have little use for such settings, although they may be important aspects of tree generation.

Category weights can be configured in both tools. These settings are similar to those in multi-criteria decision analysis tools such as Analytical Hierarchy Process (AHP) (Saaty, 1980), and Visual Interactive Sensitivity Analysis (VISA) (Visual Thinking International, 1996). In this study, the default weights (referred to by the tools as “priors”) were used in both tools.

Other settings vary between the two tools. CART provides for missing value and high level categorical penalty settings on variables, and features such as change class names, set categorical search parameters and set range of variable values used in analysis. Further, CART allows the combining of multiple trees together. None of these options were used in the analysis conducted during this project. DTREG provides similar functions, but uses different names. This further confuses things for the inexperienced user; even if one does have prior experience using similar tools, the possibility of

[2] V-fold cross validation is used to determine the optimal tree size and does not require a separate dataset for assessing the accuracy and size of the tree.

confusion is introduced by the use of different nomenclature by different vendors.

DTREG also provides features such as TreeBoost and Decision Tree Forest, which are said to be more accurate than a single tree. These two models are outside the scope of the current project as direct comparison with CART is not possible. However, the fact that such options are available raises a number of issues. First, the user is informed that they are better than a single tree, and yet such models are more difficult to understand. Second, that these options are not necessarily available in all tools gives rise to the possibility of the data-mining task producing different results simply as a consequence of which tool was chosen. Thus, complete analysis requires the user have a solid understanding of data-mining techniques, and that they select a tool that supports such techniques. It is unlikely anyone other than an experienced user will be able to make such a decision, and yet the decision may affect the outcome of the process.

3.3 Comparison of results

Figure 1 shows the tree generated by DTREG. The uppermost node is the tree root, and represents all records in the dataset. UseFibre is the target variable, and this can be predicted by constructing a tree out of the variables Budget and Points. Specifically, nodes 2 and 3 are split by the Budget variable, and nodes 4 and 5 are split by the Points variable. It is noted that each node also indicates a number of misclassification costs.

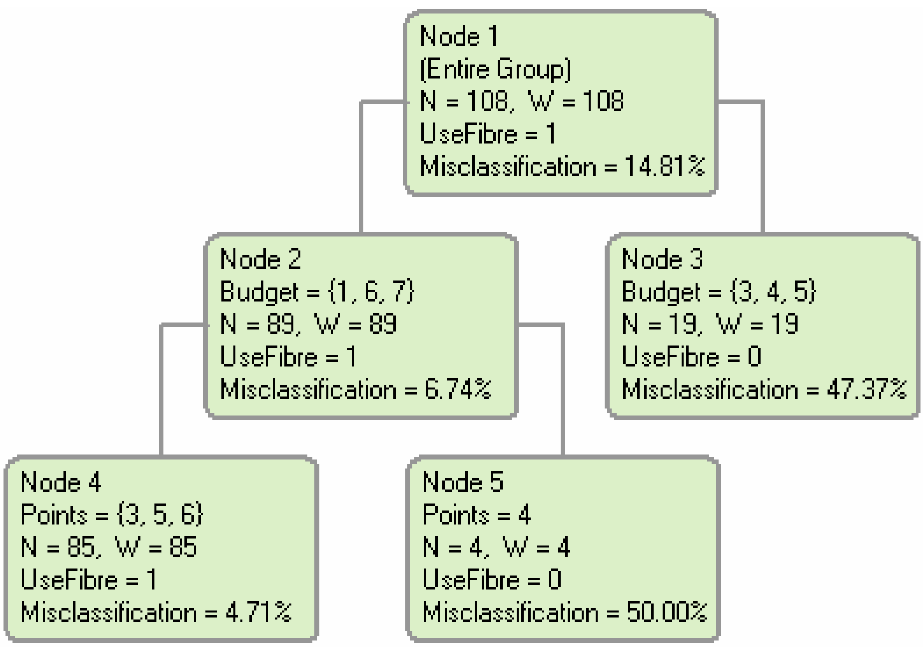

Figure 1: Tree generated by DTREG

This tree forms the basis of the following algorithm to determine the outcome of UseFibre:

```
IF Budget IN {1, 6, 7} THEN
  IF Points IN {3, 5, 6} THEN
    UseFibre = 1
  ELSE
    UseFibre = 0
ELSE
  UseFibre = 0
```

Figure 2: Algorithm based on DTREG results

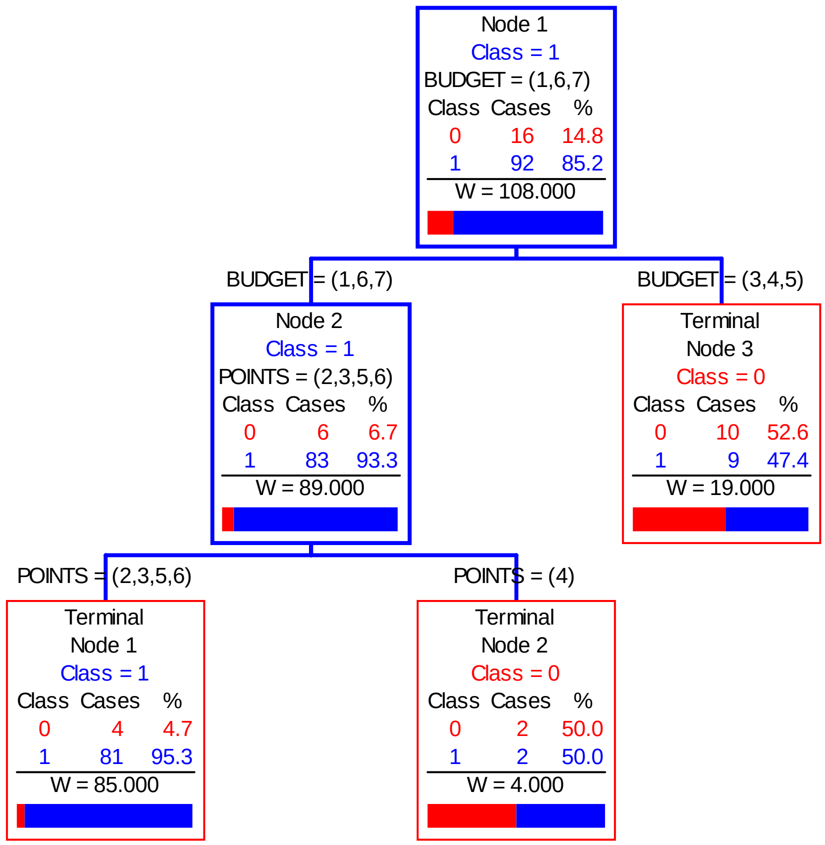

Figure 3: Tree generated by CART

However, CART produces a different tree, as illustrated in Figure 3. While at first glance this tree appears similar, close examination reveals a subtle difference. The first split in the tree is based on the budget variable and is the same as the tree produced by DTREG. However, the second split is slightly different. Again, it is based on the Points variable, however, while DTREG groups records where Points was equal to 3, 5, and 6, CART also includes Points = 2 in this group. This tree forms the basis of the algorithm illustrated in Figure 4, slightly different to that in Figure 2.

```
IF Budget IN {1, 6, 7} THEN
  IF Points IN {2, 3, 5, 6} THEN
    UseFibre = 1
  ELSE
    UseFibre = 0
ELSE
  UseFibre = 0
```

Figure 4: Algorithm based on CART results

Finally, in generating the trees, the two tools impute variables with differing importance levels. While both tools assigned an importance of 100% to Budget, DTREG assigned an importance of 51.7% to Points, while CART assigned Points an importance of 50.4%.

Such minor differences between the two trees will probably have a negligible impact on the outcome of the analysis. However, the authors flag as an issue the possibility that other tools might produce further different results that may not be so minor. This has two implications. First, an implication for anybody using data mining tools is that their analysis may be affected by the tool they choose to use, even when those tools purport to implement the same data mining techniques. A second implication is for the research community to further test related tools to determine the degree to which their results vary.

As well as comparing the trees generated by the tools, a comparison of Lift and Gain Charts created by the tools was conducted. Lift and Gain are used to assess the value of a predictive model (Sherrod, 2003), however, as with many of the aspects previously discussed, the authors feel that without prior background in data mining to aid with the interpretation of such charts, they will be of little value to the practitioner.

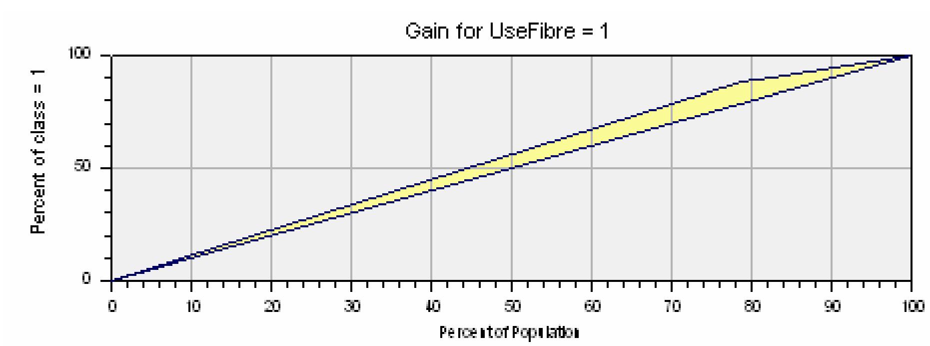

Figure 5: Gain for UseFibre (DTREG)

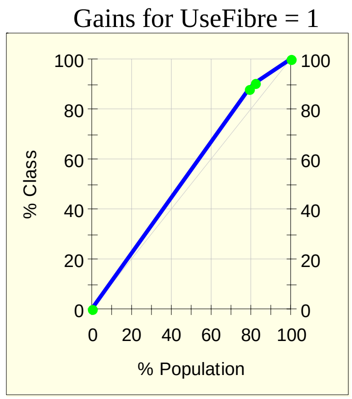

Figure 6: Gains for UseFibre (CART)

Gain charts from both tools are the same (Figure 5 and Figure 6), however, the Lift charts (Figure 7 and Figure 8) do vary slightly. Specifically, while CART presents the Lift as a series of descending steps, two sections of the DTREG graph have a gradient that is neither horizontal nor vertical. Cumulative Lift charts from both tools have a similar difference. It is also noted that both tools have some differences in terminology, for example, what DTREG refers to as cumulative gain (i.e. the ratio of cumulative percentage of class to cumulative percentage of population), CART dubs cumulative lift.

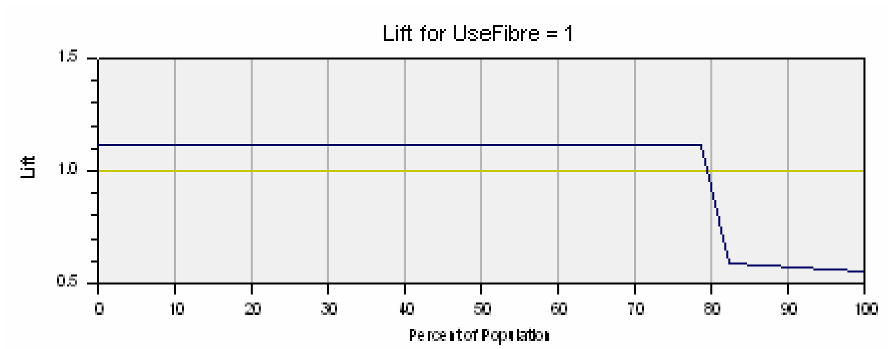

Figure 7: Lift for UseFibre (DTREG)

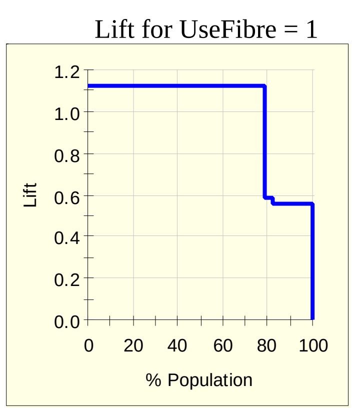

Figure 8: Lift for UseFibre (CART)

4 Usability analysis

The second objective of this study is to investigate the usability of decision tree data mining tools. Six factors are evaluated in the following sections.

4.1 *Does the tool provide a user-friendly interface to the average IT professional?*

When DTREG is opened, it displays a toolbar at the top and a navigation bar at the left side of the window, while the main area of the window is left blank. Similarly, CART displays a toolbar at the top of the window when the tool is opened and the remainder of the window is left empty.

Upon first loading the software, neither tool provides any guide, prompt or other useful message to indicate to the user how to proceed. However, DTREG does prompt the user to create a new project using the appropriate toolbar button if the user attempts to use the navigation bar on the left-hand side.

DTREG also guides the user through the process for setting up variables after loading an appropriate data file. No explanations for terms such as “Target Variable”, “Predictors”, “Weight” and “Categorical” are provided, so the user needs at least basic understanding of decision trees in order to set up their variables correctly. Following this stage the user is guided through the process of setting up class labels. At this stage, some explanation is provided to the user.

The following step is to design the model. Here, the researcher is expected to select the tree-generation algorithm, methods for testing and pruning the tree and pruning control, tree size controls, methods for dealing with missing values, and setting up surrogate predictors. No explanation for any of these settings is provided to the user; hence, any user without sufficient understanding of decision trees is likely to leave the default settings. The user in this case study was left with the feeling that these options were important, but had no idea how to use them. Consulting the user manual is somewhat helpful, however it is clearly written for an audience with prior data mining knowledge.

CART also begins the process with opening the data files. Upon completing this step, a model setting window appears, but has no immediate guide available to the user on how to proceed. CART also does not guide the user through the process in a step-by-step fashion in the manner that DTREG does. The software did not explain key data mining terms used; the user manual also did not explain any of the terms used and focused purely on the mechanical aspects of how to use the software interface.

The user’s reaction to both tools was that at least some degree of knowledge in the area is required in order to confidently set up the models used to create the decision trees. The user manuals provided with both tools did little to change this situation, as both were written for the experienced user. Thus, both tools were relatively unfriendly to the average IT professional with little prior data mining experience.

4.2 *Is the tool difficult to learn? Can the user master its procedures within a short period of time?*

Three issues lead the authors to conclude that the tools tested are difficult to master within a short period. First, the user had to spend considerable time reading the manuals to gain an understanding of how the tools worked. Given that the DTREG manual has 176 pages and the CART manual has 300 pages, this is far too large an undertaking for anybody with limited time available. Further, as stated above, the documentation provided with both tools is written from the perspective of a user with prior knowledge in data mining.

The user's experience was that considerable time was required in order to become proficient in using both tools. This time is not spent learning how to use the software, which is in itself relatively simple. Rather, time is required learning data mining terminology and gaining an understanding of the techniques, so that the tool can be used effectively and the results can be properly understood. The difficulty for the user is that until such an understanding has been reached, the will have little understanding of which options should be used.

4.3 *Does the tool give meaningful results? Does the tool present the modeling results in an easily understood format?*

The tree diagrams produced by the tools and illustrated in Figure 1 and Figure 3 are themselves easily understood. The Lift and Gain charts are also clear, although they do require deeper knowledge of the workings of decision tree algorithms.

However, the question must be asked whether the results are meaningful if the user has little confidence that the correct options were selected. If one considers the tree illustrated in Figure 9, and compares this with the tree in Figure 3, it becomes clear that the options chosen are highly important. The tree in Figure 9 was generated by incorrectly configuring the variables as continuous, rather than categorical, variables. Note that continuous variables are the default in CART, so an inexperienced user could easily make such a mistake.

Much has been made of the point that without background knowledge, the user will be unable to select appropriate options for the tree generation process. This is a further example of the importance of the user having appropriate skills and knowledge before attempting to use decision tree tools.

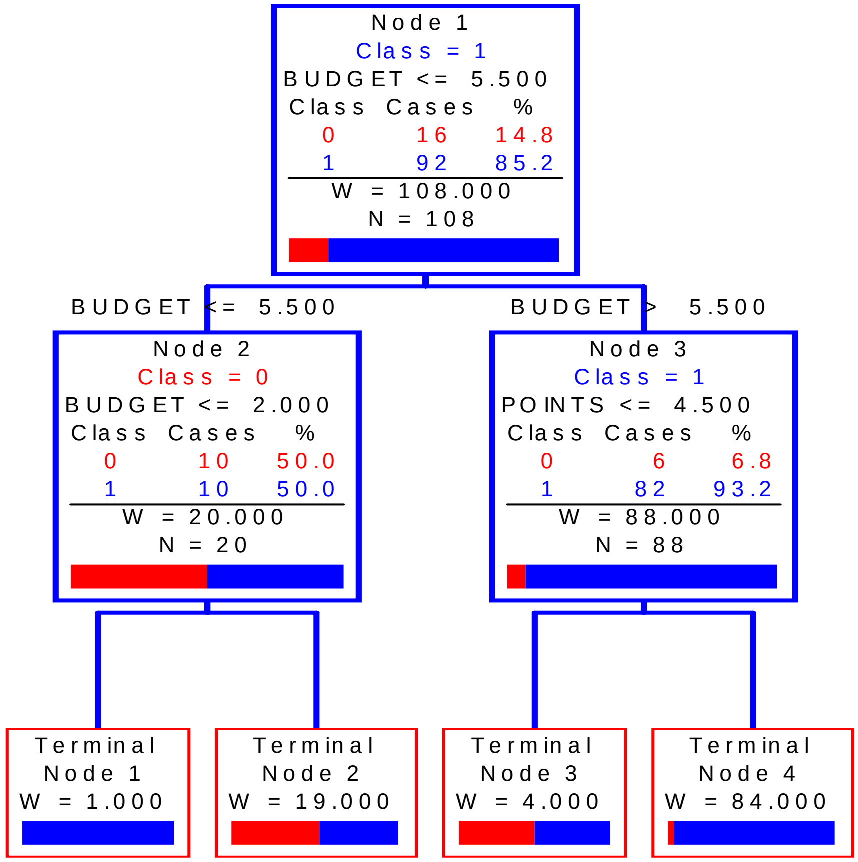

Figure 9: Tree generated with incorrect options

4.4 Does the tool provide any wizard or other aid to help users in setting up models?

CART provided no wizards or other aids to assist in creating the model, and the validation for model setup only applied to the target variable.

DTREG did provide a simple wizard to guide the user through the model creation process; however there was little or no explanation of what was required at each step. When the user's input was validated, it appears it was only partially validated, such as insisting target and predictor variables be selected before continuing, but not ensuring the user had addressed whether the variables were categorical or continuous.

The authors believe it is easy for the inexperienced user to make mistakes during the model creation process; an observation which has obvious implications for vendors.

4.5 Does the tool provide any debugging help?

The user could not determine any method in either tool to help troubleshoot problems in the model and subsequently developed tree. This is a significant shortcoming in both tools.

Presumably a more experienced user would be less likely to need such a feature in a software package, however this issue is not specific to data mining tools. It is noted here for the sake of completeness.

4.6 Can the tool be used without requiring any professional training?

The user's experience was that knowledge of data mining and statistical analysis techniques is needed in order to understand the terminology used by both packages. A user without knowledge in this area will not be able to master the software and will not be able to interpret the results generated by the tools. The user may also set up the model incorrectly, which may lead to the user being misled by incorrect results.

Thus, the authors conclude that the use of such tools without appropriate training is either risky, due to the possibility of misinterpreting the results, and inefficient, due to the time and effort required to master the software, or both.

It must be noted that these findings are based on the experience of a user with two particular tools, and may not be representative of other data mining tools. However, the point that any organization hoping to derive value from such tools should seek to ensure the users are suitably prepared for whichever package they choose to adopt.

5 Conclusions

Data mining technology has great potential. Developments in the technology have been driven by the growth in data volumes and the desire to find hidden relationships and patterns in the data. As data mining tools become more common, people with little or no knowledge in data mining will inevitably use the tools.

This becomes a serious issue when organizations may inform strategic decisions with results from tools that were perhaps used incorrectly. This paper makes a number of suggestions.

First, that vendors develop data mining tools that are increasingly user-friendly, and which include safeguards to alert users when they may be using the tools inappropriately.

Second, organizations adopting data mining tools should ensure that personnel involved in their use should be prepared with the requisite understanding of data mining and statistical methods, to ensure they use the tools effectively.

Finally, relevant education and training programs should include appropriate coverage of statistics and data mining topics in their curriculum to ensure that graduates of such programs are prepared for the use of data mining tools.

References

Agrawal, R., Imielinski, T. & Swami, A. (1993). Mining Associations between Sets of Items in Massive Databases, *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD'95)*, Washington D.C., May.

Breiman, L., Friedman, J., Olshen, E. & Stone, C. (1984). *Classification and Regression Trees*. Pacific Grove: Wadsworth.

China Data Mining Research (n.d). *All Data Mining Software*. Available: http://www.dmresearch.net/news/Abroad/1000000003.htm, Retrieved: 15 April 2005.

Collier, K., Carey, B., Sautter, D. & Marjaniemi, C. (1999). A Methodology for Evaluating and Selecting Data Mining Software. *Proceedings of the 32nd Hawaii International Conference on System Sciences*, Hawaii.

Freeman, J.A., (1994). *Simulating Neural Networks with Mathematica*. Addison-Wesley.

Groth, R. (1998). Data Mining: A Hands-On Approach for Business Professionals, *Introduction to Data Mining*. New Jersey, USA: Prentice-Hall.

Han, J.W. (1996). Data Mining Techniques, *Proceedings of the ACM-SIGMOD International Conference on Management of Data (SIGMOD'96)*. Montreal, Canada.

Piatetsky-Shapiro, G., & Frawley, W. (1991). *Knowledge Discovery in Databases*. Cambridge, MA, USA: AAAI Press / The MIT Press.

Povinelli, R.J., Johnson, M.T., Lindgren, A.C., & Ye, J. (2004). Time Series Classification: Using Gaussian Mixture Models of Reconstructed Phase Spaces. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 16, No. 6, pp. 779 – 783.

Roiger, R. & Geatz, M. (2002). Data Mining: A Tutorial-Based Primer, *Data Mining: A Closer Look*. Addison Wesley.

Saaty, T.L. (1980). *The Analytic Hierarchy Process*. New York, USA: McGraw-Hill.

Sherrod, P.S. (2003). *DTREG: Classification and Regression Trees, For Data Mining and Modeling.* Available: http://www.dtreg.com, Retrieved: 25 March 2005.

Steinberg, D & Colla, P. (1997). *CART — Classification and Regression Trees*. San Diego, CA, USA: Salford Systems.

Visual Thinking International (1996). *V·I·S·A User Guide*. Glasgow, UK: Visual Thinking International Ltd.

Watson, S.R. & Buede, D.M. (1987). *Decision Synthesis: The principles and practice of decision analysis.* Cambridge, UK: Cambridge University Press.

Yang, J., Wang, W. & Yu, P.S. (2000). Mining asynchronous periodic patterns in time series data, *Proceedings of the sixth ACM SIGKDD international conference on knowledge discovery and data mining*, Boston, MA, USA, August 20-23.

Zhou H.F, Zhu J.Q., Zhu Y.Y. & Shi B.L. (2002). ARminer: a data mining tool based on association rules, *Journal of Computer Science and Technology*, Vol.17, No. 5, pp. 594 – 602.