SAFETY SCIENCE Monitor

ISSUE 1 2005

Article 5 VOL 9

PATIENT CLASSIFICATION AND THE PORT PACKAGE

P. R. HARPER
Faculty of Mathematical Studies, University of Southampton

THE GENERAL CLASSIFICATION PROBLEM

As a direct consequence of individuality, people typically differ in a number of medical, physical and socio-economic characteristics, for example by age, profession, severity of an accident in the workplace, time off work and speed of recovery. Groups of patients with resource needs, for example representing those who are a particular age or those who have had a particular injury, are usually considered as groups of patients with similar needs. In fact these groups are typically heterogeneous and require more detailed modelling for classification. Resulting resource needs and corresponding costs vary from person to person within each group.

From both a clinical and operational perspective, it is desirable to be able to divide this heterogeneous group into smaller homogeneous (in terms of some measure) sub-groups. Homogeneity brings the benefits of increased certainty in individual patient needs and resource utilisation. For example, given an individual patient we can classify them into a patient sub-group in which we know, from past experience and data, that their risk of an accident at work is likely to be within a certain range of likelihood with a given confidence. The risk or costs for this population group will typically substantially differ from the predicted risk and costs of other groups. The purpose of classification in this example would be to produce tight bands with high confidence. Thus with the added knowledge and confidence of risk of injury and expected costs, the potential for improved efficiency and effectiveness in targeted interventions is considerable.

An important criterion for a good classification procedure is that it not only produces accurate classifiers (within the limits of the data) but that it also provides insight and understanding into the predictive structure of the data (Breiman *et al.*, 1984). For example, finding which socio-economic and other characteristics contribute to the risk of a an injury in the workplace not only provides valuable assistance in classifying individuals into risk groups with some certainty, but more generally has advanced the knowledge and understanding of who is at higher risk of injury.

There are two elements to a general classification problem. Measurements are made on some case or object (for example a person who has had an accident with measurements including age, sex, profession, type of injury, hospital treatment, time off work, outcome etc.) and based on these measurements a prediction is made as to which class a case is in. The prediction is made following a pre-defined classification rule.

In mathematical terms we define X to be the measurement space containing $x = (x_1, x_2, \ldots x_m)$, the measurement vector, where each x_i is a measurement taken on a case. The method should, given any x in X, have a classification rule to assign one of the classes $(1,2,3,\ldots J)$ to x, where J is the number of classes. The classifiers are based on past experience using a combination of expert knowledge and past data with their relevant outcomes. For example, the classifiers to be defined could come from a hospital database combined with the expert knowledge of the consultants, specialty managers and other medical staff. Each measured variable is *continuous*, *nominal* or *ordinal* in nature. A variable is continuous if the measured value is a real number (e.g. time off work, age). A variable is nominal if it is a finite categorical set with no natural ordering (e.g. sex, hospital, profession). A variable is ordinal if it is a finite categorical set with a natural order.

COMPARISON OF CLASSIFICATION ALGORITHMS

There exist many different classification algorithms, for example regression models, treebased algorithms (CART), Artificial Neural Networks (ANN), Discriminant Analysis (DA). Intrasubject comparisons have been considered in the past, for example, within statistics (Remme *et al*., 1980), within symbolic learning (Clark and Boswell, 1991) and within neural networks (Xu *et al.*, 1991). Other authors, for example King *et al*. (1995) and Harper (2002) have compared different algorithms for different types of datasets. Here the algorithms were evaluated using a number of criteria to measure the accuracy and the computing time taken to produce results, and the comprehensibility of the results and the ease of use of the algorithm to relatively naive users.

Research in this area indicates that in practice there is no single *best* classification tool but instead the best technique will depend on the features of the dataset to be analysed and any preferences of end-users. The research has made a start in investigating what these features are with particular emphasis on healthcare data. A summary of the main findings are as follows (Harper, 2002):

- Regression models consistently have fastest run-times, although the difference in times compared to CART and DA is likely to be insignificant in practice. Neural Networks require significantly more time to train and validate models.
- In general CART, Regression and Neural Network classification approaches give similar accuracies, although CART typically gives consistently good results. DA often performs poorly.
- CART is well suited to datasets with large skew (>1) and kurtosis (>7) and where there is a large proportion of categorical independent variables. CART makes no assumption about the underlying distribution, hence why CART performs consistently well. In contrast, these conditions limit the performance of discriminant and regression models, where the data is furthest from the (multivariate) normal.
- Neural Networks often produce the best accuracy when dealing with smaller datasets but perform slightly disappointingly when handling dependent variables with high levels of variability or deviance.
- If ease of use and human understanding are high priority, symbolic algorithms such as CART should be chosen.

A survey of medical users from a number of different organisations revealed that tree-based tools, such as CART, do have a greater practical appeal than that of the other tested techniques. This is a measure of the extent to which the CART algorithm produces comprehensible results that are generally easier to interpret by medical staff than the results of other algorithms, and on the time it took for hospital staff to understand the technique, prepare the data and actually perform the analysis to produce correct and meaningful results. In practice clearly a balance must be made between the accuracy and interpretability of a proposed technique. Accuracy is undoubtedly important, especially when considering a number of variables such as predicting outcome. We might however wish to avoid a situation in which we are obtaining accurate predictions but where the form of the classifier is complex and little confidence and knowledge is gained on the data structure. Such a *black box* approach is limited in producing interpretable classification rules both for understanding the prognostic structure and for the planning and management of healthcare in general.

CLASSIFICATION AND REGRESSION TREES (CART)

The foundations of CART

Classification and Regression Trees (CART) is a classification method that has been successfully used in many healthcare applications. Example applications include creating case-mix groups (Smith *et al*., 1992), minimum data requirements (Hornberger *et al*., 1995), cancer survival groups (Garbe *et al.*, 1995) and Intensive Care (Ridley *et al*., 1998). Breimen *et al* (1984), the founders of the technique, use the following University of California study as a means of introducing the reader to the technique. The study uses measurements recorded when a heart attack patient is admitted to hospital and attempts to classify the patients into low-risk and high-risk groups. Nineteen variables are recorded in the first twenty-four hours, including age, blood pressure and seventeen other ordinal and binary variables summarising the patient's condition. The CART method produces a tree that, by answering a series of yes/no questions, can be used to classify the patient. The authors found that it was possible to identify a high risk group of those patients not surviving more than 30 days based on minimum systolic blood pressure, age and whether sinus tachycardia was present. Figure 1 shows the tree produced in the study. In the tree, the letter F indicates low risk and G for high risk.

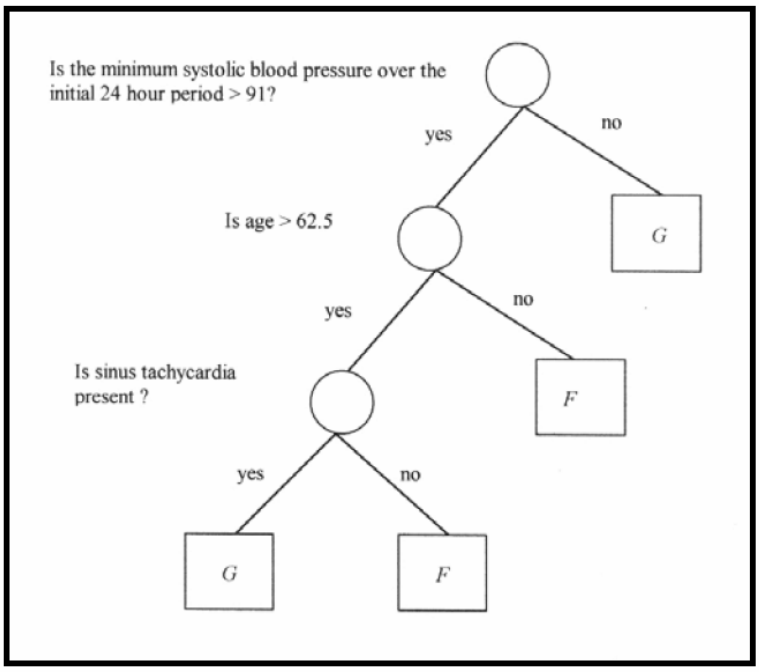

Figure1: CART analysis of cardiac patients (from Breimen *et al*., 1984)

The method

The first step in producing a tree is to decide which variable is to be predicted (LoS, operation times, survival rates etc.). If the variable is ordinal then the variance is used to measure the *purity* in the group, if the variable is categorical then deviance is used. An algorithm is used to split the original dataset into sub-populations of increasing purity (decreasing variance or deviance). At each junction of the tree is a *node*. A *terminal node* is a node at the end of the branch of the tree. At each node in the tree the algorithm searches through each of the independent variables in turn. For each variable it finds the best binary split that produces a node with the smallest variance or deviance. Then it selects the variable that has produced the best binary split (best of the best). The *parent node* will thus be split on this variable with the split as defined. This may have the effect of leaving one of the *child nodes* with a higher variance/deviance than the parent node. The algorithm however continues to branch from each of these child nodes until defined stopping rules have been fulfilled.

An issue in CART analysis is when to stop the partitioning, i.e. when do we say that the variance has not significantly reduced? It would be possible to create a tree where each terminal node has zero variance by having just one case in each node. However this would be statistically irrelevant and practically useless. It is necessary therefore to introduce *stopping rules* so that terminal nodes have sufficient size to yield reasonable and statistically robust results. Stopping rules include:

- Stop when nodes contain a certain number of cases.
- Stop when reduction of variance is below a certain threshold.
- Stop when a maximum number of terminal nodes (or layers) have been produced.

Care must be exercised when defining stopping rules and should account for the number of cases in the dataset. A terminal node with less than 30 cases, for example, can be expected to yield little predictive power and lack statistical robustness. Standard bounds are no less than 50 cases per node, a significance level of 1% on the reduction of variance in order to split a node, and a maximum of around 10 terminal nodes. Once a tree is constructed statistical summaries can be produced at the terminal nodes which can be used to form the classes.

CART components

There are 4 components required to construct a regression tree:

1. A set of questions of the form: Does x_i belong to the set A. The answer to such questions induces a split of the predictor space, cases associated with A and those with the complement of A. The sub-samples form the nodes.

2. A goodness of split criterion $\sqrt{}(s, t)$ that can be evaluated at any split s at any node t.

3. A means of determining the appropriate size of the tree.

4. Statistical summaries at terminal nodes of the tree, for example, node averages and frequency distributions.

Validating the trees

Once a tree has been produced it should be validated to give an estimate of the accuracy of its classifications. The same data that is used to construct the tree cannot be used to test the classifications, as the estimate will be over-optimistic. This is overcome by splitting the data into two sets, A and B. The cases in A must be independent and identically distributed to the cases in B. This has the drawback that it reduces the sample size used in the construction of the tree. Set A can be used to train the data and build the tree (training set) and set B to test the robustness and validity by forcing the data through the tree (testing set).

For smaller samples there is a technique called *V-fold cross validation*. This involves three stages:

1. Split data into v sub-sets

2. Classify on A-A_v for each v

3. Cross validate over all samples, combine miss-classification rates to measure accuracy Standard statistical methods can be used to compare the values of the training set and test set at any node. The significance test to compare data sets can be described as the following:

Let n_1 = number in group at node from training set

n_2 = number in group at node from test set

(A) Categorical dependent variable

Let r_1 = number responding 'yes' in group at node from training set

r_2 = number responding 'yes' in group at node from test set

Let π_1 and π_2 be the true proportions responding 'yes' in the training and testing populations respectively.

For a large sample (>20) r_1 is approximately normal distributed with mean $n_1\pi_1$ and variance $n_1\pi_1(1-\pi_1)$, and r_2 with mean $n_2\pi_2$ and variance $n_2\pi_2(1-\pi_2)$.

Under the null hypothesis of $\pi_1 = \pi_2 = \pi$ we obtain:

$$Variance\left(\frac{r_1}{n_1} - \frac{r_2}{n_2}\right) = \frac{\pi(1-\pi)}{n_1} + \frac{\pi(1-\pi)}{n_2}$$

and

$$Expectation\left(\frac{r_1}{n_1} - \frac{r_2}{n_2}\right) = 0$$

where π is estimated by $p = \frac{r_1 + r_2}{n_1 + n_2}$

A 95% confidence interval for $\pi_1 - \pi_2$ is given by:

$$\frac{r_1}{n_1} - \frac{r_2}{n_2} \pm 1.96\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

(B) Ordinal (continuous) dependent variable

Let $\bar{x}_1$ = the mean value within the group at node from training set

$\bar{x}_2$ = the mean value within the group at node from test set

s_1^2 = the estimated group variance at node from training set

s_2^2 = the estimated group variance at node from test set

Let μ_1 and μ_2 be the true means in the training and testing populations respectively. Let σ_1^2 and σ_2^2 be the true variances in the training and testing populations respectively estimated by s_1^2 and s_2^2 for large n_1 and n_2.

Under the null hypothesis of $\mu_1 = \mu_2$ we derive:

$$Variance(\bar{x}_1 - \bar{x}_2) = \left(\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

and

$$Expectation(\bar{x}_1 - \bar{x}_2) = 0$$

Hence a 95% confidence interval for $\mu_1 - \mu_2$ is calculated as:

$$\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2,(n_1+n_2-2)} \sqrt{\left(\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

Confidence intervals can be used to judge whether or not the two data sets are statistically the same. If the interval produced at any given node contains zero then it suggests the test data supports the training data; if zero is not contained in this interval then the node should be treated with caution.

The CART algorithm details

Classification

1. For the current group calculate the total variance within the group.

2. For each value of independent variable, calculate the total value of the dependent variable in that group ($\sum x$), the total value squared ($\sum x^2$), and the number of items of data in that group (N).

3. Sort the values of the independent variable into increasing order of the mean value of the dependent variable.

4. For each independent variable calculate the best point at which to split the sorted mean values to produce the minimum variance.

5. Split the data based on the best independent variable in order to reduce the total variance calculated as:

$$\frac{\sum_{\text{all groups}} \left[\sum x^2 - \frac{(\sum x)^2}{N}\right]}{\text{Total observations}}$$

6. Choose a suitable sub-group of the data as the current group, and repeat the above steps until either the data is split into groups of size less than a minimum number, or the reduction in variance obtained by a split of the data is below a minimum value.

Regression

1. For each pair of adjoining subgroups, calculate the change in variance resulting from the amalgamation of these two groups.

2. Combine the sub-groups whose combination will result in the least gain of variance.

3. Repeat the above steps until the desired numbers of sub-groups are obtained.

Computational time issues

With uncensored data numerated covariates, for example patient age, these are ordered such that the analysis is carried out on ($< x$, εx) and is performed n times, where n is the number of individual ages. We have O(n) computations to make. Enumerated covariates, A, B, C and D with X = {A, B, C, D}, are ordered such that E(A) δ E(B) δ E(C) δ E(D). Therefore it is logical to look at the groupings {A}, {B, C, D}; {A, B}, {C, D}; {A, B, C}, {D} etc., with order O(n-1)

With censored or categorical data numerated covariates, analysis is carried out with order O(n). However if we are analysing an enumerated split, for example hospital specialties A, B, C and D with X={A, B, C, D}, we

have to investigate all possible groupings: {A}, {B,C,D}; {A,B}, {C,D}; {A,B,C},{D}; {B},{A,C,D}; {B,C}, {A,D}; {C}, {A,B,D}; {A,D}, {B,C} with order $O(2^{n-1} - 1)$ for nominal and $O(n^2 - 1)$ for ordinal variables. The computation time for the construction of the tree sequence for censored data is of obvious concern, although a possible solution for large number of enumerated types would be the use of factorial designed experiments.

THE PORT STATISTICAL PACKAGE

A statistical package, PORT, has been developed. PORT incorporates a tree-based algorithm, similar to CART, that assists in the production of clinically and statistically meaningful healthcare groupings. For example, these could be patient groupings based on cost of injury, time off work, or outcomes.

PORT has been designed to enable AUVA personnel to create appropriate groupings of cases and carry out the necessary statistical analysis. At the highest level of functionality, PORT may be used as a:

- Data exploratory tool, allowing the user to explore and understand in greater detail their data. For example, PORT permits manual splitting of the data into desired groupings and the rapid extraction of a number of key statistics, time-dependent profiles and continuous distribution fitting.
- Tree-based algorithm tool for classification and prediction, allowing the user to derive statistically meaningful, easy to interpret homogeneous groupings. This aids understanding of the structure of the data, enabling the user to define interpretable classification rules as necessary.

In the context of AUVA, likely patient groups would be based on cost per case as the dependent variable. Independent variables could include any routinely collected, or other, data such as age, sex, nationality, MDE, Beza etc.

A high-level appreciation of the functionality of PORT is given in Figure 2.

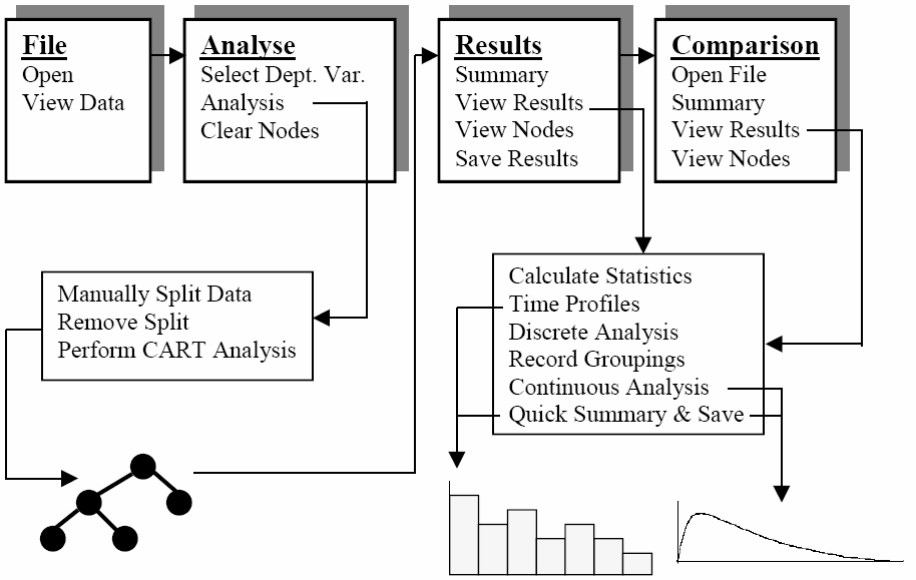

Figure 2: High-level PORT functionality diagram

The subsequent sections of this chapter illustrate different aspects of this functionality through various screen-shots and discussions.

Loading, viewing and constructing a tree

Data files may be loaded in to PORT from a current choice of .dbf or Excel formats. Figure 3 shows the main screen which forms the central control of the program. Through this menu the user can load and save data, view the data table and resulting tree and perform the necessary splits to construct a tree.

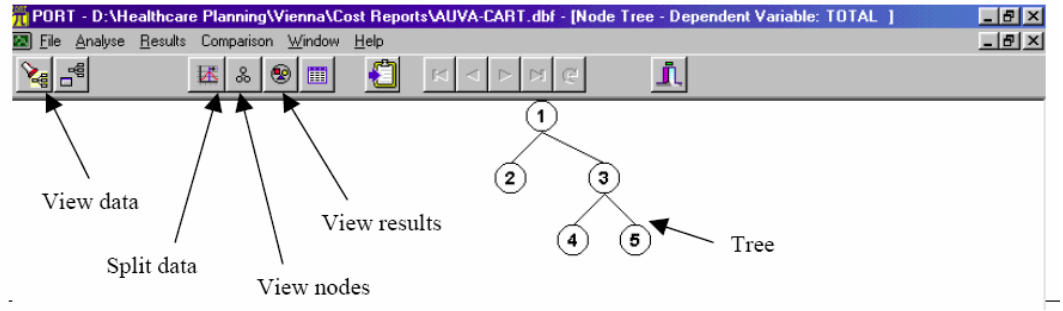

Figure 3: PORT main screen

Having chosen the dependent variable from those available in the data table, the user can commence the splitting of the data (binary splits) adopting either a manual or CART approach, or a combination of both from the list of independent variables (Figure 4). For example, splitting the whole dataset (node 1) into two subgroups (nodes 2 and 3) representing those less than 50 years of age and those aged 50 or over.

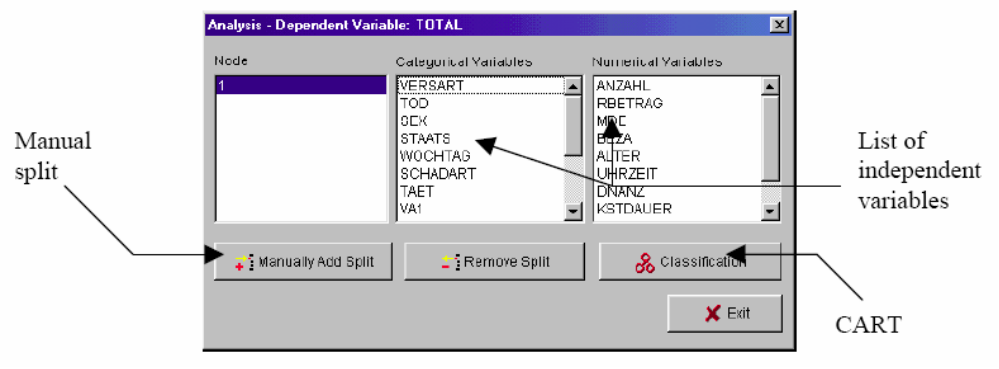

Figure 4: Splitting the data to construct a tree

CART analysis

If the tree-based algorithm is selected, the user will need to provide the necessary information in order to construct a tree. The dependent variable has already been selected. Independent variables to use in the classification algorithm must be chosen from the list of those available (e.g. sex, age, profession). These may be nominal (categorical) or scale (continuous) variables. Additional information required includes the desired number of final groups and the minimum number of patients in each group.

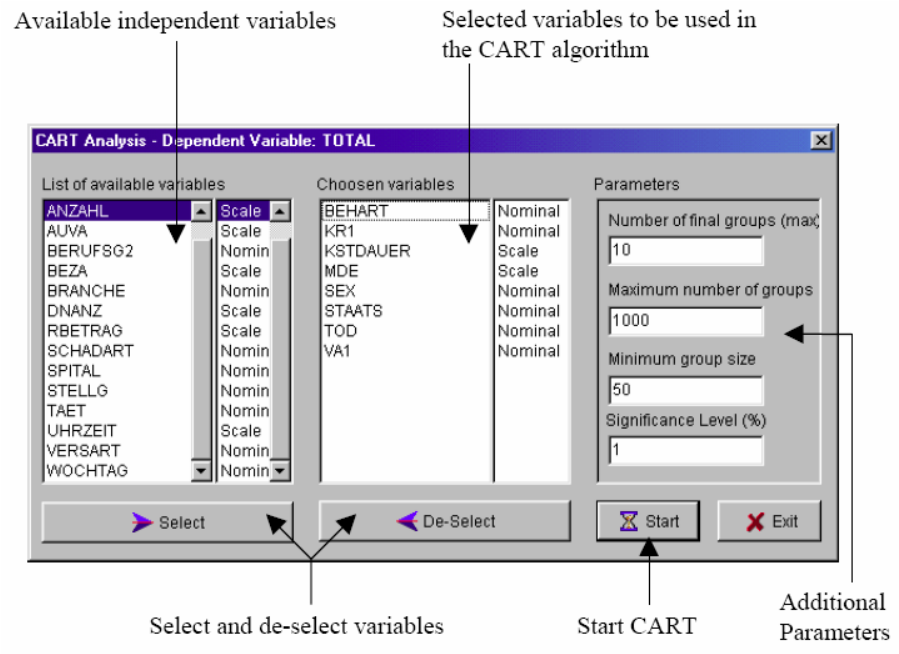

Figure 5: CART parameters

Viewing the results

A number of statistical indicators are available for each node within the tree. These include group mean, variance and inter-quartile range (for a continuous dependent variable) or percentage split and deviance (for a categorical dependent variable), together with rapid access to arrival profiles (for month, day and hour) and distribution fitting of continuous variables. A summary of nodes form provides easy access to node results (Figure

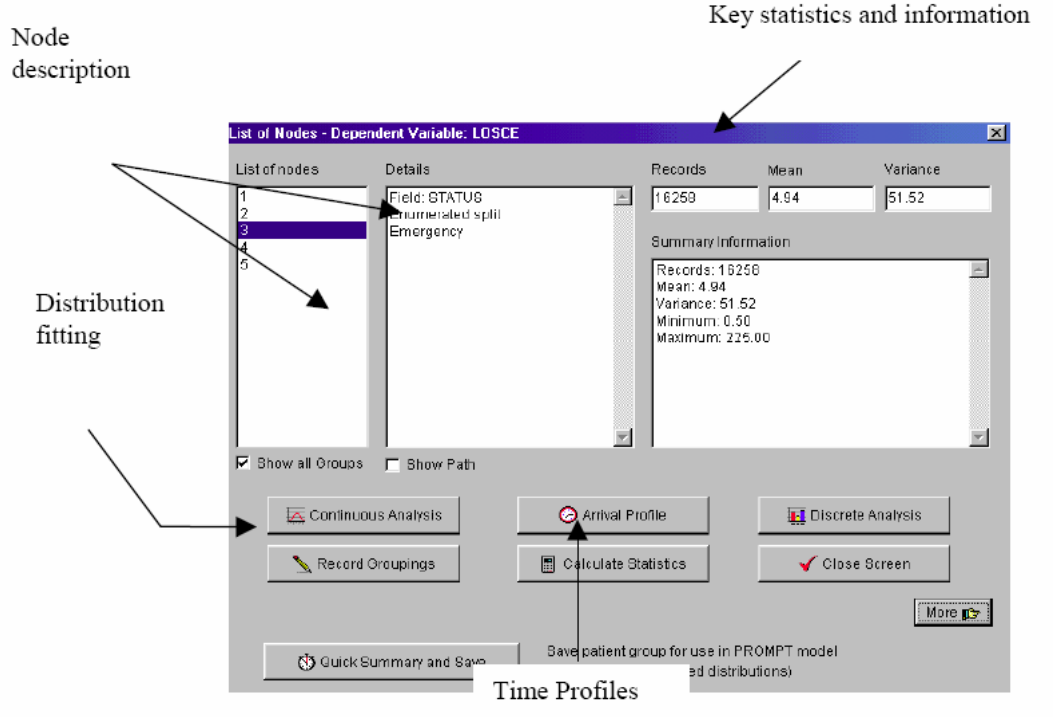

Figure 6: Summary of nodes form

The distribution and time profile buttons are particularly useful. For example, here it is possible to look at age distributions, costs distributions and MDE distributions for a particular group. The time profiles allow the user to examine when accidents occur (Figure 7).

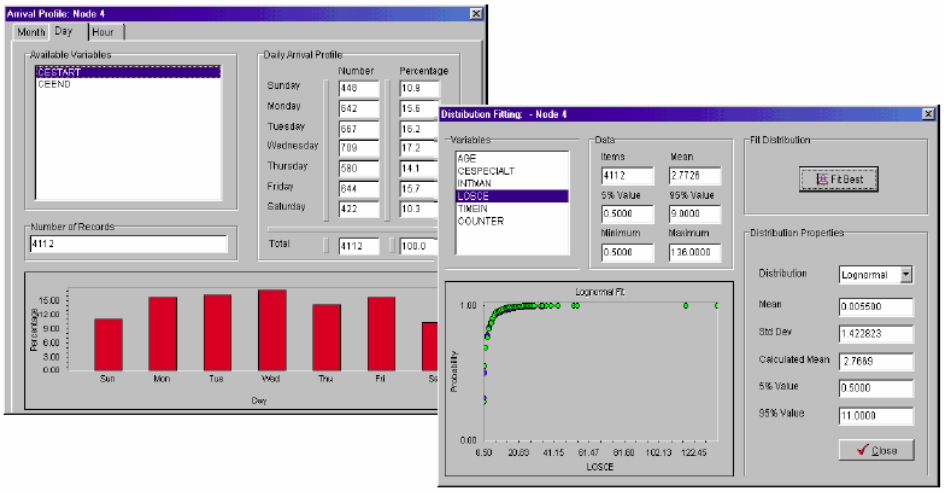

Figure 7: Time profiles and distribution fitting

SUMMARY

In order to capture the uncertainty and variability amongst the population, a number of classification techniques may be considered. Research has shown that there is not necessarily a single *best* classification tool but instead the best technique will depend on the features of the dataset to be analysed. However CART has shown to be particularly robust and user-friendly.

This is a measure of the extent to which the CART algorithm produces comprehensible results that are generally easier to interpret than the results of other algorithms and on the time it takes to understand the technique, prepare the data and actually perform the analysis to produce correct and meaningful results. A statistical package, PORT, has been developed which incorporates the CART tree-based algorithm that assists in the production of clinically and statistically meaningful groupings. For example, these could be groupings based on lifetime costs, risks of injury or outcome (survived injury or died).

REFERENCES

Breiman, L *et al*. (1984), *Classification and Regression Trees.* Chapman & Hall, London.

Clark, P and Boswell, R (1991), "Rule induction with CN2: Some recent improvements". In *Proceedings of ESWL'91* (Porto, Portugal) 151-163.

Garbe, C *et al*. (1995), "Primary Cutaneous Melanoma – Identification of prognostic groups and estimation of individual prognosis for 5,093 patients". *Cancer*. 75: 2484-2491.

Harper, P R (2002), *Operational Modelling for the Planning and Management of Healthcare Resources*. PhD Thesis, University of Southampton, UK.

Hornberger, J C, Habraken, H and Bloch, D A (1995), "Minimum data needed on patient preferences for acute, efficient medical decision-making". *Medical Care*. 33: 297-310.

King, R D, Feng, C and Sutherland, A (1995), "Statlog: Comparison of classification algorithms on large real-world problems". *Applied Artificial Intelligence*. 9: 289-333.

Remme, J, Habbema, J D F and Hermans, J (1980), "A simulative comparison of linear, quadratic and kernel discrimination". *Journal of Statistics and Computer Simulation*. 11: 87-106.

Ridley, S *et al*. (1998), "Classification trees: a possible method for iso-resource grouping in intensive care". *Anaesthesia*. 53: 833-840.

Smith, M E *et al*. (1992), "Case-mix groups for hospital-based home care". *Medical Care*. 30: 1-16.

Xu, L, Krzyzak, A and Oja, E (1991), "Neural nets for dual subspace pattern recognition method". *International Journal of Neural Systems*. 2: 169-184.