[6] B.H. Erickson and T.A. Nosanchuk. *Understanding Data.* The Open University Press, 1979.

[7] B. S. Everitt. *Cluster Analysis.* Heinemann, London, 2nd edition, 1980.

[8] B. S. Everitt and G. Dunn. *Applied Multivariate Statistical Analysis.* Edward Arnold, London, 1991.

[9] U. M. Fayyad, P. Piatetsky-Shapiro, and P. Smyth. From Data Mining to Knowledge Discovery. In *Advances in Knowledge Discovery and Data Mining*, pages 1–34. The MIT Press, Cambridge, Mass., 1996.

[10] J. Fox. *Linear Statistical Models and Related Methods.* John Wiley & Sons, New York, 1984.

[11] W. J. Frawley, G. Piatetsky-Shapiro, and C. J. Matheus. Knowledge discovery in databases: An overview. In G. Piatetsky-Shapiro and W. Frawley, editors, *Knowledge Discovery in Databases*, pages 1–27. AAAI Press / MIT Press, Menlo Park, CA., / Cambridge, MA., 1991.

[12] D. E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning.* Addison-Wesley, Reading, Mass., 1989.

[13] J. Healey. *Statistics: A Tool For Social Research.* Wadsworth, Belmont, CA., 1990.

[14] K. M. Ho and P. D. Scott. Discretization of continuous variables in bivariate relationships. Technical Report CSM-287, Dept. of Computer Science, University of Essex, Colchester, UK, February 1997.

[15] J. H. Holland. *Adaptation in Natural and Artificial Systems.* University of Michigan Press, Ann Arbor, 1975.

[16] C. Marsh. *Exploring Data: An Introduction to Data Analysis for Social Scientists.* Polity Press, Cambridge, UK, 1988.

[17] C. A. O'Muircheartaigh and C. Payne. *The Analysis of Survey Data. Volume 1: Exploring Data Structures.* John Wiley & Sons, New York, 1977.

[18] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.

[19] J. R. Quinlan. *Programs for Machine Learning.* Morgan Kaufman Publ. Inc., Los Altos, CA, 1993.

[20] P. D. Scott, M. H. Hobbs, R. J. Williams, and A. P. M. Coxon. Exploratory analysis using a genetic algorithm for multiple regression. Technical Report CSM-288, Dept. of Computer Science, University of Essex, Colchester, UK, February 1997.

[21] P. D. Scott, R. J. Williams, and K. M. Ho. Forming categories in exploratory data analysis and data mining. Technical Report CSM-285, Dept. of Computer Science, University of Essex, Colchester, UK, February 1997.

[22] J. W. Tukey. *Exploratory Data Analysis.* Addison-Wesley, Reading, Mass., 1977.

of initial data examination procedures listed in Section 3 are conventional EDA techniques: SNOUT's contribution is to automate them and apply them en masse to build an initial picture of the data set rapidly.

8.1 Further Development of SNOUT

We still consider that facilities to allow the use of domain knowledge to constrain the search will prove to be of great value in automating exploratory data analysis. Both the decision tree and genetic algorithm procedures described above enable SNOUT to build multivariate models of a domain represented by a data set. In each case the ultimate criterion of model quality is some form of goodness of fit. This is sufficient so long as nothing more than predictive accuracy is required. However it is very frequently the case that an investigator will wish to go beyond prediction and obtain some form of explanation. This distinction is important. Ownership of a large yacht may be very good predictor of high income but it is much more likely to be its consequence rather than its cause.

Finding models that have some explanatory substance requires knowledge of the domain that is not actually contained within the data. If an intelligent assistant such as SNOUT is to produce such knowledge it requires two features: a means of acquiring such knowledge from the user and a means of using that knowledge to constrain a search for good models. As a first step towards such a facility we have developed a program, based on causal order logic [3] that accepts user input specifying temporal relationships between events or attributes represented by variables in the data set and uses these to determine which variables could sensibly be used to build models of a specified variable. This module has been successfully tested but not yet integrated into SNOUT.

A second major enhancement, which is also currently under way, is the introduction of a script language allowing strategies for exploratory analysis to be specified as sets of production rules. At present this is confined to automating the initial analysis and making further use of the results of bivariate analysis. For example, there is a script that, given a variable A, will search for all pairs of variables B and C such that A is not related to B but both A and B are related to C. Such results are of interest because they provide evidence that C may be of explanatory importance. Ultimately we anticipate that the script language will be used to embody much of the domain knowledge.

Acknowledgments

Work on the development of SNOUT has been funded by the Economic and Social Research Council's programme on the Analysis of Large and Complex Datasets under grant number H519255030.

References

[1] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, Pacific Grove, CA., 1984.

[2] J. A. Davis. *Elementary Survey Analysis*. Prentice-Hall, Englewood Cliffs, New Jersey, 1971.

[3] J. A. Davis. *The Logic of Causal Order*. Sage Publications, Newbury Park, CA, 1985.

[4] L. Davis. A Genetic Algorithms Tutorial. In L. Davis, editor, *Handbook of Genetic Algorithms*, pages 1–101. Van Nostrand Reinhold, New York, 1991.

[5] J. Dougherty, R. Kohavi, and M. Sahami. Supervised and unsupervised discretisation of continuous features. In *Proc. Twelfth International Conference on Machine Learning*, Los Altos, CA, 1995. Morgan Kaufman Publ. Inc.

7 Model Selection Using the Genetic Algorithm

In developing a model of some phenomena using statistical techniques the investigator is frequently faced with the need to choose between a large number of models. For example, suppose it is desired to construct a regression model of some dependent variable using a data set that includes 100 other variables. If the objective is a reasonably simple equation, that is, one with only a small number of terms, the investigator must select a small number of variables to include in the model. This is the problem of *model selection*. It is potentially difficult because the selection must be made from a huge number of alternatives: one member of the powerset of candidate variables must be chosen.

SNOUT includes a module that enables the investigator to use a genetic algorithm ([15], [12], [4]) for model selection. It would appear to be any ideal task for genetic algorithms which are particularly good at searching large spaces for groups of elements that together contribute substantially to the fitness.

SNOUT's genetic algorithm is used primarily to search for multiple regression models. Chromosomes represent sets of variables. The fitness metric has two components: a regression is carried out using the variable set specified in the chromosome and the resulting value of R^2 used as a measure of goodness of fit. This is divided by the number of variables in the set in order to bias the system in favour of simpler models.

This technique for finding regression models has been tested extensively using real and artificial data. It is effective in that almost always finds the best model and errors are always either one missing or one superfluous variable.

Statistical packages often supply one or more stepwise procedures for building regression models. Comparisons between SNOUT's genetic algorithm technique and the stepwise methods available in SPSS showed that the latter are faster because they require fewer regressions to be computed. However it is known that stepwise methods do not necessarily find optimal models [10] because they are essentially hill climbing methods.

A more complete account of the use of genetic algorithms in SNOUT appears in Scott, Hobbs, Williams, and Coxon [20].

8 Discussion

When we begin work on SNOUT we anticipated that most of the innovatory features would concern the use of domain knowledge to constrain the search for relationships. We were therefore surprised to discover that there was much scope for new techniques to be employed in the early stages of exploratory analysis.

The two most important innovations are the category formation techniques discussed in Section 4 and the variable grouping procedure described in Section 5. Category formation is a fundamental aspect of knowledge discovery and there is much scope for further work on this topic. The dichotomisation procedure has been shown to be an effective tool but many improvements are possible: most obviously extending it to allow more than two partitions. The procedure for finding groups of related variables has proved to be an extremely fast and effective way of uncovering the basic structure of a complex data set.

The applications of decision tree techniques (Section 6) and genetic algorithms (Section 7) have much in common. Both provide the investigator with means to explore multivariate relationships and both are applications of well established machine learning techniques. The decision tree module is unorthodox because the investigator may collaborate in constructing the tree, while the genetic algorithm offers an alternative to established techniques for finding multiple regression models.

The development of heuristic methods for variable typing (Section 2) was only necessary because it has not become established practice to include this information as part of a data set. The battery

more children. Not surprisingly responses to these questions are highly correlated. Consequently the variable similarity tree identifies them as a distinct cluster. Similarly a group of variables related to political affiliation form another cluster. However the tree also shows that these two groups are not strongly related to each other, since their subtrees are widely separated.

A hypothetical investigator interested in accounting for attitudes to abortion could immediately draw two conclusions. First, attitudes to abortion in various circumstances are strongly related: hence it is likely that any factor that contributes to opinions about one circumstance contributes to opinions about the others. Second, there is little point in looking for a relationship between abortion attitudes and political beliefs because any relationship that may exist is very weak. Further examination of the variable similarity tree reveals that a third cluster of variables pertaining to religious belief and behaviour is a close neighbour of the abortion attitude group. Hence the investigator can immediately conclude that a more detailed examination of the relationships between religious belief and attitudes to abortion would be very worthwhile.

6 Interactive Decision Tree Construction

Decision tree construction ([1],[18],[19]) is probably the most widely used data mining technique. SNOUT provides the investigator with a module for building such trees that is based on ID3 [18] but it is intended to be used in somewhat different way.

Decision trees are typically constructed because the investigator wishes to develop an accurate and reliable way of predicting the values of some dependent variable: for example, whether a loan applicant is likely to default. In contrast SNOUT is intended as a tool to aid investigators to come to an understanding of the relationships within the data set.

SNOUT therefore has a facility that allows the investigator to collaborate in the construction of the decision tree. In a normal decision tree program the selection of the attribute to be used to construct a subtree is made by the system on the basis of some appropriate criterion such as maximal information gain. SNOUT's decision tree module can also be used in this way. Alternatively, the investigator may specify which variables are to be used to construct the initial stages of the tree.

This has proved to be a very useful facility as the following example illustrates. Suppose the hypothetical investigator of abortion attitudes discussed in Section 5 wanted to know how men and womens' attitude to abortion on demand varied. By specifying that the first attribute chosen should be gender, the investigator is effectively getting two distinct subtrees, one for men and one for women. If men and women's attitudes to abortion are influenced by similar factors, each subtree will be built using the same set of variables; conversely if they have different origins then different sets of variables are likely to appear in the respective subtrees.

The investigator may also specify that certain variables should not be used in tree construction. This is important when prior knowledge or the variable similarity tree (see Section 5) suggests several variables may be measuring essentially the same thing. A decision tree to predict attitudes to abortion on demand that allowed all the variables to be used would be dominated by the other abortion attitude variables. Such a result is unsurprising and ultimately uninformative and uninteresting. Much more insight into the factors associated with beliefs about abortion is obtained by building a decision tree that excludes them. Alternatively the investigator may decide that all variables measuring attitudes should be excluded so that the resulting tree will only containing variables measuring tangible socio-economic attributes. This can readily be accomplished since the attitudinal variables will have already been identified at an earlier stage (see Section 2).

It should be clear that this simple approach will produce fairly uneven results: some of the dichotomies formed will appear reasonable while others will be rather arbitrary. In the exogenous phase SNOUT therefore proceeds to use them as a starting point for an improved set of dichotomies derived by maximising both the number of variables associated with each variable and the strength of those associations. The investigator may specify Yule's Q [2] or Cramer's V [13] as the measure of association. A detailed description of the procedure used in the exogenous phase is given in [21].

The following illustrates the type of result these dichotomisation techniques can achieve. SNOUT dichotomised all the variables in a set of about 100 taken from the 1991 British Household Panel Survey (BHPS). One of the questions was "Which daily newspaper do you purchase or read first?". SNOUT partitioned the response to this as follows:

Group A: Mirror, Star, Sun

Group B: Express, Financial Times, Guardian, Independent, Mail, Telegraph, Times, Today

Anyone familiar with the British press will immediately recognise that SNOUT has identified the most fundamental distinction: Group A contains the tabloid papers while Group B contains the broadsheets plus the "middlebrow" Express and Mail. Note that this dichotomy is based entirely on the other responses of the readers: SNOUT has no other information concerning the properties of these newspapers. Note also that the resulting groups are far from equal in size: those in Group A have much larger circulations than those in Group B.

A more complete discussion of category formation in SNOUT appears in Scott, Williams, and Ho [21].

5 Identifying Variable Groups

Once the endogenous dichotomisation phase is complete, the investigator will typically wish to know how particular dichotomies were derived. SNOUT can supply a list of all those variables that contributed to the partitioning of a specified variable. For example, SNOUT partitioned the regions of the UK in the 1991 BHPS data into two groups that corresponded roughly with a South/North divide. Subsequent listing of associated variables revealed that Northerners are more likely to smoke, read tabloid newspapers and live in council housing.

This notion of exploring related variables has been generalized in SNOUT to automate the discovery of sets of related variables. The method used is similar to agglomerative hierarchical techniques for cluster analysis ([7], [8]) but is applied to the variables themselves rather than the the items in the data set. Thus where cluster analysis attempts to identify groups of similar individuals, SNOUT is attempting to find groups of related variables.

During dichotomisation the system will have computed the association between every pair of variables using Yule's Q or Cramer's V (see section 4). These are used as similarity measures to construct a *variable similarity tree* in a bottom-up fashion. Nodes of the tree represent groups of variables; the leaf nodes represent a group of only one variable. The similarity of two nodes is defined as the mean similarity of all variables in the first group to all those in the second. Tree building proceeds by repeatedly joining the most similar pair of groups to form a new larger group.

SNOUT displays the resulting tree graphically, thus rapidly providing the investigator with an overall picture of the relationships present in the data. Closely related variables are clustered on subtrees, so it is immediately apparent which relationships are worth further investigation.

This technique was used to construct the variable similarity tree of data set taken from the 1990 US General Social Survey. It comprised 38 variables, the majority of which were attitudinal indicating the respondents political and ethical beliefs. Seven of the questions asked whether abortion was acceptable under various circumstances ranging from rape to a simple wish not to have

4 Category Formation

One of the most common techniques that people use to deal with the enormous variety and complexity of their experiences is to form equivalence classes or categories. Partitioning the set of values taken by a variable in a data set is often a very useful step for at least four distinct reasons:

1. *To enable an analytic procedure that works with discrete values to be employed using a variable whose values are continuous.*

   For example, in recent years several research groups have developed techniques for discretisation of continuous variables before or during the construction of decision trees ([1],[18]). Dougherty, Kohavi & Sahami [5] have produced a valuable comparative study of these techniques.

   As part of the SNOUT project a new method of partitioning numeric variables, called *zeta discretisation*, has been developed based on maximising the prediction accuracy for a categorical variable. Initial results suggest that this avoids the accuracy/time trade-off apparent in Dougherty *et al's* [5] comparisons. An account of this method appears in [14].

2. *To enable the investigator to develop a crude picture of the major relationships in a data set with relatively little effort.*

   This is exemplified by Davis's [2] approach to exploratory analysis of survey data in which all variables are dichotomised and Yule's Q used as a common measure of association between all pairs of variables.

   SNOUT incorporates a dichotomisation module (described below) that takes Davis's methods as a starting point.

3. *As an end in itself because the discovery of a 'good' partitioning tells the investigator something about the structure of the world that the variables represent.*

   For example noticing that the heavenly bodies fell into two groups, those that remained fixed relative to the others and those that wandered about, was an important step towards discovering the structure of the solar system and its relationship to the stars.

4. *As an expedient for dealing with data shortage.*

   If there are very few instances of some of the values of a nominal variable it may not be possible to apply particular analytic procedures. Grouping the values into a smaller number of categories is an obvious solution.

4.1 Variable Dichotomisation

SNOUT attempts to dichotomise all the variables in a data set using a two phase procedure. During the first *endogenous* phase each variable is partioned using only information concerning the distribution of values of that variable. During the second *exogenous* phase, each variable is partitioned so as to maximize its association with all other variables in their dichotomised forms.

The endogenous phase uses simple but crude techniques, derived from techniques suggested by Davis [2] for manual data exploration, to form initial dichotomies. Interval variables are divided at the median while ordinal variables are divided at the point which minimizes the difference in size between the two resulting components. If the largest category of a nominal variable exceeds 35% of the total then it forms one component and the remaining categories are grouped as the other; if not then no attempt is made to dichotomise the variable during this phase.

It is often useful to distinguish variables representing subjects opinions or attitudes from those representing more tangible attributes such as employment, housing, and marital status. SNOUT identifies such opinion variables using a keyword technique similar to that used to detect ordinal variables. Indicative words for attitudinal variables include 'favour', 'oppose', 'agree', and 'important'.

These heuristic classification techniques constitute a rudimentary attempt to discover some of the semantics of the data set. The keyword approach appears to work well and could almost certainly be extended to identify other interesting classes of variable.

3 Initial Data Examination

Once the variables have been classified subsequent steps of exploration and analysis are selected by the investigator from a menu of options.

3.1 Univariate Analysis

Given a new data set an investigator's first step is usually to get some understanding of the distribution of the individual variables before attempting to discover any relationships between them. SNOUT calculates a wide range of descriptive statistics for each variable in the data set. The particular statistics computed depend on the level of the variable and include mean, median, standard deviation, maximum, minimum, range, quartiles, deciles, number of distinct values and number of missing cases.

The investigator is presented with all this information in tabular form. In addition a wide variety of graphical displays are also available including frequency histograms, pie charts, box plots and stem and leaf displays ([22],[6],[16]). The box plot facility also incorporates a simple tool for inspecting bivariate relationships: separate box plots of an interval variable for each value of a nominal or ordinal may be displayed side by side.

3.2 Initial Bivariate Analysis

The next step in the exploration will depend on the investigator's objectives. If the investigator has an interest in some particular variable and wishes to identify which other variables are related to it, SNOUT can be asked to search for all such relationships. For each of the other variables tests are carried out to assess the strength and statistical significance of the relationship with the dependent variable. The choice of tests depends on the variable levels: if both are nominals chi-squared and Cramer's V [13] are used; if both are intervals correlations is calculated; otherwise dummy regression is used[2]. The results of all tests are listed: statistically significant relationships are highlighted and categorised according to the strength of the relationship. The use of different measures for different levels of variable is unavoidable but does have the disadvantage that it is difficult to compare the relative strengths. The dichotomisation of all variables (see Section 4.1) allows a uniform metric to be applied.

At this point the investigator will have invested very little time in interacting with the systems and should already have a clear idea of how the variables are distributed and know the strength and direction of all the significant relationships involving the specified dependent variable. If the aim of the investigation is a predictive model of the dependent variable the next step would be to use this bivariate information to develop a multivariate model.

[2]In the current version of SNOUT ordinals are treated as nominals for the initial bivariate analysis, but may be coerced to interval variables by the user.

The most obvious source of information on the level of the variable is the representation used to store it in the data file. This will normally be either a string or a number. A string may denote either a nominal variable or a free text response. SNOUT uses both the length of the string and the number of distinct values to distinguish these cases: only those whose values are less than 9 characters and that have fewer than 25 distinct values are classed as nominal variables; the rest are excluded from further analysis.

Variables represented by numbers may be nominal, ordinal or interval [1] because numbers provide a convenient and compact coding for all sorts of information. If some of the values of the variable in the data file are non-integral SNOUT concludes that it must be at the interval level. However, if all the values are integers then it could denote an interval variable such as age, an ordinal variable such as degree of support or opposition to a particular policy, or a nominal variable such as favourite newspaper.

Various heuristic techniques are therefore used to determine the most probable level of such a variable. The first depends on the number of distinct values. Interval variables are likely to have many distinct values; nominals and ordinals rather fewer. Hence if the number of values exceeds 25, SNOUT concludes that it is very likely to be an interval variable. A variable with less than three distinct values is treated as a nominal.

If there are between 3 and 25 distinct values more information is needed to distinguish ordinals and nominals. Fortunately data sets often include text defining the responses encoded by each value: for example 0 = Male, 1 = Female; or 1 = Strongly Agree, 2 = Agree etc.. Although interpreting such text and hence deducing the variable type would require extensive linguistic and world knowledge it is possible to derive a great deal of information from such value labels using keyword techniques. The occurrence of certain words as part of a value label are strongly indicative that the variable is at the ordinal level. Examples include 'most', 'least', 'strongly', 'slightly', and 'few'. SNOUT currently employs a list of about 20 such words to identify ordinal variables.

When SNOUT has read in a new data set, all variables are scanned and assigned to a level using such heuristic techniques and ignoring codes for missing values. The investigator is then given the opportunity to scan the results and override them if necessary. The level classifications can be saved so that this procedure need only be carried out once, the first time a data set is examined using SNOUT.

Experiments using a variety of data sets have shown that SNOUT typically assigns 1-2% of variables to the wrong level: the commonest mistake being confusion of nominals and ordinals. These errors are readily detected and corrected so the net saving in the investigators time is considerable.

2.2 Heuristic Variable Classification

In addition to determining the level of each variable SNOUT also attempts to identify a number of important classes of variable. This is because knowledge of the meaning of a variable can be used to restrict the relationships that are explored in later analysis. For example, while it is quite plausible that age may have a causal influence on earnings, it is logically impossible for earnings to have a causal influence on age.

SNOUT therefore attempts to identify variables representing gender and age by considering the values labels. Any variable with less than three values whose value labels include some of the keywords 'male', 'female', 'men', and 'women' is deemed to represent gender. Value labels are less helpful in identifying interval age variables since they are often absent because the values are self explanatory. SNOUT must therefore look for the occurrence of 'age' in the variable name.

[1] No attempt is made to distinguish ratio and interval variables because the difference is of no importance for the procedures employed in SNOUT.

suggests that it should be possible to use data mining techniques to automate EDA, thus solving both of the problems arising from the limitations of conventional EDA techniques.

1.2 SNOUT: An Intelligent Assistant for EDA

We have developed a system called SNOUT (SNiffing Out Useful Things) that is intended to be used by social science researchers engaged in exploratory analysis of survey data. Such data sets are often very large; a typical example might comprise the responses of several thousand subjects to several hundred questions. The objective of investigations of this type is more likely to be an understanding of a set of interacting relationships rather than accurate prediction of a dependent variable. Hence the emphasis in developing SNOUT has been on building an interactive tool that assists the investigator to build an understanding of the structure of the data rather than a system in which the user sits back and waits for the system to produce an answer.

SNOUT runs under Microsoft Windows and is written mainly in Visual C++. Early versions used calls to SPSS for all standard statistical computations but these have been replaced in later versions so it is now a standalone program. However, some of the features currently under development are written in LPA Prolog. Upon launching the program the user is presented with a standard Windows graphical menu driven interface. Normally a user's first action will be to open a data set which must be in SPSS file format. Once this has been done some initial analysis takes place automatically: subsequent steps are selected by the investigator

In the rest of this paper we present an introduction to the distinctive features of the current version of SNOUT. Features found in other statistical or data mining systems are mentioned briefly while those that are novel are described at greater length. However, in a paper of this nature it is not possible to give a detailed account of any one aspect of the system. The interested reader is referred to existing ([14],[21],[20]) and forthcoming papers devoted to particular modules of SNOUT.

2 Heuristic Variable Level Inference and Classification

In order to decide what statistical or machine learning procedures should be applied to a set of data it is necessary to know something about the types of variables to be analysed. In particular it is almost always necessary to know their level: *nominal*, *ordinal*, *interval*, or *ratio* [2]. For example, applying multiple regression to data sets made up of nominal variables produces meaningless results. In contrast, decision tree induction algorithms [1] such as ID3 [18] can only be applied directly to nominal variables; other levels of variable must be converted to nominals by a discretisation process [5].

Unfortunately the data sets that SNOUT is intended to investigate typically contain very little information about variable types. Furthermore the values are normally stored as numbers irrespective of the actual level of the variable concerned. Consequently many statistical packages raise no object if the user instructs them to perform a multiple regression on nominal variables!

2.1 Heuristic Level Inference

SNOUT needs to know the level of each variable in a data set in order to decide what tests and measures are appropriate when investigating particular relationships. Since this information is not available in the data set two alternative sources must be used: either the investigator must supply the variable types or the system must infer them from the limited information available. The former solution is tedious in the data sets we are considering which may have several hundred variables. SNOUT therefore uses a collection of heuristic techniques to determine the type of each variable.

SNOUT: An Intelligent Assistant for Exploratory Data Analysis

P.D. Scott, A.P.M. Coxon, M.H. Hobbs, R.J. Williams
University of Essex, Colchester CO4 3SQ, UK.
scotp@essex.ac.uk

March 7, 1997

Abstract

In this paper we present an account of the main features of SNOUT, an intelligent assistant for exploratory data analysis (EDA) of social science survey data that incorporates a range of data mining techniques. EDA has much in common with existing data mining techniques: its main objective is to help an investigator reach an understanding of the important relationships in a data set rather than simply develop predictive models for selected variables. Brief descriptions of a number of novel techniques developed for use in SNOUT are presented. These include heuristic variable level inference and classification, automatic category formation, the use of similarity trees to identify groups of related variables, interactive decision tree construction, and model selection using a genetic algorithm.

1 Introduction

This paper describes the principal novel features of SNOUT a system currently under development, that uses both established and novel data mining techniques to serve as an intelligent assistant for exploratory data analysis.

1.1 Exploratory Data Analysis and Data Mining

There are two main modes of data analysis: confirmatory analysis, which an investigator uses to determine whether or not the data provide evidence for some particular hypothesis (or class of hypotheses) concerning a relationship between variables represented in the data set; and exploratory data analysis (EDA) which is used to discover regularities in the variables in the data that are of interest to the investigator [17]. In general EDA will be used in new areas of research where the aim is to find interesting patterns or structures that require explanation, thus generating hypotheses for later confirmatory study [8]. Thus the primary goal of EDA is to lead the investigator to an understanding of what relationships appear to exist within a data set.

In this early phase formal statistical methods designed to test specific relationships are not relevant: instead flexible methods for finding possibly unanticipated patterns in the data are required. Following the pioneering work of Tukey [22], a range of techniques have been developed that enable the investigator to examine sets of data with the goal of exposing regularities and patterns [6]. Many of them exploit the remarkable pattern recognition capabilities of the human visual system by presenting aspects of the data in graphical form. Although powerful, these techniques suffer from two serious limitations. First, their graphical basis limits their usefulness in discovering multivariate relationships, and second, because they require a lot of work by the investigator it is tedious and very time consuming to apply them to data sets with large numbers of variables.

Exploratory data analysis therefore has much in common with data mining ([11],[9]). The fundamental objective is clearly the same: discovery of relationships in large bodies of data. This