To appear in *Proceedings, IDA-95 Symposium of the International Conference on Systems Research, Informatics and Cybernetics.*

A Case Study of Planning for Exploratory Data Analysis

Robert St. Amant and Paul R. Cohen

Computer Science Technical Report 95-56

Experimental Knowledge Systems Laboratory
Computer Science Department, Box 34610
Lederle Graduate Research Center
University of Massachusetts
Amherst, MA 01003-4610

Abstract

To develop an initial understanding of complex data, one often begins with exploration. Exploratory data analysis (EDA) provides a set of statistical tools through which patterns in data may be extracted and examined in detail. We briefly describe an implemented planning representation for Aide, an automated EDA assistant. We discuss the analysis of a small dataset, in which exploration is driven by a set of variable selection algorithms. Our results, derived through a semi-autonomous process in which the user oversees the decisions Aide makes, are comparable with those produced by earlier researchers.

This research is supported by ARPA/Rome Laboratory under contract F30602-93-C-0100.

Exploring Data

Exploratory data analysis (EDA) transforms data to make features and patterns stand out clearly from surrounding noise. Exploratory procedures complement modeling and hypothesis testing procedures, guiding or revising the analysis in response to unexpected findings in the data. EDA enables a researcher to isolate suggestive features of data, to identify potential structure indicated by the features, and to generate plausible hypotheses to explain the structure.

A wide array of techniques have been developed to support this process. Exploratory results include histograms of discrete and continuous variables, box plots of relationships, partitions of relationships that distinguish different modes of behavior, functional simplification of low-dimensionality relationships, and two-way tables such as contingency tables. These are just a few simple examples; a sophisticated analysis combines these and many others in the construction of a global picture of the data [Tukey77, Hoaglin83].

We are developing an assistant for intelligent data exploration, AIDE, to assist human analysts with EDA [StAmant95]. AIDE takes a script-based planning approach to EDA. Data-directed mechanisms extract simple observations and suggestive indications from the data. Scripted combinations of operations are then applied in a goal-directed fashion to generate simpler, deeper, or more comprehensive descriptions of the data. Control rules guide the operations, relying on intermediate results for their decisions. The system is mixed-initiative, capable of autonomously pursuing high- and low-level goals while still allowing the user to guide or override its decisions.

Through AIDE we make two claims about data exploration. First, a mixed-initiative organization, in which the user and a semi-autonomous system share control, produces better results than either acting alone. The system is responsible for generating opportunities for exploration and following through with the analysis. The user, with external knowledge about context and the goals of the analysis, steers and critiques the process. A mixed-initiative approach to control balances the two extremes of autonomous machine learning systems and user-driven statistics systems.

Second, planning is an appropriate and powerful means of managing the automation of EDA. Exploration and planning share many characteristics: relatively few general principles guide exploratory procedures; difficult problems are decomposed into smaller or simpler parts; exploration is constructive, often relying on partial results and incremental improvement to reach solutions. Planning captures these strategic aspects of exploration. A planning framework provides a higher level of abstraction for reasoning about data manipulation, while still allowing specific decisions to take advantage of local context.

Controlling Exploration

In terms of search, exploration can generally be viewed as a local data-driven process. Specific combinations of indications, or suggestive characteristics of the data [Mosteller77], lead to appropriate actions being taken. Strong skew in a batch of data points, for example, indicates that a transformation for symmetry may be appropriate. An indication of clustering in a relationship may lead to a consideration of local behavior within each cluster. Findings can establish intermediate goals, such as trying to account for a pattern in one variable by looking for similar structure in related variables. Nevertheless the opportunistic nature of exploration requires that the process be driven by the data.

In practice, however, we often have two guides in exploring data: the user's knowledge of context, and the choice of a specific type of model. For example, suppose that our task is to describe the relationships between variables A, B, and C. We find approximately linear relationships between

A and B, B and C, that there is clustering in C, and so forth. All findings seem to be equally interesting. If we are now informed, however, that a regression model is appropriate with A as the response, then context forces us to reinterpret the results. Linearity between A and B becomes much less interesting than between B and C, if both B and C are to be included as regressors. We will need to look more closely at the latter relationship. User knowledge and modeling context often help constrain and guide exploratory search.

As a starting point we use a study due to Fowlkes, Gnanadesikan, and Kettenring [Fowlkes87], concerning techniques for describing a dataset by a selected subset of the observed variables. The authors outline a set of forward selection algorithms that apply in three different contexts: discriminant analysis, multiple regression analysis, and clustering.[1] Their work provides a useful example of the interaction between the user and a set of computer procedures for incremental model construction.

Each procedure begins with a set of variables and p, the number of variables to be included in an initial model (p starts at a value of 1.) The procedure iterates, at each step selecting a model that contains $p+1$ variables and incrementing p to the new level. The model is evaluated, taking into account the new value of p, the variables included in the model, and the evaluation of the model at the earlier stage. In the context of regression, variable subsets are evaluated by Mallows's C_p. In clustering the evaluation is more complex but essentially similar, relying on a measure of group separation, $S^d(K)$.[2] At each step a decision is made whether to continue or to halt with the model generated so far.

An example of an exploratory procedure is the straightening of a bivariate relationship, using Tukey's ladder of transformation [Tukey77]. One begins by splitting the relationship horizontally (along the x-axis) into into three partitions. Each partition is reduced to a representative $\langle x, y \rangle$ coordinate. We use the medians of each partition to determine these points. This gives a three point summary of the data. If these points fall along a straight line, then the relationship is held to be approximately linear. If the three points are not linear, then we can compute the deviation from linearity by the change in slope moving from one pair of points to the next. If the ratio of slopes is sufficiently different from 1, a power transformation is applied. A new ratio is computed, and another transformation applied if appropriate. This procedure is characteristic of exploratory procedures in general: it involves a heuristic evaluation of a feature of the data; it deals with a reduced summary rather than all of the data; and it depends on iterative improvement for a final result.

Exploration and modeling procedures are represented as plans. A plan consists of a goal form, a set of input and output variables, a set of applicability constraints, and a grammar of subgoals. Grammar constructs combine subgoals in sequence, iteratively, in parallel, and by test conditions. The plan below implements a general transformation procedure. The plan establishes subgoals to determine the direction of the transformation and the variable (or variables) to be transformed. The iteration form causes the transformation to be repeated until a test for straightness is met. The plan is activated to satisfy a subgoal established by a higher level plan, activated by an indication of curvature in the relationship. All modeling and exploration and plans are syntactically similar to this one.

---

[1] AIDE's processing does not rely only on the modeling procedures we discuss in this paper; we use the forward selection approach here as an easily grasped example.

[2] The evaluation of a cluster model is based on a candidate assignment of points into K clusters, K varying over $1, 2, \ldots, K_{max} = 8$. Each assignment is evaluated by $S^d(K) = S(K) - E(S(K))$. $S(K) = tr(T^{-1}{}_p B_p)/p$, where $B_p = T_p - W_p$, W_p and T_p being the within group and total sums of cross-product matrices respectively, tr denoting the trace statistic. $E(S(K))$, the null expected value, is estimated by simulation.

```
(define-plan* transformation-strategy
  :goal        (linearize ?structure ?directives ?context ?result)
  :variables   (structure directives context result)
  :features    ((:data-type relationship)
                (:cardinality 2))
  :internal    (transformed-relationship direction iteration-record)
  :grammar     (:SEQUENCE
                select-transformation-variable-subgoal
                extract-bindings-subgoal
                (:ITERATE (call-iteration-test ?iteration-record)
                          step-transformation-subgoal)
                finalize-transformation-subgoal))
```

A planning representation has a number of advantages over imperative or rule-based implementations. The grammar of a plan makes sequences of decisions explicit: a plan can enforce a sequence that always checks the residuals after fitting a line to a relationship, for example. Moreover, incremental plan expansion gives a natural hierarchical breakdown of the process into different levels of abstraction. Plans also provide explicit processing context for interpreting results. Finally, because plans are matched with goals instead of called directly, the planner can delay decisions until sufficient information is available to make them.

Data Analysis

The dataset to be explored contains nine variables that describe 56 automobiles. The variables are make/model, price, mileage, rear seat space, trunk space, weight, length, turning radius, engine displacement, and gear ratio. The results produced by AIDE do not greatly differ from those given by Fowlkes et al. The significance lies rather in the cooperative decision-making process in which AIDE and the user participate over the course of the analysis. Because of space constraints, we can only sketch a description of the entire analysis.

AIDE begins by collecting information about each variable. This includes generation of statistical summaries, testing of data types, testing of whether the data might be considered continuous or discrete, generation of hierarchical clusters for numerical data, and so forth. Indications are then generated. An indication of skew is computed, for example, from the midsummaries of a variable. An indication of clustering tests for outliers in the distance array associated with a hierarchical clustering. AIDE reports these observations:

Variable	Indications
Price	Five high outliers (Audi, BMW, Datsun 810, Peugeot, Volvo.)
Mileage	Right skewed. Four high outliers (Datsun 210, Subaru, VW Diesel)
Seat	Mean = 26 inches; standard deviation = 3 (VW Dasher at 37.5.)
Weight	Partitionable into two classes.
Turn	Five distinct values: might treat as discrete.
Gear	Continuous, but three recurring values (3.08, 2.73, 2.93.)
	Additional indication of clustering into three groups.

While AIDE can run autonomously, a mixed-initiative modeling task gives a better example of the system's use. We begin by directing AIDE to perform a forward selection clustering. The procedure involves generating a set of possible models, evaluating each model, and deciding on one to pursue. AIDE produces a set of candidate cluster assignments for each variable. The variable

gear, for example, has four distinct assignments that break the data into 2, 3, 4, or 5 clusters. In one case, price, AIDE is unable to make any cluster assignments because of a threshold on minimum cluster size.

Within the forward selection plan, each of these possibilities is considered a potential successor to the null model starting state. The plan establishes goals to evaluate the successors. The default evaluation in this case is a simple computation of the evaluation heuristic $S^d(K)$. Evaluation is not limited to a single numerical summary, however; another plan sensitive to the evaluation context examines each variable in more detail. The initial shallow exploration produced indications of clustering in weight and showed that gear might plausibly be partitioned into recurring values and non-recurring values. During this stage all relevant information is associated with each successor model.

The plan next establishes the goal of selecting the successor model from the alternatives. In the absence of independent indications of clustering, the system selects the model with the highest relative value of $S^d(K)$. Across all models, the median value is 0.04; there are two outliers, associated with displacement (0.27 for two clusters) and gear (0.18 for two clusters). In this situation, however, there are other possibilities raised by the initial exploration: gear with two, three, or four clusters, and weight with two clusters. Though AIDE has no formal way to combine evidence here, the system can take two courses: to prune the possible alternatives initially, and to save the remaining possibilities to pursue later. A heuristic contingency table test shows in this case that the clustering of weight is almost identical to that of displacement, which means that both need not be considered. AIDE presents the relevant information graphically by plotting the clusters, and makes its decision for displacement. Goals of exploring the other possibilities are suspended.

Examining the plots, we might decide that weight, or possibly gear, gives a better separation of the data. We can take this course of action: we pause AIDE's processing, return to the most recent decision point, and examine the set of alternatives AIDE considered in selecting the initial model. This is made possible by the plan representation, which gives an explicit structured record of the analysis. We halt the current plan of exploring the model ⟨displacement⟩, and direct AIDE to return to the suspended model ⟨weight⟩. AIDE then resumes with an exploration of the new model.

In the next stage all two variable models that include weight are generated. The analysis follows a similar path from this point, computing $S^d(K)$ for each model and taking into account the indications produced by independent exploration. Eventually the model includes three clusters using just the variables weight and price. The high price outliers produced by the original exploration end up treated as a cluster, though there is little cohesion to the group. Exploration leads to classification of two points, Opel and Chevette, as outliers in the weight/displacement relationship, and the VW Dasher as an outlier with respect to rear seat space. We leave the modeling process now to examine part of the exploratory process.

The decomposition of a relationship into partitions triggers exploration with the goal of describing the behavior in each partition. One of the plans to explore relationships can be described loosely in this way: "If the data can be naturally partitioned, see if the behavior is different in each partition." A closer examination of data within partitions is comparable to examining residuals of a fit; the notion, basic to exploration, is that it is always worthwhile to look for structure beneath the surface. In the case of the relationship ⟨weight, displacement⟩, AIDE has already established that there are similar clustering effects in each variable; a similar clustering indication is present in the combination of the variables. A plan for differentiation compares the partitions of ⟨weight, displacement⟩ in terms of the indications present and the type of descriptions that the deeper exploration has produced. In this case the plan finds a high correlation—0.82, above a heuristic threshold—between weight and displacement in one partition, while in the other partition the correlation is below 0.1. This indication acts as evidence that the clustering is meaningful.

Further exploration contributes to the results described above. This brief example should illustrate two important considerations in building a system for EDA: the uses to which modeling context and user knowledge can be put, and how structure can be imposed on the analysis. The example demonstrates a useful cooperation between a partially autonomous system and a human user, in which the strengths of each—computational power on one hand, contextual knowledge on the other—are put to good use.

Acknowledgments

Thanks to an anonymous reviewer for comments about the direction this work should take, and to Dave Hart for a close reading of this paper. This work is supported by ARPA/Rome Laboratory under contract F30602-93-C-0010. The U.S. government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation hereon.

References

[Fowlkes87] Edward B. Fowlkes, Ramanathan Gnanadesikan, and Jon R. Kettenring. Variable selection in clustering and other contexts. In C. L. Mallows, editor, *Design, Data, and Analysis: by some friends of Cuthbert Daniel.* Wiley, 1987.

[Hoaglin83] David C. Hoaglin, Frederick Mosteller, and John W. Tukey. *Understanding robust and exploratory data analysis.* Wiley, 1983.

[Mosteller77] Frederick Mosteller and John W. Tukey. *Data Analysis and Regression.* Addison-Wesley, 1977.

[StAmant95] Robert St. Amant and Paul R. Cohen. Preliminary system design for an EDA assistant. In *Preliminary Papers of the Fifth International Workshop on Artificial Intelligence and Statistics*, 1995.[3]

[Tukey77] John W. Tukey. *Exploratory Data Analysis.* Addison-Wesley, 1977.

[3] Available at http://eksl-www.cs.umass.edu/eksl.html