Data Exploration Systems for Databases

Richard J. Greene and Christopher Hield
Environmental Assessment and Information Sciences Division
Argonne National Laboratory
9700 South Cass Avenue * EID/900
Argonne, Illinois 60439-4832

ABSTRACT

Data exploration systems apply machine learning techniques, multivariate statistical methods, information theory, and database theory to databases to identify significant relationships among the data and summarize information. The result of applying data exploration systems should be a better understanding of the structure of the data and a perspective of the data enabling an analyst to form hypotheses for interpreting the data. This paper argues that data exploration systems need a minimum amount of domain knowledge to guide both the statistical strategy and the interpretation of the resulting patterns discovered by these systems.

1. INTRODUCTION

Data exploration systems apply machine learning techniques, multivariate statistical methods, information theory, and database theory to databases to identify significant relationships among the data and summarize information. The result of applying data exploration systems should be a better understanding of the structure of the data and a perspective of the data enabling an analyst to form hypotheses for interpreting the data. In a sense, data exploration systems are a tool of the "scientific method": raw data is collected, laws describing the principal features of the data are hypothesized and tested, and theories explaining the laws are hypothesized and tested by using the theory to predict new information.

The benefits of data exploration systems can be significant. An analyst deluged with data can greatly reduce the time needed to understand the meaning of the data, and the accuracy of data analysis can be greatly increased. One might think of a data exploration system as a tool for examining large and complex data for meaning that might normally go unnoticed.

The purpose of data exploration systems is to reveal structure (or equivalently "pattern") in data. The basic operations of a data exploration system consist of describing, detecting, and searching for structures. Ideally, the detected structure did not arise by chance and can be described in the terms of the subject matter that produced the original data. The statements above naturally imply issues regarding the role of domain specific (i.e., subject matter) knowledge in data exploration.

Collecting data is often much easier than transforming the data into useful knowledge. The result is a lag between the time the data becomes available for analysis and the time knowledge can emerge from an analysis of the data. This problem is aggravated in areas where data is collected about phenomena to discover what factors affect performance and the relationships among the factors. The goals of analysis are not only to improve future performance but also to use the data to understand the underlying principles governing the observed behavior. The time and effort required to analyze a large amount of data representing a variety of qualitative and quantitative attributes is often quite prohibitive. It is from within this context that data exploration systems have emerged.

A previous NASA application of data exploration in 1983 used the AUTOCLASS system for the analysis of the Infrared Astronomical Satellite (IRAS) Data. For one year, the system sampled 94

spectral intensities and 2 celestial coordinates. In all, there were 5,425 records of data. Human data analysts spent two years analyzing and classifying the data into a known, but inadequate, taxonomy. In 1987 the AUTOCLASS program, a domain independent program based on Bayesian statistics, was applied to the IRAS data [Denning]. The program ran for 36 hours and created a new classification scheme and detected statistical patterns in the data that humans interpreted as "discoveries". The overall response to the AUTOCLASS system, however, has been reserved.

This paper presents our research into data exploration systems. The research is oriented toward analyzing databases of historical performance data for patterns indicative of success and failure. The findings presented here are applicable to any data exploration system. The first area discussed is capabilities: which patterns are sought in the data, which techniques identify the patterns, and how the data patterns are presented to a human investigator. Next, a methodology is discussed for effectively applying a data exploration system (i.e., how do the system's characteristics affect how it is applied). The research is a timely contribution to data exploration, as it identifies weak and missing capabilities of these data discovery systems, and, in some cases, the recommendations have been implemented and investigated.

This paper will show that induction and generalization over a database cannot be completely free of domain knowledge. At a minimum, the data exploration system must account for the semantics of numeric data. Next, classical statistical methods must be employed with great care because both the statistical hypothesis and the method of data collection strongly affect the validity of induction and the interpretation of the resulting generalization. Finally, the architecture of a data exploration system reflects the state of knowledge about the problem domain : when little is known about the domain being explored, the system is a loosely coupled set of tools supported by metadata. The more known about the domain, the less exploration there is and the more predictable the analysis becomes.

2. EXAMPLE of DATA EXPLORATION

Consider the following hypothetical database containing data regarding the past performance of a pre-launch rocket fuel monitoring subsystem:

Table 1. Sample Database

Manufacturer	Sensor Type	Launch Time	Number of Sensors	Sensor Indication
ABC Corp.	pressure	morning	6	anomalies
XYZ Inc.	temperature	afternoon	12	clear
ABC Corp.	density	night	6	clear
XYZ Inc.	volume	morning	6	anomalies
XYZ Inc.	volume	morning	12	clear
XYZ Inc.	pressure	afternoon	6	anomalies
ABC Corp.	density	night	12	clear
XYZ Inc.	volume	afternoon	6	clear
ABC Corp.	temperature	morning	6	clear
XYZ Inc.	pressure	night	6	anomalies
ABC Corp.	temperature	morning	12	clear

For simplicity's sake, there are only eleven launch descriptions. Sensors are manufactured by either ABC Corp. or XYZ Inc. The sensors measure one of four possible fuel-related factors: the temperature, pressure, density, or volume. Launch times are categorized as either morning, afternoon or night (after-dark) launches. Sensors are installed in batteries of six, with a maximum of two batteries or twelve sensors. Finally, each data record is "classified" by its reading indication. The sensor indication exhibits two possible readings : clear (i.e., a "successful"

reading) or anomalies (i.e., inconsistent sensor reports). This gives a possibility of (2*4*3*2) 48 different attributes to describe the environment of a sensor reading. The structure of the data is shown in the decision tree in Figure 1. The attributes are shown as nodes and the edges are labeled with the attribute values. The basis for classification is the indications "clear" and "anomalies", shown as plus and minus signs. The ID3 induction algorithm was used to create the tree.

One hypothesis from this analysis is that the sensor type was the most important factor in successful sensor readings, followed by the number of sensors, the manufacturer of the sensors, and finally the external temperature. This observation implies that some launches had only this one factor in common and that this one factor might have contributed significantly to the success or failure of the sensor readings. In this example, the induction algorithm is identifying factors that seem to be responsible for sensor success or failure. Specifically, pressure sensors tended to account for anomalous readings while sensors monitoring the fuel temperature and density tended to read successfully. This observation remains constant regardless of the values of the other attributes. Likewise, sensors monitoring fuel volume with twelve sensors were likely to detect fuel status correctly. However, when monitoring fuel status with six of XYZ Corp's fuel volume sensors, the external temperature became a decisive factor. Morning launches were likely to exhibit anomalous sensor readings while afternoon launches did not. Once the factors are identified, we now must seek verification of the patterns found by generating an explanatory hypothesis (i.e. a hypothesis that explains why the observed pattern is true and thus how to encourage or avoid the pattern in similar situations -- a "lesson learned"). Hypothesis generation is discussed later.

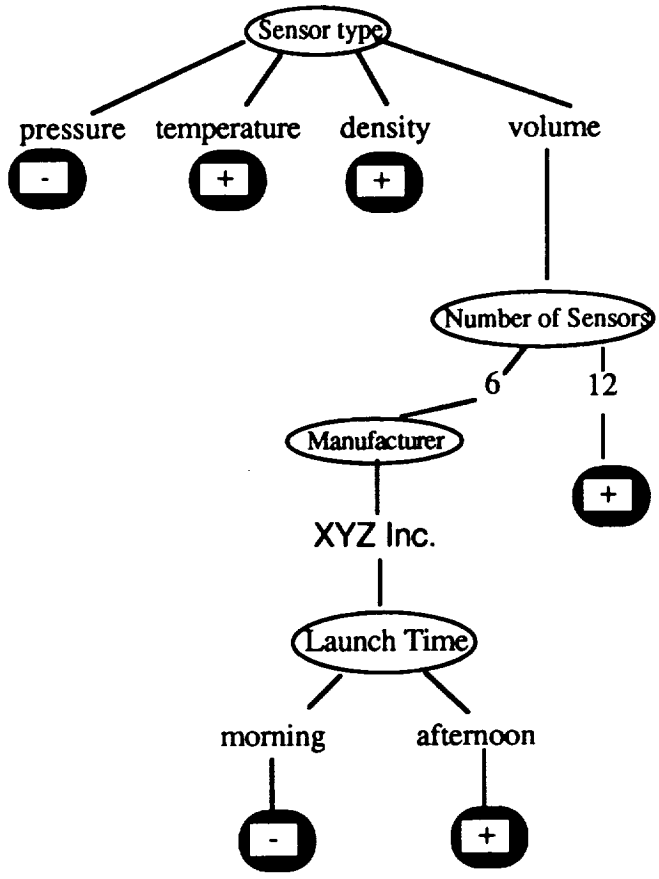

Figure 1. Decision Tree Induced from Table 1.

The structure of the decision tree above can be automatically generated and may serve as a basis for hypothesis generation. For example, why were pressure sensors such a determining factor? One hypothesis is that the fuel pressure varies within the shuttle's rocket boosters between the sensors located at the top of the booster and those at the bottom. This hypothesis could provoke a deeper investigation. It may describe a significant pattern of sensor failure: launches with sensors from different manufacturers, performed at various times of the day, using a different number of sensors, all exhibited anomalous fuel status readings when using fuel pressure sensors. The type of sensor was their only common factor. Likewise, why were morning launches subject to invalid sensor readings? One factor might be the number of sensors. When the number was twelve, readings in similar launches were successful. Thus, a possible explanation of the detected pattern can serve as a hypothesis and lead to important yet subtle new lessons learned which the data supports and yet, perhaps, goes unnoticed or unmentioned by any human analysts. One reason patterns may be hard to detect is the sheer volume of data coupled with the large range of attribute values. Without some form of statistical summary, how can an analyst extract as much information as possible from such high volume data? The automatic inductive analysis described above applies such analyses to the data and creates a discrimination tree as depicted graphically in Figure 1. In the discrimination tree, an analyst can rank the attributes with respect to their ability to classify sensor indications in order to determine which play significant roles in correlating the data as success or failure. In short, the induction algorithm above answers the questions "which attributes were the most significant" and "how are the data related?"

One can see the value of the analysis presented above. Given larger data sets with more attributes and greater attribute ranges, the analysis task would become insurmountable without a tool such as the induction algorithm. It is important to note that the induction algorithm will identify any pattern supported by the data set. However, the pattern may be only coincidental or trivial. If the data have cause/effect, correlations, or other useful information, then the induction algorithm will find it and make the relationships explicit for use in hypothesis generation [Parsaye, Hoaglin].

3. CAPABILITIES of DATA EXPLORATION SYSTEMS

The main issue of data exploration systems is the type of regularities the system can detect. Each system's approach can detect some regularities but is ignorant of other types of regularities. In short, data exploration systems detect regularities that their designers deem important. The domains of applications however, may exhibit certain types of regularity. To what degree can data exploration systems remain domain independent and "generalized"? How should a "regularity" be defined and how can we ensure that the assumptions for its detection are satisfied by the data in question? And, once a pattern has been detected, how should the pattern be interpreted by the user? Hamming's motto for those applying numerical methods also suits those applying data exploration systems : the purpose of computing is insight, not numbers. The choice of computing technique affects how we understand the results [Hamming].

As stated, each system's approach can detect some regularities but is unable to detect others. For example, consider a data exploration system determining the relationship between two variables by applying the chi-square test on a five-by-five contingency table of data values. For the sake of example, assume there are only two real-valued attributes X and Y, and the technique is applied to discover if these two attributes are related and, if possible, describe the relationship. Let the graph of the actual values be the sawtooth wave shown in Figure 2.

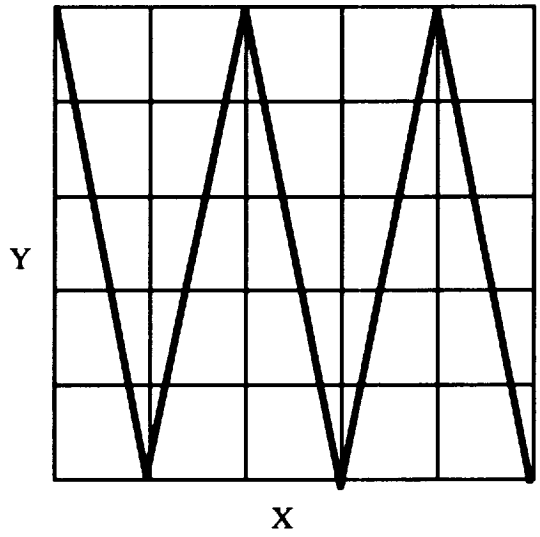

Figure 2. Chi-square test for attribute relationship.

The null hypothesis is that the variables are unrelated; i.e., each bin (or square) is uniformly distributed with data points. To reject the null hypothesis, the observed values must be different from the expected values appropriate to the desired significance level. Yet, as Figure 2 shows, the points are uniformly distributed, as the bins contains approximately the same number of data points. The chi-square test as described would not allow us to reject the null hypothesis. In other words, the attributes are judged to be unrelated because, in this case, the data exploration system fails to detect periodic behavior. If the amplitude or frequency of the wave is varied, the dynamic behavior of the data exploration system reaches critical points where the system can detect a regularity using the chi-square test, but as the amplitude or frequency of the wave crosses critical thresholds such as the one depicted in Figure 2, no relationship is detectable even though the essence of the regularity remains the same. It appears as though the phenomena under examination experience moments of chaos when, in fact, the chaotic behavior is inherent in the detection technique.

As this example suggests, the capabilities of the data exploration system should match the features of the explored context or the user risks "discovering" artifacts of the discovery mechanisms themselves and, thus, inducing invalid generalizations. Yet, matching a technique to problem context can be paradoxical: isn't one of the purposes of data exploration to discover precisely these descriptive features?

Nevertheless, several capabilities can be enumerated. First, the data exploration system should offer a variety of techniques to detect potential regularities and constrain their application appropriately. Without this capability, one cannot be sure that the regularities detectable are the only ones in the data. Multivariate statistical methods offer such a variety and, in addition, constrain when a method applies. Data exploration tools often incorporate many of these sophisticated methods but fail to detect when the techniques do not apply. Thus, the results of applying these techniques can be invalid yet appear as authoritative. The problem of invalidity is discussed in Sections 5 and 6.

Second, the data exploration system should support a flexible and appropriate strategy for data exploration. A *statistical strategy* is a formal justification for the selection, ordering, and application of techniques made during the course of data exploration. The data exploration system should in some sense act as an intelligent practitioner of data analysis [Hand]. Some systems

simply provide a "grab bag" of exploratory techniques that the user selects and applies ad hoc and without a sensible strategy. The worth of hypotheses resulting from such methodology should be suspect. The next section addresses this topic in more detail.

Next, the data exploration system should permit exploration sessions to be "frozen" and resumed. Without this capability, analysis must be conducted in a single sitting. However, an exploration strategy may induce a potentially complex series of probes into the structure of the data and it may not be practical to perform these operations in a single sitting. Furthermore, the system should record the lines of exploration and provide a "replay" capability of the last n operations.

Fourth, the data exploration system should provide a facility for semantic data modeling, as the semantics of the data do affect data analysis. Without this capability, two situations can arise: the relationships and generalizations discovered may be invalid or the system may fail to pursue a semantically rich relationship because two semantically related pieces of data are treated as unrelated by the system. A potentially rich semantic link will be lost. For example, the system can formulate relationships between two numerical variables if these numbers are related in the modeled domain; e.g., pressure and temperature. The importance of data modeling cannot be overemphasized in data exploration and is discussed in more detail in section 6.

Finally, there is the issue of missing or noisy data. How should this data be incorporated into the exploration process? Should there exist a three-valued logic for data values: unknown, not detected, or present? Similarly, some data may not be noisy at all, and other data must be smoothed before analysis. Clearly, a theory of data exploration should address these issues.

4. CONTEXT of DATA EXPLORATION

Next, how does a data exploration system support the analysis process? Some systems such as BACON use empirical numerical data directly to induce a law supplemented with theoretic variables that both simplify the representation of the law and serve as a conceptual aid [Langley]. For example, BACON induced Black's heat law and the concept of "specific heat". Other systems such as IXL operate more interactively by placing a human in-the-loop [Parsaye]. In this case, the data exploration system is more an integrated set of loosely-coupled tools. BACON gives the user an end result while IXL gives the user a set of intermediate results. The style of investigation imposed by the data exploration system should be appropriate to the problem under investigation.

At best, the exploration tool should guide the user in selecting an exploration strategy and then ensure that the data satisfies the assumptions underlying the selected mathematical techniques. Given that the data represents some aspect of reality, the strategy for seeking out implicit or hard-to-perceive relationships should make sense within the specific context of exploration.

The data exploration environment can be categorized as either "supervised" or "unsupervised". In a supervised learning environment, a human analyst supplies the exploration tool with metadata, or information that describes the various data attributes that are found in the data records. This gives the exploration tool knowledge of the environment the data is supposed to describe. With this data, an exploration tool can determine the appropriate set of tools that it can validly apply to the data. An unsupervised learning environment exists when no metadata is supplied. In this case the exploration tool must examine the data using the widest variety of tools, though the validity of the application of these tools to the database is up to a human analyst to determine. This mode of learning is valuable when minimal knowledge concerning the nature of the data is available.

Consider, for example, a database containing numerical data attributes. If nothing is known about this data, a numerical induction system such as BACON would apply its analysis heuristics in an attempt to determine wheher the terms are numerically related. This will very often provide an analyst with valuable information regarding the relations between terms. As stated earlier,

BACON has discovered and "rediscovered" many valuable numeric laws. However, if the data represents attributes such as a zip code, a year-of-birth, a social security number, or a yearly income, any numerical relationships discovered that relate these terms has no meaning in the real world. The purpose of data exploration systems is just the opposite: to discover trends that can be used to make generalizations or predictions about the real world. Even when the numerical data found in a database does lend itself to numerical analysis, the nature of these "measurements" must be known in order to perform valid analyses. This is discussed in more detail in Section 6.

Data collection issues are key to data exploration as well. In many instances, large volumes of historic data exist that can be readily applied to data exploration. However, many situations exist in which data is constantly being collected. Consider the amount of data that is constantly being transmitted regarding the status of the many subsystems aboard the space shuttle. If a data exploration system is to be used to detect patterns and correlations for one-time-only analysis, then this is not an important consideration. Other applications however, will require the constant digestion of data streams that describe a real-time environment. A data exploration tool should be able to analyze a fixed sample or to accept incoming data and continuously modify the detected patterns to reflect the current information (i.e., sequential analysis strategy). Without this ability, an accurate determination of the significance of the various attributes being collected cannot be made.

5. INTERPRETING the RESULTS of DATA EXPLORATION SYSTEMS

Data exploration systems can serve as a powerful instrument enabling an analyst to perceive structure in a seemingly dense forest of data. Ideally, the derived perception of the data reflects relationships and generalizations actually present and not due to chance. The data exploration system merely enhances the analyst's perception much the same way as a telescope enhances an astronomers perception. This section outlines the technical difficulties in achieving this goal. The crux of the matter is validity. Under what conditions might a data exploration system offer invalid results? The thesis is stated simply: the interpretation of results can be complex because a statistical strategy and the semantics of numerical data can strongly influence the interpretation of the results.

For the sake of illustration, consider a simple hypothetical database adapted from [Berger] containing data regarding two identical subsystems serviced by different maintenance teams (i.e., paired observations). It will be shown that the statistical strategy can be subjective and can strongly influence the discovered structure. Thus, a data analyst cannot interpret the results without knowing precisely the details of the statistical strategy used to discover the patterns. The conclusion is that the consumer of the data should control the strategy of analysis, not the data exploration tool. Assume that before a launch, each subsystem is assigned a maintenance team at random and after the flight the performance of the subsystem is evaluated. Next to the paired observation is the outcome stating which subsystem performed best (for the sake of clarity, assume ties are not allowed). Let attribute labels 1 and 2 represent the subsystems, and let their values represent which of the maintenance teams, A or B, serviced the respective subsystem. Attribute 3 represents the post-mission evaluation of which subsystem/maintenance team performed best. The database appears in Table 2. Attributes 4 and 5 are discussed shortly.

Table 2. Sample Database

1	2	3	4	5
A	B	A	42	30
B	A	B	39	41
A	B	A	39	23
B	A	A	43	33
A	B	A	40	25

A	B	A	46	40
A	B	A	44	35
B	A	B	37	39
A	B	A	54	60
B	A	A	52	54
A	B	B	42	40
A	B	A	52	54
B	A	A	50	50
A	B	A	45	38
A	B	A	47	44
B	A	A	52	54
A	B	B	38	39

The first issue regarding the interpretation of the results is very complex: what are the assumptions of the statistical strategy, and given that these assumptions hold for the data, how do these assumptions influence the interpretation? To illustrate the effect of exploration technique on interpretation, two separate statistical viewpoints are assumed and their implications examined. First, consider the view of classical statistics. Classical statistics assumes a model responsible for the data prior to examination, and that any deviation from the assumed model is caused by chance. Given the sample database, consider the following strategies:

- Strategy 1 assumes that the data is generated by a binomial distribution with no difference between the maintenance teams. The probability of the outcome of 13 favorable outcomes for Team A is 0.182 and the P-value is 0.049. One is tempted to assume that there is no difference in maintenance teams and that the pattern of 13 favorable outcomes for team A is due to chance.

- Strategy 2 assumes that the data is generated by a negative binomial using a sequential sampling plan: (i.e., data was collected until both Team A and Team B both had 4 favorable outcomes). It so happened that the last favorable outcome for Team B occured on trial 17. In this scenario the probability of the pattern of 13 favorable outcomes for Team A is 0.0085 and the P-value is 0.021. The conservative judgement is that the pattern in favor of Team A is not due to chance. The "discovery" is that Team B needs training.

What is the cause of this ambiguity? The ambiguity is caused by assuming a statistical hypothesis that in turn results in P-values for values that are unobserved yet theoretically possible.

Which of these interpretations is true and how can an analyst communicate the sampling plan and assumptions to the data exploration tool? More importantly, does the data discovery tool accommodate such metadata prior to exploration? From the above example, one can conclude that the sampling plan and assumed distribution strongly influence the interpretation of results. These assumptions are inherent in the classical statistical approach. The two distributions are sketched in Figure 3 for comparison. The leftmost distribution results when a coin is flipped 17 times and the probability of the number of heads is calculated. The rightmost distribution results when a coin is flipped until four tails appear, the last tail occurring on toss 17. The P-values are also indicated in the shaded areas; these indicate the "confidence level" of the null hypothesis.

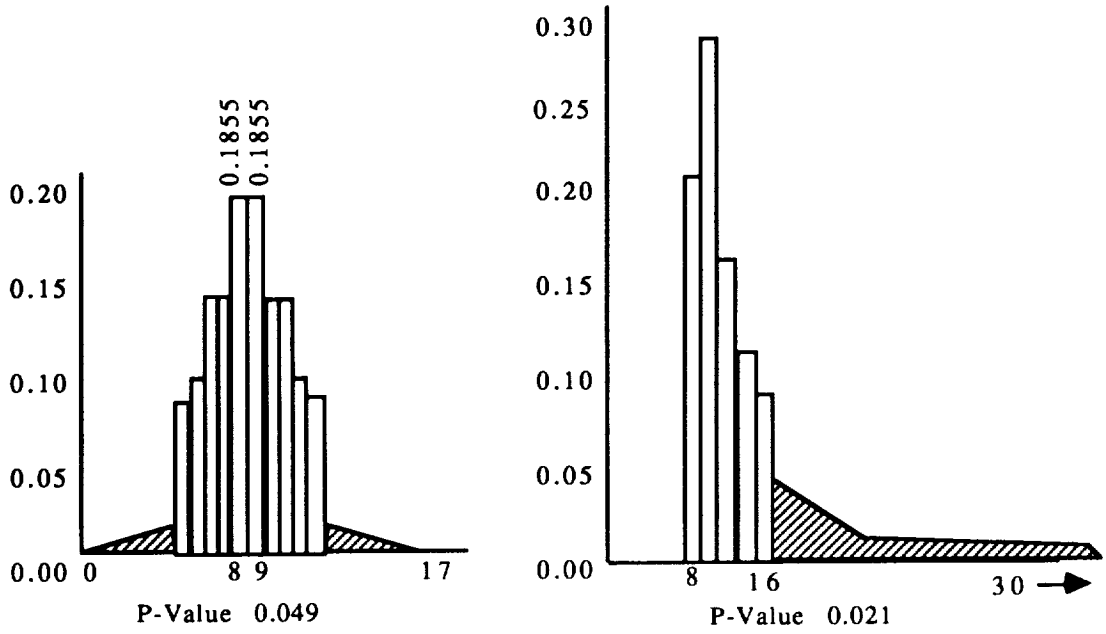

Figure 3. Two Statistical Models for Sample Database

The role of P-values is fundamental in the classical approach to data analysis, yet it is often misunderstood. P-values consist of both observed data and hypothesized (unobserved) data; these represent the probability that the null hypothesis is falsely rejected *assuming that the null hypothesis is true*. P-values represent the probability of obtaining data that casts at least as much doubt on the hypothesis as on the observed data itself. However, the P-values do not indicate the truth of the null hypothesis in the face of the data. P-values essentially give indirect evidence against the null hypothesis.

An alternative way of exploring the data is a Bayesian approach. On the surface, the Bayesian approach seems more appropriate to data exploration. First, the Bayesian approach allows the formal integration of prior knowledge into the hypothesis generation process. Next, the Bayesian approach permits sequential sampling and the accumulation of evidence. The ability to accumulate evidence and make decisions when the evidence becomes strong enough is appropriate to many experimental engineering analysis problems. The classical approach requires sample sizes to be specified in advance and the sampling conditions to remain uniform for the duration of the sampling. Then the data is analyzed. However, this type of sampling and decision regimen may not be suitable for the analysis of real-world, mission-critical operations. More appropriate for hypothesis generation is the integration of the best engineering judgement into the analysis of the arriving data and making the best decisions possible given the data at hand. Ideally, the set of hypotheses can be scored and the most probable hypothesis selected based on direct evidence plus prior knowledge. Classical data exploration techniques are of limited value in this situation. Bayesian techniques on the other hand, directly estimate the truth of a hypothesis given the data. In short, the Bayesian approach addresses itself directly to the issue of how degrees of belief are to be altered by the *observed* data. Contrast this approach with the concept of P-values, which account for *unobserved* data. Bayesian techniques are unaffected by unobserved data and thus permit a sequential sampling plan; in addition Bayesian techniques are not invalidated by nonuniform sampling conditions or affected by experimental design. Data exploration in mission-critical contexts can benefit from a Bayesian approach, yet few data exploration systems fully accommodate this approach.

6. HOW the SEMANTICS of NUMBERS AFFECTS INDUCTION and GENERALIZATION

As discussed earlier, a key issue in data exploration is the amount of domain specific knowledge necessary to apply mathematical pattern detectors. This again touches on the nature of supervised learning. A tentative answer is that data exploration does require at least a "modest" amount of semantic metadata. To see this, consider the two integers 20 and 45. If their mean value is to be used for statistical purposes, one might use the arithmetic mean to obtain an average of 32.5. However, if these integers represent a "rate" such as kilometers per hour, then the mean value should actually be 27.77 (assuming an equal weighting). When determining the mean of rates involving unit ratios (miles/hour, ohms/meter, etc.), the *harmonic* mean must be used to determine the true mean. For example, if a vehicle goes one kilometer at 20 km/hr and the next kilometer at 45 km/hr, the average velocity is 27.77 km/hr, not the 32.5 k/hr that the *arithmetic* mean would indicate. Now, if this mean is to be used to calculate other values, such as times of arrival given certain time intervals, this difference in velocity will quickly cascade through further calculations, most likely unnoticed. Though simple, this example demonstrates that all numeric data cannot be handled uniformly. A data exploration system must have some metadata describing what the numbers represent.

One important shortcoming of most data exploration systems is that numerical values *are* treated uniformly. Numbers represent some aspect of reality and the results of numerical operations are assumed to make true statements about this reality. However, our research indicates that, in general, numbers cannot be treated uniformly and mathematical operations cannot be applied without empirical justification. Specifically, if a user cannot communicate the semantics of numerical values to the data exploration system, the corresponding results may be suspect. "Measurement theory" offers important design guidelines detailing the qualities of measurements that can be used to avoid the above problems.

Consider attributes 4 and 5 in the sample database in Table 1. If numerical values are treated simply as "numbers", then the pattern "when attribute 3 = 'A', then the value of attribute 4 = 2/5(attribute 4) + 30" might be detected. Is this valid? At first glance it seems that, indeed, this generalization is true. Yet further examination demonstrates that the validity of this generalization depends on the semantics of the numbers themselves. First, assume that the numbers represent the payroll numbers of the two supervisors of Team A and Team B. The generalization in this case is probably senseless. However, if the two attributes represent outside temperature and the temperature of a subsystem, the generalization is valid only within the scale on which the temperatures were measured. In other words, the generalization is invalid if, in the future, temperature is measured in Fahrenheit instead of Centigrade or vice versa. The co-ordinate system itself induced the generality. If we change the co-ordinate system, the generalization disappears. However, if the attributes represent the weight of two related subsystems, then the generalization is valid regardless of the co-ordinate system in which the original measurements were made. The reasons for these assertions are grounded in measurement theory.

Briefly, measurement theory is a branch of mathematics that formalizes the practice of associating numbers with objects and empirical phenomena and the interpretation of those numerical values. Numbers can take four different meanings, and these meanings constrain the types of operations that result in valid application. First, a number can be "nominal". This means the number represents a qualitative symbol such as a name. In the example above, employee number is nominal. Next, a number can be "ordinal". This means the number represents a location in a ranking but not magnitude (e.g., the object ranked fourth does not necessarily have twice the ranked property as the item placed second even though (2)*2 = 4). Third, a number can be a "ratio" measurement. This means that the number represents a measurement with an arbitrary scale and origin. In the example above, temperature is such a measurement. Note that if temperature X is twice temperature Y in Fahrenheit, this is not necessarily true if the numbers are converted to

Centigrade. Finally, a number can represent an "interval" measurement. This means that the property measured has a "natural" absolute origin. In the example above, weight is such a measurement because at zero G, the object measured has no weight regardless of the specific measuring scale. A table of allowable transformations is given below.

Table 3. Valid Operations Based on Measurement Type

Type of Measurement	Admissible Transformation	Example
nominal(symbolic)	any 1-1 transform	numbers used as labels
ordinal	x >= y iff f(x) >= f(y)	preference rankings
interval	f(x) = ax + b	temperature in C or F
ratio	f(x) = ax, a > 0	weight

Table 3 indicates that the generalization "X is cY where c is constant" is an invalid operation on interval measurements such as temperature because generalizations induced in one temperature scale do not necessarily hold in another scale. The admissible transformations for a measurement type are the "litmus test" for the validity of an inductive generalization. See the Appendix for proof of this assertion.

The interested reader is referred to [Roberts] for details, implications, and formal proofs about the validity of numerical inferences. The importance of measurement theory to data exploration cannot be overstated. This theory shows that numerical values cannot be treated uniformly, because the results can be invalid. On the positive side, measurement theory also indicates that data exploration systems need not possess a great deal of domain specific knowledge. All that needs to be insured is that the appropriate transformation is applied to a numerical value. If the dictums of measurement theory are obeyed, then the opportunities for invalid generalization are reduced.

7. INTEGRATION of SYMBOLIC and NUMERIC DATA

A related issue is the integration of symbolic and numeric data. Most systems use one type of data exclusively and fail to exploit the information carried by both types. Yet, many databases contain both types. A single coherent computational framework is needed for using both qualitative and quantitative data in searching for potentially meaningful patterns. One framework for integrating both symbolic and numeric information is to cluster the numeric information, assign symbolic cluster names, and use the cluster names in an algorithm such as ID3, which discovers low entropy attributes with respect to a given taxonomy. This technique was tried on our prototype system using a simple Euclidean distance measure and the *maximin-distance* clustering algorithm. The key issue in this approach is to make an informed guess about the numeric data and select a suitable distance measure. Once again, a modest amount of domain-specific knowledge must be applied. If one knows nothing at all about the data as shown in this section, the results of any exploration approach must be suspect.

Cluster analysis organizes data by uncovering underlying structure in data either as a grouping or a hierarchy of groupings. The analyst can use the grouping as confirmatory evidence of suspected structure or as fertile ground for further experimentation to explain the discovered taxonomy [Everitt]. The AUTOCLASS program mentioned in the introduction is based on cluster analysis. An important theoretical consideration for employing cluster analysis is that it is free of the ambiguities induced by P-values and other statistical assumptions previously discussed. Thus, cluster analysis can serve as both a remedy for the problems of classical or Bayesian statistical methods or as an additional validation technique to supplement these methods. The database of Table 2, when clustered, appears in Figure 4:

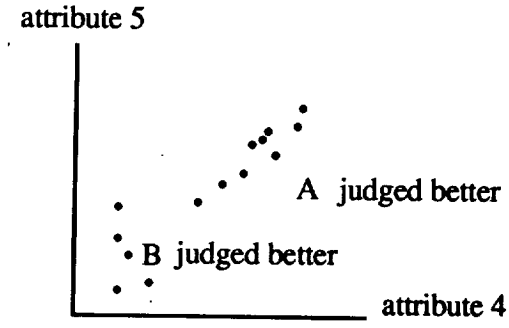

Figure 4. Clustering

One hypothesis from the clusters might be that Team A actively trains, while Team B is shamed into training also. When Team A does not train, Team B relaxes and then wins on talent alone.

There are several issues involved with cluster analysis. First, how are clusters defined? Exactly what shapes and distances define a cluster are domain dependent parameters and are often unknown before analysis. Next, research suggests that clustering without some domain knowledge is still a weak technique. The reason is that clusters are defined by distance measures that themselves have meaning within a domain. The Euclidean distance is just one example of a potential distance measure useful in clustering. Finally, which attributes should serve as the clustering attributes? These all depend on the goals of the analysis. Consider the different ways of clustering a deck of cards: one may form clusters based on numeric value, suit, color, etc. The data exploration system should provide a mechanism for defining distance measures and allowing the user to search for structure based on different grouping criteria.

Our experiments use clustering and distance measures not only as stand alone techniques (as in the example) but in data reduction as well. Clusters of numeric data are tagged with names, and the names serve as an additional attribute that represents and classifies a group of the numeric values. Then a symbolic induction technique such as ID3 finds those factors correlated with the cluster. A similar technique is used for curve fitting. When a set of numeric data is fitted with a curve, the coefficients of the curve are compared with the curves fitted to related data. Using the Chi-Square Test, a distance measure is defined, and a determination regarding the two sets of numeric data is made. If the data are close, they are grouped within the same cluster. The same maximin algorithm can be applied with a different distance measure.

8. CONCLUSIONS AND RECOMMENDATIONS

Our research suggests several conclusions. First, data exploration is based on the detection of regularities in data. Therefore, the data exploration tool should provide a variety of detection techniques for the types of regularity likely to occur in the data. Without this, the system is blind to potentially important and characteristic patterns. Next, symbolic data patterns can be detected using a variety of statistical techniques. Various entropy measures have been shown to be useful in this endeavor. Decision trees constructed from the output of algorithms such as ID3 based on entropy measures offer a visual representation for the structure of the data. Third, data exploration on numeric values is very complex due to the semantics of numbers. Care must be taken to avoid transforming the data in meaningless ways and deriving invalid patterns on the data. Finally, more research is needed into integrating symbolic and numeric data into a coherent framework for data exploration.

Several recommendations are made for future directions for research. Some enhanced data exploration techniques have already been implemented and are undergoing experimentation.

Experiments are underway to integrate symbolic and various types of numeric data into an overall methodology for data exploration. Promising techniques include the incremental integration of domain-specific knowledge into cluster analysis and curve-fit analysis.

The role of conjecture in the discovery process is well-recognized [Polya]. As stated previously, data exploration tools can support the formulation of a conjecture by examining the data and elucidating the "structure" of the data. Structure in the sense used here means "regularity" or generalization exemplified by the data. Regularities are described and hence detected mathematically. The user must verify that the assumptions underlying the application of the mathematical technique are satisfied, and if so, then the data exploration tool can perform a great deal of statistical analysis and uncover the structure of the data. Certainly, powerful tools such as multivariate statistics and information theory can provide the data exploration tool with a sturdy vehicle for exploration.

However, a paradox arises. On one hand, an analyst applies a data exploration tool *because* so little is known about the data. But, on the other hand, the very exploration techniques require a certain amount of knowledge about the data before the results can be validated. In other words, if a data exploration tool presents us with a host of conjectures, what can be said about the potential validity of the conjectures, all else being equal? Ideally, one does not need a tool to manufacture blind alleys, smokescreens, and distractions for the analyst.

What is needed is an incremental approach to integrating domain knowledge as it is acquired into the exploration process. This paper has shown that induction and generalization over a database cannot be completely free of domain knowledge. At a minimum, the data exploration system must account for the semantics of numeric data. Next, classical statistical methods must be employed with great care because both the statistical hypothesis and the method of data collection strongly affect the validity of induction and the interpretation of the resulting generalization. Finally, data exploration systems can be structured in two ways: when little is known about the domain being explored, the system should be a loosely coupled set of tools supported by metadata. The more known about the domain, the less exploration is needed, the more predictable the analysis becomes, and the more domain specific knowledge should be infused into the analysis process. The tool supporting the analysis should be flexible enough to accommodate a variety of data and analysis strategies.

APPENDIX

Proof that the generalization of f(a) = cf(b) is invalid for interval measurements (e.g., mass, temperature on F or C, etc.) where $c > 0$.

Proof: Assume f(a) = cf(b) for some a,b,c where f(x) is a quantity assigned to x on an interval scale. The generalization is valid iff it is invariant under all admissible transformations. Since f(x) is assumed to be an interval measurement, let g(x) = kx + b where $k,b > 0$, the general admissible transformation for interval scales. If f(a) = cf(b), then (g*f)(a) = c[g*f](b) --> g(f(a)) = c[g(f(a)]. But, k f(a) + b <> c[kf(x) + b]. Hence f(a) = cf(b) is an invalid generalization in an interval scale. Note, however, that if b = 0 (i.e., the measurement type is ratio such as weight), then the generalization is valid.

REFERENCES

Berger, J.O. and Berry, D.A. (1988) Statistical Analysis and the Illusion of Objectivity. *American Scientist* , 76(2)

Denning, P.(1989) Bayesian Learning, *American Scientist* , 77(3)

Everitt, B. *Cluster Analysis* . Wiley and Sons, New York, N.Y., 1974

Hamming, R.W. *Numerical Methods for Scientists and Engineers.* Dover Publications, Inc., New York, N.Y., 1973

Hand, D.J. Emergent Themes in Statistical Expert Systems in *Knowledge, Data, and Computer-Assisted Decisions*, Springer Verlag, Berlin 1989

Hoaglin,D.C., Mosteller, F.P., and Tukey , J.H. *Exploring Data Tables, Trends, and Shapes*, Wiley and Sons, New York, N.Y., 1985

Langley,P. and Zytkow, J.M. (1989) Data-Driven Approaches to Empirical Discovery.*Artificial Intelligence* 40

Parsaye, K. ,Chignell, M, Khosafgarian,S., and Wong, H. *Intelligent Databases*, Wiley and Sons, New York, N.Y., 1989

Polya, G. Mathematical Discovery, Wiley and Sons, New York, N.Y., 1981

Roberts, F.S. *Measurement Theory* Addison-Wesley, Reading, Mass., 1979