Benchmark Development for the Evaluation of Visualization for Data Mining

Georges G. Grinstein[1*], Patrick Hoffman[1*], Sharon J. Laskowski[2], Ronald M. Pickett[1]

[1]Institute for Visualization and Perception Research
University of Massachusetts at Lowell, Lowell, MA 01854
{grinstein, phoffman, pickett}@cs.uml.edu

[2]The National Institute of Standards and Technology, Gaithersburg, MD 20899
sharon.laskowski@nist.gov

Abstract: New sets of powerful data visualization tools have appeared in the marketplace and in the research community. This, combined with readily available computer memory, speed, and graphics capabilities, makes it possible to explore larger and larger data sets. However, it is difficult to judge the effectiveness of these tools for supporting large scale information exploration and knowledge discovery. In this paper, we describe a set of issues critical to benchmarking and evaluation in this domain. We then propose an approach to constructing an evaluation environment and report on initial results from a prototype environment in which we tested five visualization approaches against nine existing data sets.

1 Introduction

We are currently seeing a rapid growth in the development of tools and techniques for supporting knowledge discovery in databases (KDD). New sets of powerful data visualization tools have appeared in the marketplace and in the research community. This, combined with readily available computer memory, speed, and graphics capabilities, makes it possible to explore larger and larger data sets. While this trend has served to increase the interest and effort of corporations in exploring their data for hidden nuggets of information, these visualization tools are not well integrated with data mining software, and it is difficult to judge the effectiveness of either the visualizations or the data mining.

To remedy the situation, it is becoming increasingly important to develop appropriate data sets and reproducible benchmark tests to identify the current best practices and to steer development of future systems.

In this paper, we discuss some of the issues that need to be addressed in order to provide benchmark testing and evaluation to the visualization and data mining communities. We survey evaluation approaches that have been applied in other

[*] This work was funded, in part, with research funding from the National Institute of Standards and Technology.

information technology domains and then describe a basic framework in which to perform evaluations. We conclude with a discussion and examples of various visualization techniques, each exercised on several different data sets. These examples comprise illustrate the kind of environment for testing that is critically needed to advance the development of visualization for data mining. Such an environment, when fully developed, should provide a broad array of tests for comparative evaluation on a common set of criteria and provide for comparisons across systems on the same data and tasks.

2 Background

In a Meta Group, Inc. survey, "Data Mining: Trends, Technology and Implementation Imperatives", it was found that the data mining market will grow 150% to $8.4 billion by 2000. Half of 120 companies surveyed believe that data mining will be critical for their businesses in the next two years [Wong97]. Some expect that visualization software will proliferate even on Wall Street over the next few years [Yras96] in response to its special needs to comprehend complex data. What is often missing in all this talk, however, is a recognition of the difficulties to be faced. Without the appropriate visualization techniques, these data mining approaches will remain difficult to use and require a great deal of expertise. Corporations understand the promise of data mining to wade through large amounts of data, but they are not adequately aware of the human limitation in grasping what the analyses show [Mart96], and many are finding that the tools just cannot handle the volume of data they are gathering [Sted97].

It is clear that no one general set of visualization tools will be suitable to address all problems. Different tools must be chosen based on the task and data. Currently, there is little guidance for these choices. The only way to address this problem is through the development of evaluation methodologies and benchmarks that show the strong and weak features of specific classes of visualizations. Then we can begin to answer the question, "How does effectively slice, dice, plot, color, and interact with data in a visualization?"

A recent special advertisement in Computer World describing a data mining "face-off" provides a good example of the need for such guidance. Five companies participated in a "competition" in which they described how they would respond to two hypothetical Requests For Proposals that could be solved by a data mining and/or data warehouse solution. The solutions varied widely, from a total data management strategy to the data mining alone, and ranged in cost from $150K to over $1M. This wide discrepancy in approach and cost makes clear the importance of being able to sharply evaluate the solutions to choose the best.

Usama Fayyad, in a recent editorial [Fayy97], makes the point that the database and information retrieval communities have met with great success in advancing algorithm performance by establishing benchmark data sets, and he believes that the KDD community could benefit as well.

Evaluations that produce clear benchmarks are also needed to steer development toward optimal solutions, models and theory. In other areas of information technology, such as speech recognition, image recognition, and information retrieval, benchmarks and evaluation metrics have clearly helped to move new technology into useful, reliable, and predictable products. We believe they are critical in this research area as well.

Before we specifically address the question of evaluation for visualization in the context of data mining, we first look at some of the successful approaches to evaluation in these other areas of information technology.

2.1 Benchmarking and Evaluation for Information Technology

The Information Technology Laboratory at the National Institute of Standards and Technology [NIST97] has been supporting and contributing to the development of tools to measure the effectiveness of information technology applications. The goal is to provide researchers, developers and users objective criteria for understanding how products and techniques perform and for assessing their quality. These tools include test and evaluation methods, metrics, and reference data sets. For example, NIST provides large unstructured text collections and uniform scoring procedures for the Text Retrieval Conference [TREC97]. This annual event, now in its sixth year, has proven to be an invaluable resource to the information retrieval research and development community. Its activities have enabled great strides in improving the search engines and in speeding the transfer of this technology. Large test corpora, queries, and associated pooled evaluations are made available to participants who are required to submit the output of their search engines for evaluation before the actual workshop. [Voor97] contains an overview and proceedings from the 1997 conference. TREC has encouraged research in text retrieval, increased communication among industry, academia, and government, sped the transfer of technology from research labs into commercial products by demonstrating improvements in retrieval methodologies, and increased the availability of appropriate evaluation techniques.

A similar effort by NIST has been the development of test corpora and evaluation methods for spoken language recognition. NIST has been involved in the creation and distribution of speech corpora—nearly 30 of them—and associated benchmark evaluations. These evaluations have proved to be critical in the recent commercialization of speech recognition. Details are contained in [NIST98a].

These efforts and others, in such areas as fingerprint recognition and optical character recognition [NIST98b], have been very successful. Hence, it is reasonable to assume that such an approach can also be applied to improving the quality of the next generation of tools that integrate data visualization into the KDD process.

2.1.1 State of the Art in Benchmarking and Test Data Sets for KDD

There has already been a substantial amount of effort in the area of benchmarking and test data sets for KDD. See, for example, the data sets identified in the Knowledge Discovery Nuggets web site [Kdnu98] such as those in the Machine Learning Database Repository and the Neural Networks Benchmarking web sites, which provide good starting points for reproducible experiments, especially for neural net algorithms. However, these sets suffer a variety of limitations. Many are very small or for very specific learning algorithms. Some of these collections are synthetic, that is, they were designed *a priori* to stress prediction algorithms in predetermined ways. Many of the large sets are from the statistical community rather than the visualization community and typically do not include benchmarks.

The Information Exploration Shootout [IESH97] developed at the University of Massachusetts at Lowell and the MITRE Corporation has begun to address the need for more serious comparative evaluations of the various data exploration techniques. The first two data sets, network intrusion and online daily news archives of Web pages, were chosen because of their timely subject matter and for their size (200 Mb, 1.2Gb respectively), as well as for their potential to have synthetic (planted) intrusions and to deal with "free-form" patterns of information (typical news and large amounts of other unstructured data). However, there has been no agreed upon set of metrics or evaluation criteria on which to judge and compare approaches to exploring these data sets visually.

Finally, in 1997, the Knowledge Discovery in Databases Conference organized its first Knowledge Discovery and Data Mining tools competition, the KDD Cup. This competition was aimed at demonstrating and comparing the effectiveness of tools in the area of supervised learning. The winners were determined on the basis of a weighted combination of classification accuracy (predictive power or "lift"), software novelty, efficiency, and the data mining methodology used. Note, that to properly evaluate the competition, entered data sets had to be analyzed ahead of time. For large data sets this is very time consuming. The emphasis here was more on data mining algorithms than on visualization. It is easier to measure accuracy of the classifications than to measure and compare one visualization with another. Visualization has a number of dimensions to be measured and is highly dependent on the user, the task, and the structure of the data. It is difficult to pull these out to identify an optimal method.

2.2 Issues in Benchmarking and Evaluation

In this section we discuss three major issues that contribute to the difficulty of creating benchmarks and evaluation methodologies for visualization techniques aimed at supporting data mining of large data sets.

2.2.1 Dependence of Performance on User Knowledge and Expertise

To illustrate the importance of factoring in user knowledge and expertise into a benchmarking effort, we relate some of our experiences with the Information Exploration Shootout's [Grins97] first exercise, which involved the detection of intrusions into a computer network. The two major challenges for the participants were the complexity of the problem domain and the size of the database. Details of the internet protocol and internet operations are arcane, and to adequately address, let alone solve the problem, an expert in the field is necessary. The central need for a domain expert is typically a common feature of real-world knowledge discovery problems. The skills that we found to be required in our approach were: 1) domain knowledge of computer network security; 2) experience with visualization software; and 3) statistical expertise.

The first task of the shootout was a large preprocessing activity. We grouped the individual packet records into natural clusters of communications sessions. The resulting reduction in size was substantial. For example, the baseline data set contains over 350,000 records. The corresponding session-level data set contains approximately 16,000 records. We then analyzed the processed data sets predominantly with visualization techniques. For many we used parallel coordinate plots and conditioning. Even in this step, the visualization is driven by domain knowledge. With this approach, several anomalies were identified, and these turned out to be network intrusions when interpreted by a system administrator aware of various network attacks.

This experience showed that there are a number of aspects of the process that have to be evaluated, much depends on domain expertise, and on the amount of data involved. Even with this large but not huge set, the visualization required a scaling down of the problem. Any benchmark testing methodology must consider these complex requirements. Testing must also include a good understanding of the perceptual issues involved, as discussed in this next section.

2.2.2 Perceptual Issues in the Evaluation of Visualization Systems

The challenge in conducting an evaluation of any system is to ensure that the evaluation is both valid and discriminating, and, where one system is to be compared to another, that the comparison is fair. By fair, we mean that testing must occur under controlled conditions: the challenges put to the systems must be equivalent and each system (or system variant, if incremental tuning or adjustment is being investigated) must be operated under similar conditions. In the case of comparing system speed of performance, obviously the systems must run on platforms with equivalent speed so that ensuring fairness with respect to purely computational operation of a visualization system is a non issue. Also, it is assumed that a system performs deterministically, that is, in exactly the same way computationally every time it operates on the same data. However, that is definitely not the case with respect to those aspects of the operation of a visualization system that involve human sensory, perceptual and cognitive processes. These can vary widely in their operation from one test of the system to another, and the fairness of comparisons can be undermined, unless care is taken to ensure comparable operation from test to test. We propose there are three basic ways in which comparability must be protected.

There is first the need to ensure comparability of performance at the sensory level. Several factors have to be considered, including display calibration, control of lighting and other viewing conditions, and adequate testing and selection of observers, particularly with respect to such critical aspects as color vision and stereoscopic vision where that might apply. Comparability at the level of perceptual processing must also be ensured, and that will depend in large part on whether the required perceptual processing of the visualization is pre-attentive or not. By pre-attentive, we mean that the process runs off automatically, that it requires no conscious analysis, only that the observer attends to the display. The possibility of encoding data into forms that elicit such automatic processing has been demonstrated in several exploratory visualization systems, see [Pick95]. To the extent that a visualization system depends on purely pre-attentive perceptual processing, the problem of ensuring comparability from test to test devolves to ensuring only that the various determinants of comparable sensory processing mentioned above are adequately controlled. But it is not likely that we will ever be able to depend entirely on pre-attentive perceptual processing in visualization systems.

We should aim to exploit pre-attentive perceptual processing as much as possible, but visualization will probably always have to depend on perceptions that require a large component of consciously controlled, deliberate analysis. This implies that, to a large degree, the effectiveness of the perceptual processes will depend on what is termed perceptual learning. The effectiveness is related to the degree to which the observer can learn how to look at, how to see, and how to assemble the various components and features of structures potentially visible in the display before they are adequately perceived. Perceptual learning has received extensive attention in the psychology literature; see [Gibs69], for example. Perhaps the best examples of dependence on perceptual learning for effective performance come from fields of medical image analysis, where, for example the pathologist in training only slowly learns with coaching to differentiate, say, the malignant from the benign specimen under the microscope, yet when experienced sees the difference instantly. Many visualization systems will depend on the observer's having learned how to perceive what is there. This means that when competing systems are tested and compared, the evaluators must ensure that the observers in each test have had adequate perceptual training and experience.

A third area important to consider in protecting comparability among visualization systems has to do with the methodology provided for the observers to conduct their analyses and report their findings. Alternative methods for laboratory testing of sensory and perceptual performance have received extensive development in experimental psychology, see, e.g., Chapter 2 in [Schi96]. The strengths and weaknesses of these alternatives have also received extensive study, as have the implications for their use in evaluating systems in real world settings. A good case in point has been the extensive debate and development of techniques appropriate for testing medical imaging and diagnostic systems, see e.g. [Swet82]. The particular methodology will affect not only the richness and precision of the analyses conducted and the reports produced, but can shape the basic nature of the game that the observer plays. It is vastly important in comparing one system to another not just that that the same methodology be applied, but that it be one that ensures, within its own operation, comparable figures of merit from test to test.

The best way to ensure that testing is comparable among the different systems to be evaluated is to have the evaluations done together in the same laboratory and, with appropriate protections for independence of the tests, on the same observers. This would suggest that ultimately one would like to develop a central testing laboratory. But, it would be possible, and perhaps more practical, to develop standard procedures that provide for testing in different settings. Either approach would require a potentially large investment. But the pay back could be very high. The potential value of visualization techniques will continue to grow as the capability for gathering and exploring large amounts of data continues to expand, and the development of approaches that can make those techniques as effective as possible will be well worth the investment.

2.2.3 Issues in Acceptance by the KDD Community

Grinstein et. al. [Grins98] suggest other thorny issues that must be addressed. One is that any benchmarking effort has to be able to produce credible subjective measures of effectiveness and be able to reconcile them with an adequately broad spectrum of objective measures. Another is that the effectiveness, and in turn the broad acceptance and use of the benchmarking enterprise will depend on how well it can support modeling and steer development of improved techniques. The whole enterprise depends on consensus in the visualization and KDD communities to cooperate and participate in the process, and that in turn depends on building up its credibility to produce sensible measures and ultimately more and more effective systems.

3 Proposed Characteristics of an Evaluation Environment

An evaluation methodology for data visualization techniques within the KDD process is different than pure (non-applied) visualization. The visualization community recognizes that good visualizations are those that are designed for the task and domain. Similarly, any specific visualization or visualization technique must be judged in the context of the step in the KDD process and the domain where it is being applied. However, even in the visualization community there is no on-going, comprehensive evaluation effort, so we cannot look to that community for any systematic collection of tasks, data sets, or benchmarks on which to base KDD visualization evaluation.

In order to develop an evaluation methodology, we must, then, develop a taxonomy of tasks and data sets that support evaluation of a specific visualization system or approach in the context of tasks and data sets that have some relevance to the steps in the KDD process and application. We envision an evaluation environment that contains numerous data sets and application-based tasks which feed into a repository of evaluation outcomes and guidelines. This environment would support an ongoing effort to systematically develop benchmark data sets and outcomes against which evaluation methods and sets can be validated and visualization techniques tested. Figure 1 outlines the structure of such an environment.

This approach does require the development of tasks which have outcomes that can be evaluated. While the KDD process has been described in terms of tasks such as data warehousing, target data selection, cleaning, preprocessing, transformation and reduction, data mining, model selection, interpretation, etc., the granularity of these processes is to large to be useful in defining such tasks. In the sections that follow, we attempt to define a set of testing criteria, and then we describe a preliminary set of lower level tasks that we think have been useful in prototyping an evaluation environment.

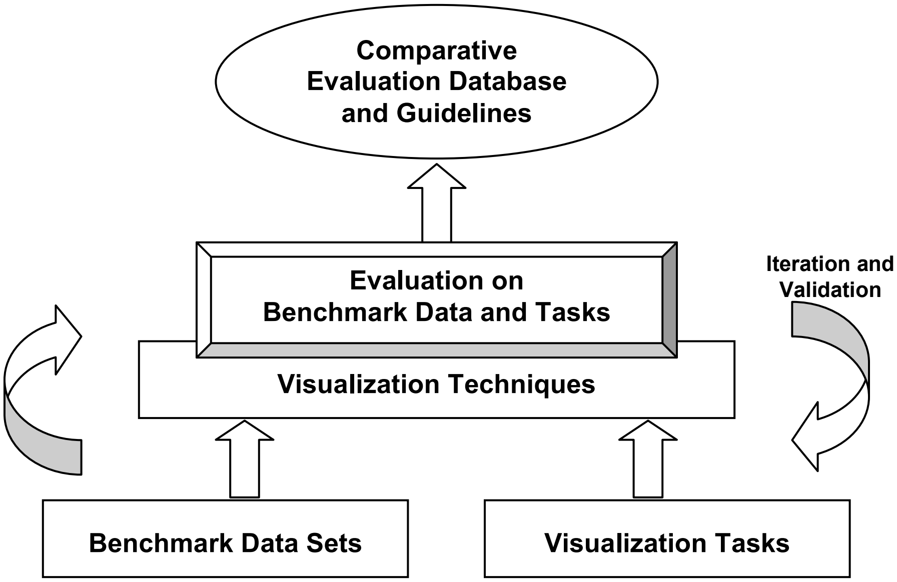


Figure 1 Evaluation Environment for Visualization

3.1 Basic Testing Criteria and Measures

In this section we discuss some basic features of measurement in the context of visualization which could possibly lead to an evaluation methodology that allows for controlled, repeatable test and evaluation.

3.2 Basic Testing Criteria

Visualization techniques can be judged on a number of criteria across the data with respect to the types and amounts of data that can be handled and with respect to the type and quantity of human interactions it can support. Across the data, these include: scalability, dimensionality, structure, and noise. Across the human interactions, these include various aspects of the techniques' capabilities to support interaction with the system and with the data at various stages in the visualization process. These include degree of interactivity, flexibility, ease of expression, and query functionality. Each of these require some sort of metric assignment so that these features can be compared across visualizations. Furthermore, a systematic, controlled approach is required to take into consideration, not just the algorithms but also the interactive qualities of a particular visualization and the perceptual capabilities of the users.

To summarize, any benchmark testing for visualization in the data mining process needs to address criteria such as:

- Scalability; time to process, time to visualize large amounts of data
- Ease of expressing and integrating domain knowledge
- Dealing with uncertain or incorrect, "dirty" data
- Ease of classification and categorization
- High dimensionality
- Flexibility of visualization
- Query and database functionality, and
- Summarization of results.

Visualization techniques need to be characterized according to a set of features derived from these criteria. Only then can they be evaluated against data sets and associated tasks that explicitly exercise them against these criteria. This approach to "benchmark" testing ensures that the results of evaluations can be compared across different visualization techniques or systems. We are assuming, however, that there has been some control for different user populations and usability of the tools themselves. This can be addressed in several ways. Users must either be trained on the systems, or the demographics of the users, for example, whether novices or experts in the domain, must be controlled for and specified. It should be noted that while we have used some of these criteria informally in the prototype environment, integrating them in a systematic way for use in evaluation is an open problem. We also have not specified the human interaction and perception characteristics needed for collection for the repository and guidelines.

3.3 Measures

There are a number of different types of measures to consider in an evaluation. Technology-based measures look at the degree to which a system can handle data sets of varying sizes. This could be tested with a series of data sets of increasing size. Task-based measures depend on the task for the domain and KDD process. For example, one could measure the output of the task of finding the outliers in a set. These measures must be designed for each task category. User-based measures include items such as time to set up, and run a data set and degree of user satisfaction.

In Section 4, a basic set of measures, such as ability to identify outliers and clusters has been applied in the comparative evaluations done for the prototype. Once again, these have not been formalized in any systematic manner, but are simply used as a proof of concept for the approach to evaluation suggested here. They are based on the known features of the benchmark data sets.

3.4 Common Test Data Sets and Tasks

A key component of this evaluation approach is the construction of test data sets. The data sets alone are not sufficient: they must be accompanied by tasks so that the evaluation measures can be applied. This invites a tradeoff in using synthetic vs. real data. Synthetic data is harder to construct, but the "correct" answers are known. Real data is easier to collect, but it is harder to evaluate performance, because it is nearly impossible to "know" the correct output to a task in any reasonably sized data set.

One idea that has been applied in the TREC conference to address this problem is that of "pooling" results to estimate the correct answers. The "findings" over the course of multiple evaluations could be collected and pooled to create a set of "best" answers. Alternatively, the group that constructs a data set could be assigned to find the answers before the release of the data, but this is quite resource intense for any one group.

4 Implementing a Prototype Evaluation Environment

Any evaluation methodology needs to provide cheap, reproducible metrics-based evaluation methods and tools plus common data sets and tasks. It is difficult to measure across low-level support technology (e.g., database capabilities), visualization capability, user interaction, and data mining component interaction simultaneously.

One solution is to develop some basic test data sets and start with some single component tasks. This can form the basis on which to develop a set of validated measures. Having such data sets and measures should support repeatable experiments. Such collective measures, developed for each system, would allow for comparative evaluation.

Ultimately, such an environment would build up a comprehensive record, composed of results collected over time on different sets and systems, that would eventually yield some guidelines for choosing visualization techniques.

In an effort to formalize a benchmark environment for visualization and data mining, a prototype effort has begun at the University of Massachusetts at Lowell. Several machine learning data sets (primarily from UC Irvine Machine Learning Repository [UCI97]) are used as input to a range of multi-dimensional visualizations. The data sets are ordered by increasing size and complexity. The five visualizations, described in detail in the next section, were chosen for their apparent usefulness in exploring large data sets, are:

- Parallel Coordinates
- Scatter Plot Matrix
- Survey Plot
- Circle Segments
- Radviz

By using specific data set examples with known features, various limitations of the visualizations can be demonstrated. These data sets can also be used to test various data mining algorithms, such as classification or clustering. Most data mining software packages include some of these data sets as examples or demos to illustrate the features of the package. More and much larger data sets will have to be included in a full evaluation environment. The data sets, the visualizations and the Java application used in the analysis can be accessed from [Hoff98].

4.1 Overview of the Visualizations to be Compared

We begin with short descriptions of the five chosen visualization techniques. The examples shown are meant to be representative of the output of these techniques but, for the purposes of this paper, are not meant to be analyzed in detail. In particular, color obviously cannot be used to discriminate among the sample data if you are reading a black and white copy of this paper.

4.1.1 Parallel Coordinates

First described by Al Inselberg [Inse85], Parallel Coordinates are a simple, but powerful way to represent multidimensional data. Each dimension or attribute is represented by a vertical line. The maximum and minimum value of that dimension is usually scaled to the upper and lower points on these vertical lines. An N-dimensional point is represented by N - 1 line segments connected to each vertical line at the appropriate dimensional value.

In Figure 2, automobile data are displayed using Parallel Coordinates, with the American cars represented with red lines, the Japanese cars with green lines and the European cars with blue lines. (Again, note color not observable in a black and white copy of the paper. Red shows up darker; hence the higher weights among the American cars, showing darker lines towards the bottom of the Weight coordinate.)

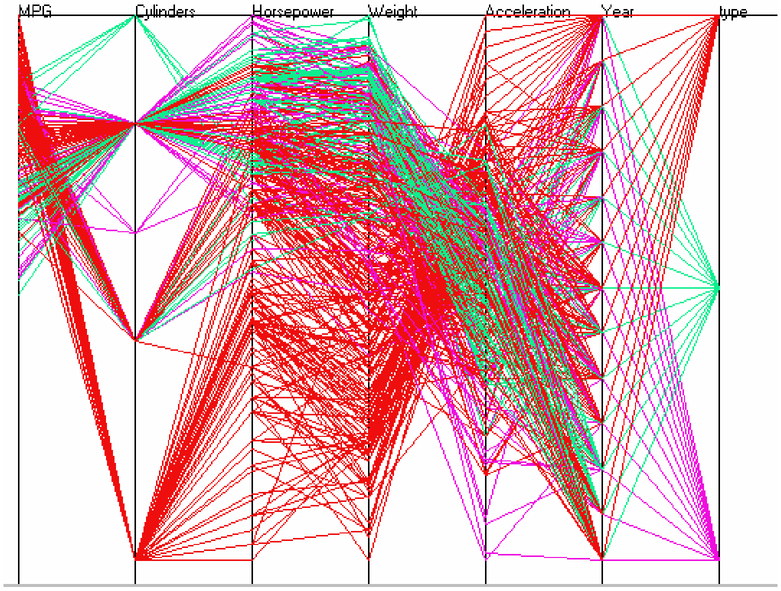


Figure 2 Parallel Coordinates - Car Data Set

4.1.2 Scatter Plot Matrices

Grids of two-dimensional scatter plots are the standard way of extending the scatter plot to higher dimensions. For example, if one has 10 dimensional data, a 10 X 10 array of scatter plots is used to look at each dimension versus every other dimension. This is useful for looking at all possible two-way interactions or correlations between dimensions.

Figure 3 shows a scatter plot matrix of the Iris Flower Data Set.

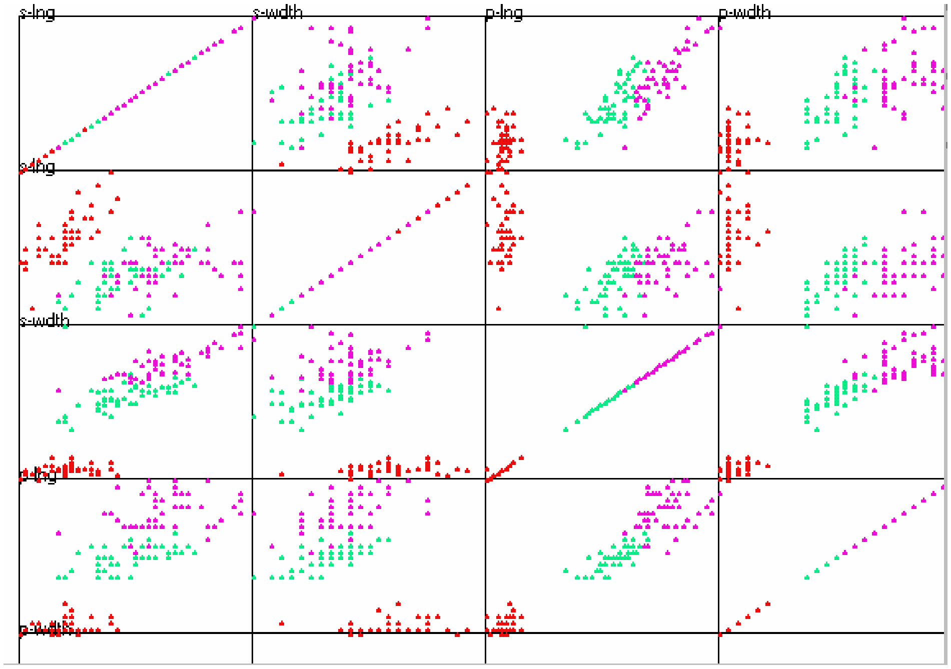


Figure 3 Scatter Plot Matrix - Iris Data Set

4.1.3 Survey Plots

A simple technique of extending a point in a line graph (like a bar graph) down to an axis has been used in many systems such as the Table Lens at Xerox PARC [Rao94]. A simple variation of this extends a line around a center point, where the length of the line corresponds to the dimensional value. This has been called a Survey Plot in the program Inspect [Lohn94]. It is a visualization of N-dimensional data that allows one to quickly see correlations between any two variables especially when the data are sorted on a particular dimension. When color is used for different classifications, a sort can sometimes make it easy to see which dimensions are best at classifying the data.

The survey plot in Figure 4 shows American (red-darkest), Japanese (green-lightest) and European (blue) cars. The data are sorted by cylinders and miles per gallon.

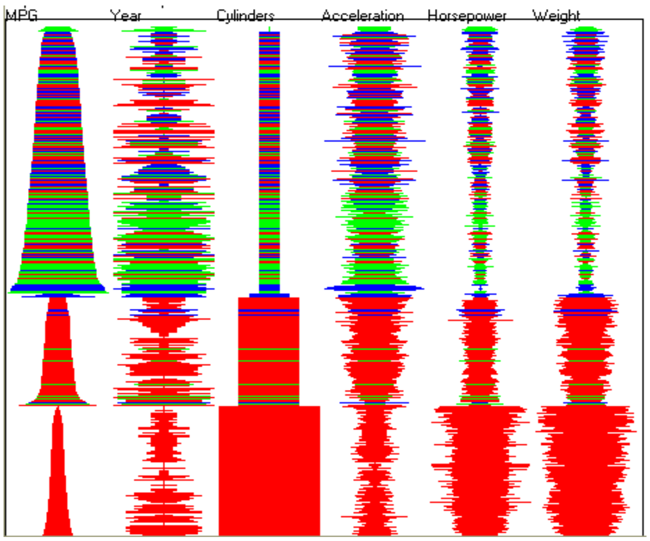


Figure 4 Survey Plot – Car Data Set

4.1.4 Circle Segments

The idea of Circle Segments originated from Ankerst and Keim[Anke96]. It is similar to the Survey Plot. However, the data start from the center of a circle and radiate to the perimeter. A gray scale is used to show the value of a particular dimension, while the class value is colored in pie segments sandwiched around the dimensional values. (This idea of gray scale between class colors is different from the original circle segments.) In Figure 5, a Circle Segments visualization of the Congress Voting Data Set is shown.

4.1.5 Radviz

Spring constants can be used to represent relational values between points [Olse93]. [Hoff97] developed a radial visualization (Radviz), similar in spirit to parallel coordinates (lossless visualization), in which n-dimensional data points are laid out as points equally spaced around the perimeter of a circle. The ends of each of *n* springs are attached to these *n* perimeter points. The other ends of the springs are attached to a data point. The spring constant K_i equals the values of the i-th coordinate of the fixed point. Each data point is then displayed where the sum of the spring forces equals 0. All the data point values are usually normalized to have values between 0 and 1.

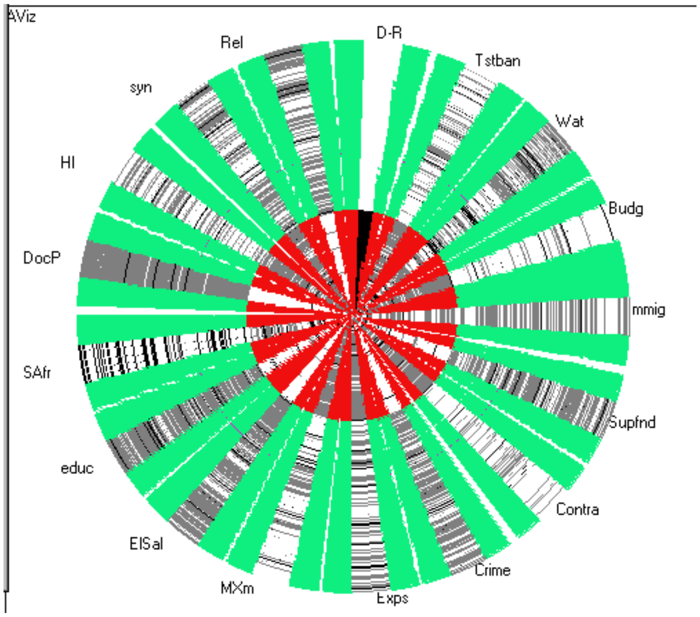


Figure 5 Circle Segments -Congress Voting Data Set

For example if all *n* coordinates have the same value, the data point will lie exactly in the center of the circle. If the point is a unit vector, then that point will lie exactly at the fixed point on the edge of the circle (where the spring for that dimension is fixed). Many points can map to the same position. This represents a non-linear transformation of the data, which preserves certain symmetries and which produces an intuitive display. Some features of this visualization are:

- Points with approximately equal coordinate values will lie close to the center
- Points with similar values whose dimensions are opposite each other on the circle will lie near the center
- Points which have one or two coordinate values greater than the others lie closer to those dimensions
- An n-dimensional line will map to a line
- A sphere will map to an ellipse
- An n-dimensional plane maps to a bounded polygon

In Figure 6, an example of the Radviz visualization is shown using the Wine Data Set. Three types of wine can be seen.

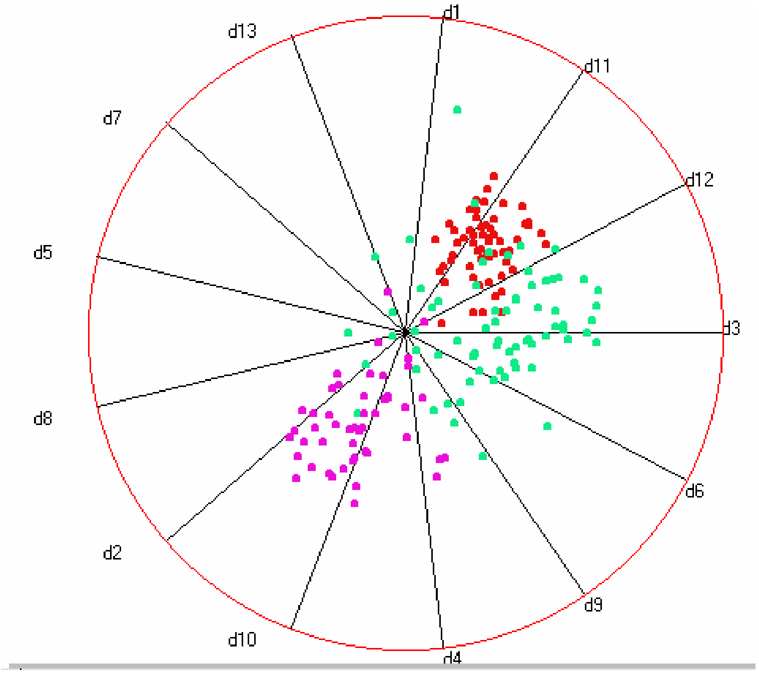


Figure 6 Radviz - Wine Data Set

4.2 Overview of the Data Sets Used in the Comparisons

The ten data sets (Simple Seven, Balloons, Contact Lenses, Shuttle O-rings, Monks problem, Iris Flower, Congress Voting, Liver Disorders, Cars, Wines,) were analyzed with the data mining tool Clementine[Clem98], as well as the five visualizations. The Clementine results and the data sets can be accessed from the web site [Hoff98a]. Two rule-based classifiers (based on Quinlan's C4.5 algorithm), a neural net, and statistical tools were used on the data sets for comparisons with the visualizations.

4.2.1 Description of the Data Sets

All of the data sets except the Simple Seven set, are from the UC Irvine Machine Learning Repository [UCI97]. The first seven-point data set was created to illustrate the features of Radviz compared with the other visualizations, however, it is a useful data set to show the basic features of a visualization. Two of the data sets, the automobile, and the Iris flower data set, were used because of their familiarity. Nearly every data mining package comes with at least one of these two data sets. The other seven data sets were chosen by increasing complexity from the UC Irvine collection. A short description follows with some detail provided later when comparing the visualizations:

- Simple Seven – seven data points used to show point overlap and normalization
- Balloons – data for demonstrating a rule for inflating balloons
- Contact Lenses – data illustrating a complicated rule for prescribing what types of contact lenses to wear
- Shuttle O-rings – the data concerning the Shuttle Challenger failure
- Monks Problems – several data sets implementing rules to test machine learning algorithms. The dataset was designed specifically to be difficult for the algorithms
- Iris Plant Flowers – from Fischer 1936, physical measurements from three types of flowers.
- Congressional Voting Records – Democrat and Republican votes on 16 issues from 1984
- Liver Disorders – a data set that can possibly predict liver disease from blood tests and consumption of alcohol.
- Car (Automobile) – data concerning cars manufactured in America, Japan and Europe from 1970 to 1982
- Wine Recognition – data of 13 chemical attributes measuring 3 types of wines

4.2.2 Complexity of Data Sets

The complexity of a data set depends on many factors which include the number of records, the number of dimensions (or attributes), the cardinality of each dimension, the independence of each dimension, and the underlying function or model which produces the data.

One measure of complexity, the Algorithmic Measure [Chai66], [Kolm65], [Solo64], says the complexity is reduced to the size of whatever algorithm can be used to create the data set. In data mining, this becomes the "model" used to describe the data set, and, finding this model (such as a rule, or a neural net) is often the main problem.

One idea of complexity would then be “how difficult is it to find a rule explaining the data set?” Another definition could be simply the “information content” or entropy of the data set. If certain fields or dimensions can be used to predict other fields, what is the highest classification achieved from a machine classification algorithm? How long and how much memory does it take for certain data mining algorithms to operate on the data set? Answering the last question may be the most practical measure of the complexity of a data set used in data mining. Building a statistical model of the data set has the problem of “the curse of dimensionality” where the joint probability calculation is related to the product of the cardinality for each dimension. Many data mining packages automatically bin continuous fields to reduce the cardinality of each dimension. As one possible measure of complexity, we have included the log of product of the cardinality (PoC) in the data set description.

Although the data sets are listed in order of increasing complexity, there does not seem to be much correlation with how well a visualization performs. Larger data sets would probably start showing a correlation, but this needs to be investigated.

4.3 Comparisons of Different Visualizations for each Data Set

In this section, we compare the visualizations across the ten data sets that were described in the previous section.

4.3.1 Simple Seven Data Set

This is a very simple data set that can be used to illustrate several features of multidimensional visualizations. It was created to show specific differences in various visualizations. It contains 7 instances, 7 classes and 4 numeric attributes. There are four dimensions (Dim1, Dim2, Dim3, Dim4) and Class. Dimension Cardinality is 4,6,5, 4, 7 respectively (number of cases). The PoC is 3360; log of PoC is 8.12.

The 7 points are listed in Table 1.

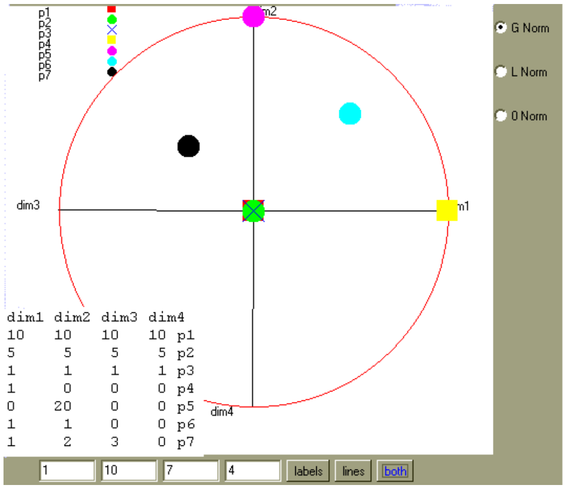
Dim1	Dim2	Dim3	Dim4	Class
10	10	10	10	p1
5	5	5	5	p2
1	1	1	1	p3
1	0	0	0	p4
0	20	0	0	p5
1	1	0	0	p6
1	2	3	0	p7

Table 1 Simple Seven Data Set

The features this data set can illustrate are:

- Global/local Normalization (see later description)
- Point Overlap
- Jittering Features
- Categorical to Numerical Mapping (7 class attributes)

Figure 7 shows the Radviz visualization on the Simple Seven Data Set. Points P1, P2 and P3 lie exactly in the same spot (center) on the display. Jittering the position helps this point overlap problem. Or by using different colors and shapes we can just notice the point overlap problem. In a standard scatter plot display jittering is a standard visualization technique to help show that many points might have the same exact value, or map to the same display point. In the Radviz display, notice that points 4 and 5 lie on the circle, since only one dimension has a non-zero value (the springs pull the data point to the edge). In the current spring paradigm, there is no distinction between the value of 1 and 20 (if no other dimensional values exist). Points 6 and 7 (light blue and dark blue) lie in spots where the combined spring forces are zero.



```
dim1 dim2 dim3 dim4
10   10   10   10 p1
5     5    5    5 p2
1     1    1    1 p3
1     0    0    0 p4
0    20    0    0 p5
1     1    0    0 p6
1     2    3    0 p7
```

● G Norm

○ L Norm

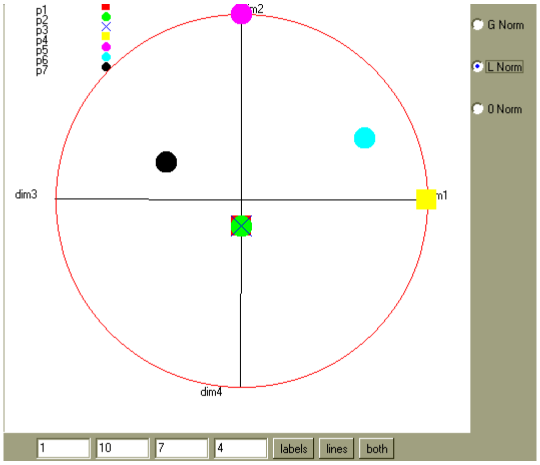
○ 0 Norm

Figure 7 Simple Seven - Radviz – Global Normalization

In Figure 8, the Simple Seven Data Set is shown using local normalization instead of global normalization. Local normalization means that each dimension is scaled from its maximum and minimum to between 1 and 0. Global normalization scales all values from an overall max and min to values between 1 and 0. Clearly this changes the location of the points in this visualization.

The Global/Local normalization problem is clearly seen in all the visualizations, Radviz, Parallel Coordinates, Survey Plot and Circle Segments. (See Figure 7 through Figure 14.)

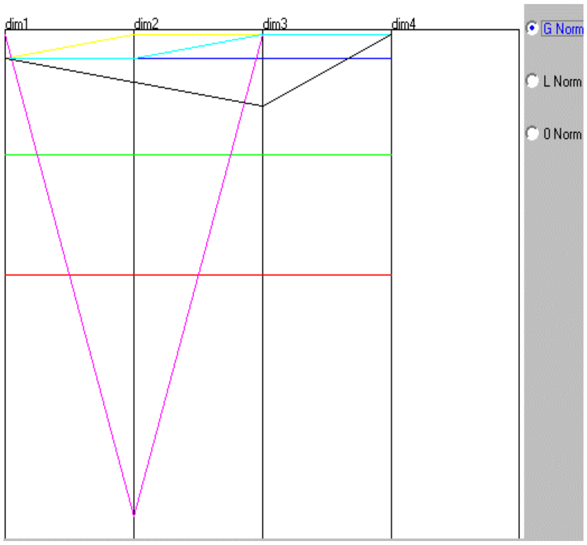
In Circle Segments only the tones of the gray scale change with local/global normalization.



○ G Norm
● L Norm
○ 0 Norm

1 | 10 | 7 | 4

Figure 8 Simple Seven - Radviz - Local Normalization



● G Norm
○ L Norm
○ 0 Norm

Figure 9 Simple Seven - Parallel Coordinates – Global Normalization

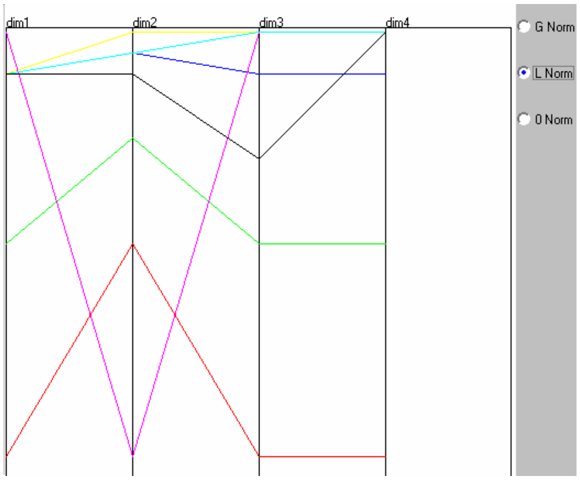


Figure 10 Simple Seven - Parallel Coordinates – Local Normalization

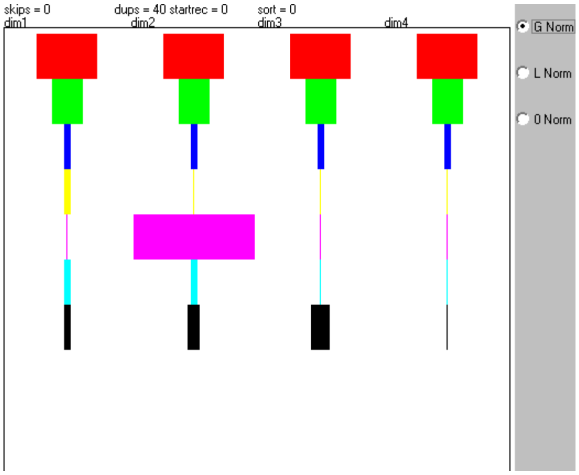
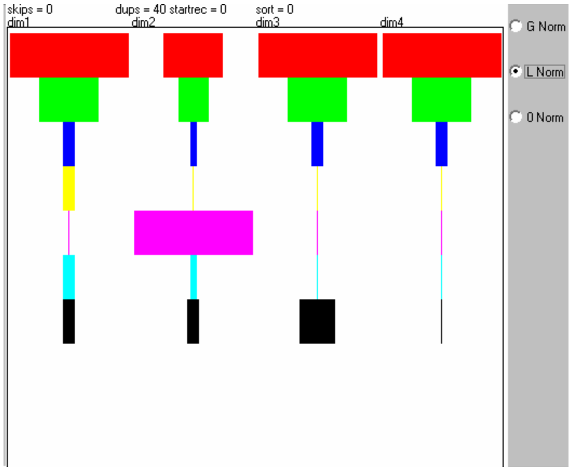

skips = 0 dups = 40 startrec = 0 sort = 0

Figure 11 Simple Seven - Survey Plot - Global Normalization


skips = 0 dups = 40 startrec = 0 sort = 0

○ G Norm
● L Norm
○ 0 Norm

Figure 12 Simple Seven - Survey Plot – Local Normalization

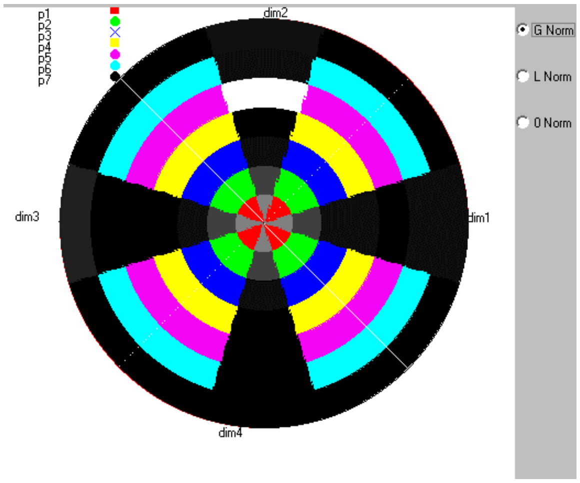
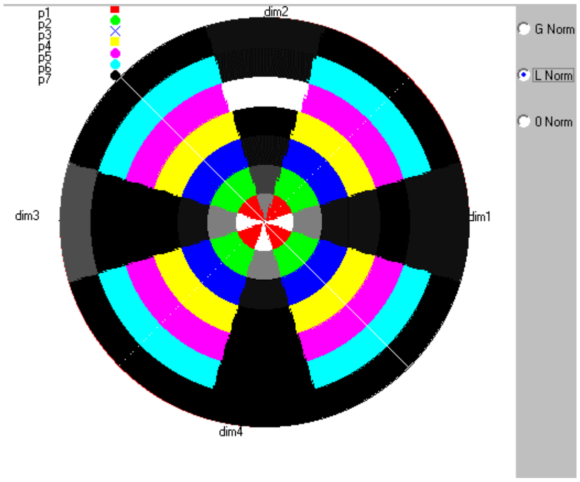

● G Norm
○ L Norm
○ 0 Norm

Figure 13 Simple Seven - Circle Segments – Global Normalization



○ G Norm

● L Norm

○ 0 Norm

Figure 14 Simple Seven - Circle Segments – Local Normalization

When the class dimension is used in the visualization, it can sometimes have a powerful affect. In the next group of Figures (15 to 18), the categorical variable (“p1”…”p7”) is converted to a number and used as part of the visualization. In Radviz, it has the effect of pulling the later points to the "type dimensional radius". Obviously, in many data analysis activities such as clustering by class, removing the "class dimension" in Radviz would be desirable. However, in Parallel Coordinates and Survey Plot, it actually seems to enhance the visual data analysis process.

Thus, in visual data mining, it is desirable to have the ability to remove one or more dimensions from the visualization. Visualizations should have various jittering and normalization options, and visualizing a particular dimension or class attribute by means of color, or as a one of the normal dimensions, is also a desirable visualization feature.

This data set is not applicable to any data mining algorithm, since there is no underlying model of the data. In the rest of the data sets, the visualization techniques will be compared with some data mining algorithms (such as C4.5 and a Neural Net from Clementine).

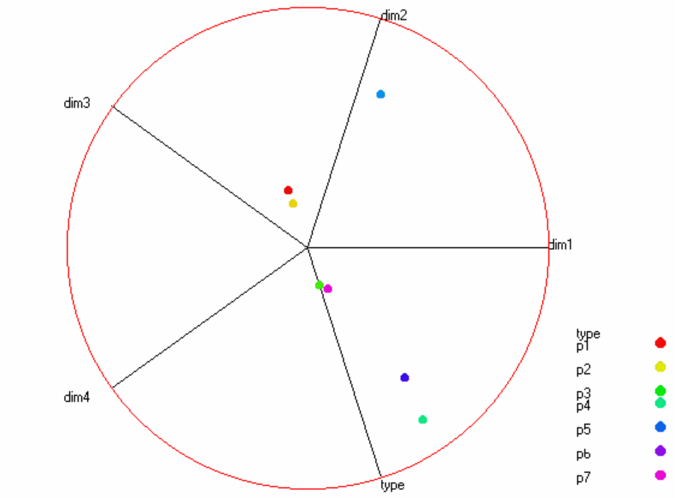


Figure 15 Simple Seven - Radviz -using Type as a Dimension

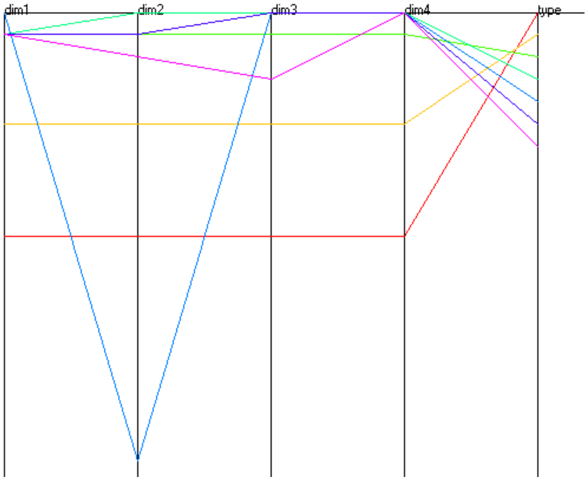


Figure 16 Simple Seven - Parallel Coordinates - Using Type as a Dimension

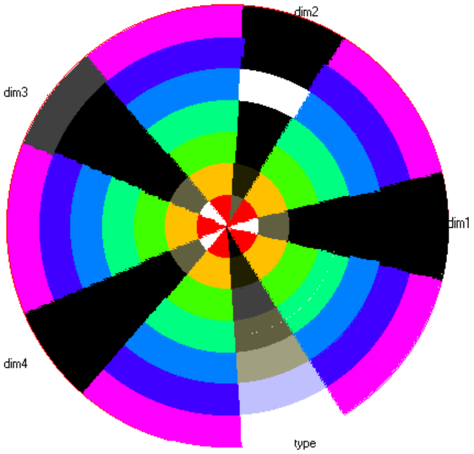


Figure 17 Simple Seven - Circle Segments - using Type as a Dimension

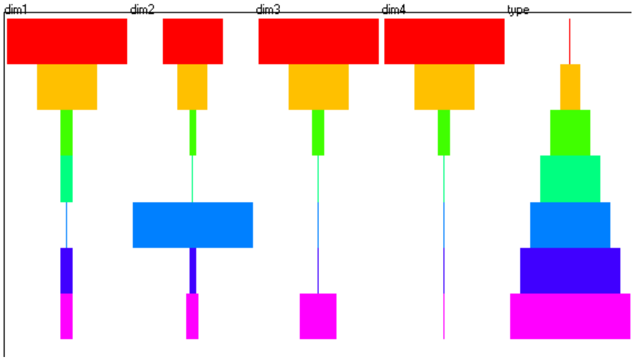


Figure 18 Simple Seven - Survey Plot - using Type as a Dimension

4.3.2 Balloons, Inflated or Not Inflated

The balloon database is a good example of purely categorical data. Each attribute can take on only 1 of 2 values: stretched or dipped; adult or child; yellow or purple; large or small. The class or the value to be predicted is whether the balloon can be inflated or not. There are actually 4 data sets corresponding to 4 rules on how a balloon is inflated. The data set used in the examples, uses the rule: if an "adult" AND "stretched", the balloon is inflated. There are 20 instances (4 repeated), 2 classes, 4 binary-categorical attributes. The data set is #9 from the UCI collection. The dimensions are color, size, act, age, and inflated. The cardinality is 2,2,2,2,20, respectively, with the PoC equal to 640, and the log of PoC equal to 6.46.

The features this data set can illustrate are:

- Properties of an All-categorical Data Set
- Categorical (binary) to Numerical Mapping
- Visual "Rule Discovery"

This data set also illustrates how a categorical dimension should/could be expanded or flattened to a new dimension for each value that the categorical dimension can take. Each dimension can be two values, but when this is visualized, it is not clear what "number" represents yellow/purple or stretch/dip etc. The original Balloon data set is shown in Table 2.

Color	Size	Act	Age	Inflated
YELLOW	SMALL	STRETCH	ADULT	T
YELLOW	SMALL	STRETCH	ADULT	T
YELLOW	SMALL	STRETCH	CHILD	F
YELLOW	SMALL	DIP	ADULT	F
YELLOW	SMALL	DIP	CHILD	F
YELLOW	LARGE	STRETCH	ADULT	T
YELLOW	LARGE	STRETCH	ADULT	T
YELLOW	LARGE	STRETCH	CHILD	F
YELLOW	LARGE	DIP	ADULT	F
YELLOW	LARGE	DIP	CHILD	F
PURPLE	SMALL	STRETCH	ADULT	T
PURPLE	SMALL	STRETCH	ADULT	T
PURPLE	SMALL	STRETCH	CHILD	F
PURPLE	SMALL	DIP	ADULT	F
PURPLE	SMALL	DIP	CHILD	F
PURPLE	LARGE	STRETCH	ADULT	T
PURPLE	LARGE	STRETCH	ADULT	T
PURPLE	LARGE	STRETCH	CHILD	F
PURPLE	LARGE	DIP	ADULT	F
PURPLE	LARGE	DIP	CHILD	F

Table 2 Balloon Data Set

The data set can expanded (or flattened) as show in Table 3.

yellow	purple	small	large	stretch	dip	adult	child	T	F
1	0	1	0	1	0	1	0	1	0
1	0	1	0	1	0	1	0	1	0
1	0	1	0	1	0	0	1	0	1
1	0	1	0	0	1	1	0	0	1
1	0	1	0	0	1	0	1	0	1
1	0	0	1	1	0	1	0	1	0
1	0	0	1	1	0	1	0	1	0
1	0	0	1	1	0	0	1	0	1
1	0	0	1	0	1	1	0	0	1
1	0	0	1	0	1	0	1	0	1
0	1	1	0	1	0	1	0	1	0
0	1	1	0	1	0	1	0	1	0
0	1	1	0	1	0	0	1	0	1
0	1	1	0	0	1	1	0	0	1
0	1	1	0	0	1	0	1	0	1
0	1	0	1	1	0	1	0	1	0
0	1	0	1	1	0	1	0	1	0
0	1	0	1	1	0	0	1	0	1
0	1	0	1	0	1	1	0	0	1
0	1	0	1	0	1	0	1	0	1

Table 3 Expanded (Flattened) Balloon Data Set

The number of dimensions has doubled, however, some visualizations (Radviz) can better illustrate the categorical nature of the data set. Clusters and rules can sometimes be easier to find with this “dimensional expansion”. In Figure 19, a cluster or possible rule seems evident in the Radviz display. In the Survey Plot (Figure 20) with flattening, the rule for inflation can be seen. Hence, “Categorical Expansion” or flattening should be a standard feature of visual data mining. As will be shown in other data set examples, some visualizations (e.g. Radviz) can demonstrate that patterns exist, but other visualizations (e.g. Survey Plot) or data mining algorithms are needed to find the exact rule or pattern. In the data mining program, Clementine, a simple rule and neural net was easily found to classify the data; however, the C4.5 rule found was not quite as simple as it could be.

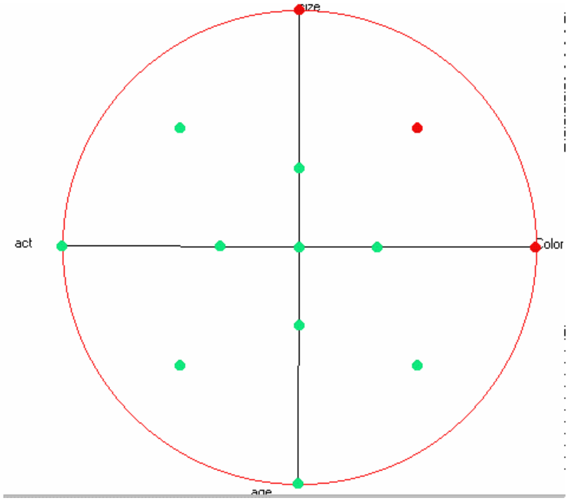


Figure 19 Balloons - Radviz - Inflated points in Red (dark)

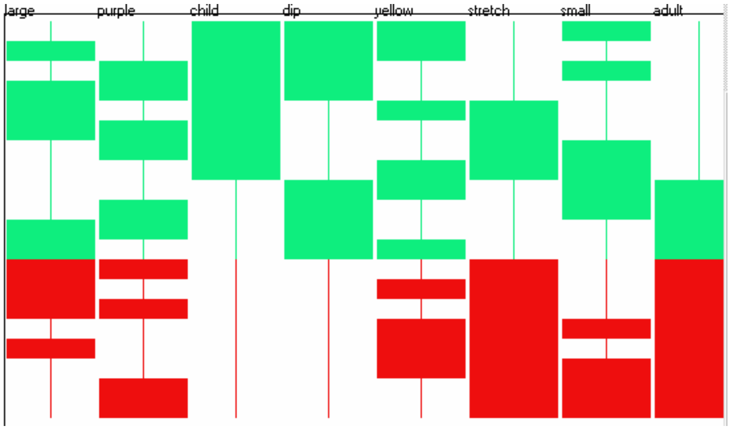


Figure 20 Balloons (Flattened) - Shows Adult & Stretch = Inflated (red-dark)

4.3.3 Contact Lenses

This data set has some complicated rules on prescribing whether a person should wear hard, soft or no contact lenses. The description of the database does not list the rules. There are 24 instances, 3 classes, and 4 discrete attributes. The data set is #52 from the UCI collection. The dimensions are age, prescription, astigmatic, tear production rate, and class (hard, soft, no). The cardinality is 3, 2, 2, 2, 3, 24 (cases) respectively with the PoC equal to 1728 and the log of PoC equal to 7.45.

The mapping of categorical data for each dimension is as follows:

3 Classes
1 : the patient should be fitted with hard contact lenses,
2 : the patient should be fitted with soft contact lenses,
3 : the patient should not be fitted with contact lenses.

1. age of the patient:	(1) young, (2) pre-presbyopic, (3) presbyopic
2. spectacle prescription:	(1) myope, (2) hypermetrope
3. astigmatic:	(1) no, (2) yes
4. tear production rate:	(1) reduced, (2) normal

The features this data set can illustrate are:

- Properties of a Categorical Data Set
- Categorical (binary & tertiary) to Numerical Mapping
- Partial Visual "Rule Discovery" from a complicated rule

Using the Survey Plot visualization (with appropriate sorting) it is fairly easy to find a few rules:

1. If the tear production rate is reduced, do not prescribe contact lenses
2. If the patient is astigmatic, then prescribe hard or no contact lenses.

With the Radviz Visualization, and using random dimensional layout, one can find some non-linear clustering of the three classes (hard, soft, no). (See Figure 21.)

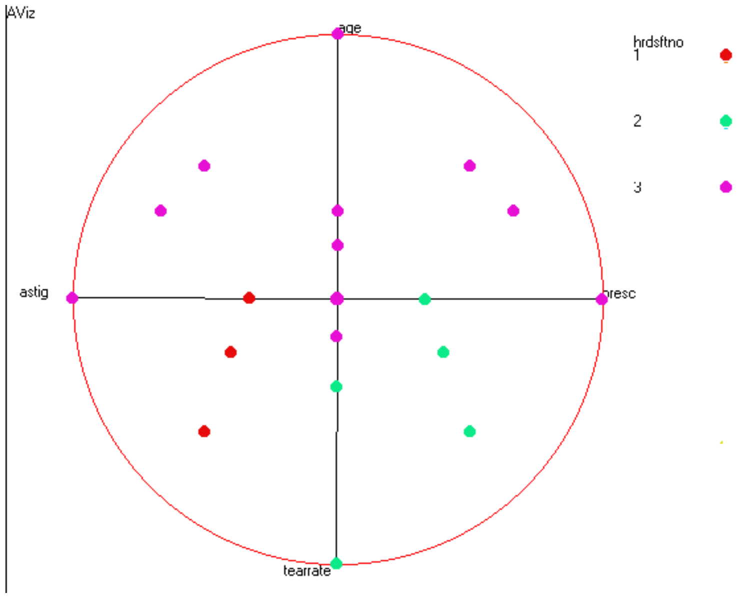


Figure 21 Contact Lenses - Radviz – (The pattern suggests some rules are present)

However, the documentation says there are 9 rules covering the data set. Clementine's neural net and C4.5 only achieved accuracy of 73 and 81% with simple default settings.

The original data set maps the categorical dimensions to numeric values (probably for some "numerical" data mining algorithms). When the "categorical" dimensions are expanded (flattened), the visualizations become more meaningful.

In this data set, the Radviz visualization hinted at the classification rule, and the Survey Plot visualization came closer to finding the rule. However, machine learning (C4.5) did best at finding this complicated rule.

4.3.4 Shuttle O-rings

This is the infamous data set concerning the Shuttle disaster. Does the data set allow one to predict the failure of the O-ring? It contains 23 instances, and 5 numerical attributes (only 4 with different values). The data set is #81 from the UCI collection. The dimensions are: number of O-rings, number of O-rings w/ thermal distress, launch temperature, leak check pressure, and flight number. The cardinality is 1, 3, 1, 6, 3, 23 (cases) respectively with the PoC equal to 3312 and the log of PoC equal to 8.11. The features this data set can illustrate are:

- Simple Regression Prediction on 1 Variable
- Outlier versus Part of the Model
- Trying to Make Predictions from Too Little Data

In various visualizations (Parallel Coordinates, Survey Plot, Radviz - see Figure 22 to Figure 26), it is easy to see a correlation with lower temperature and an increase in number of O-rings under thermal distress.

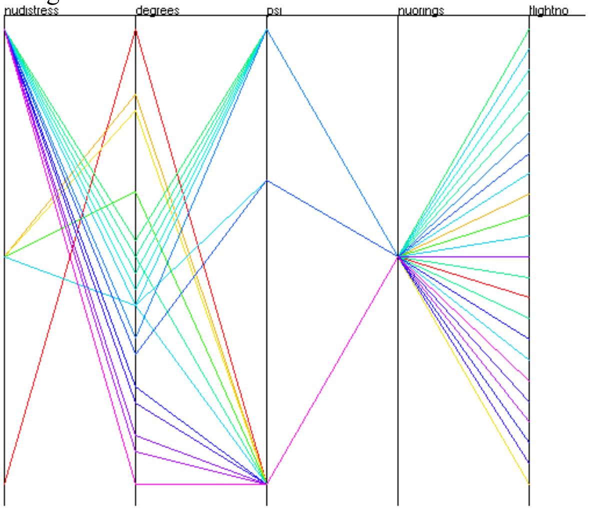


Figure 22 Shuttle O-rings - Parallel Coordinates

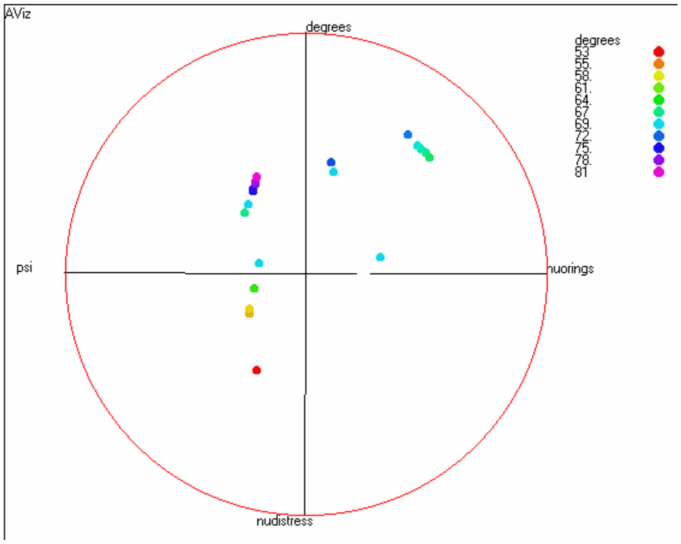


Figure 23 Shuttle O-rings - Radviz

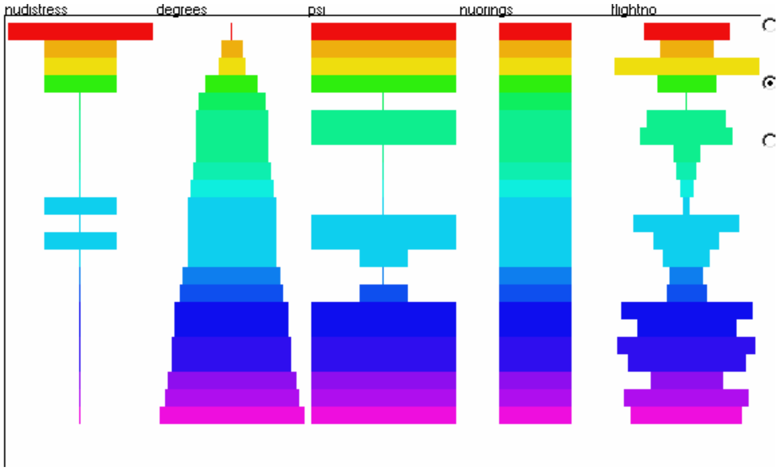


Figure 24 Shuttle O-rings - Survey Plot

4.3.5 Monk's Problems (monk1 - training data set)

This data set was specifically created to test induction algorithms and has sometimes been encoded as monks wearing 6 different articles of clothing with various colors. There are 24 instances, 1 class (0,1) and 6 nominal attributes. The data set is #65 from the UCI collection. The dimensions are class (0,1), a1, a2, a3, a4, a5, and a6. The

cardinality is respectively 2, 3, 3, 2, 3, 4, 2, and 124 (cases) with the PoC equal to 107136 and the log of PoC equal to 11.58.
The dimension information is:

class: 0, 1
a1: 1, 2, 3
a2: 1, 2, 3
a3: 1, 2
a4: 1, 2, 3
a5: 1, 2, 3, 4
a6: 1, 2

The features this data set can illustrate are:

- Visual "Rule Discovery"
- Properties of a Categorical Data Set
- Categorical to Numerical Mapping

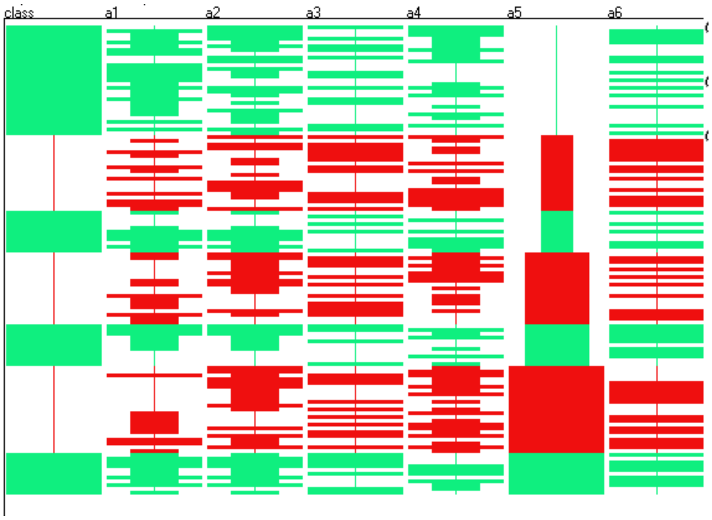


Figure 25 Monks Training Set 1 - Survey Plot – rule on (green –light)

There are actually three data sets, which implement 3 different rules. Each data set contains a training and test set. In Figures 25 to 27, we are looking at the 1st training set. The rule for the first data set can be found in a Survey Plot visualization (Figure 25) where the attributes are sorted by A5 and then the class value. It is clear that the rule (green values) is when A5 is at its smallest value or when the values of A1=A2. This rule was difficult to find visually. However, in some layouts of Radviz, it was hinted that a rule might involve A5 and A1&A2. (See Figure 26.) The rule was also evident in some layouts of Parallel Coordinates. (See Figure 27.)

The simple default values of Clementine (C4.5 & NN algorithms) found a rule and net that were only 92% and 97% accurate, and the C4.5 rule was more complicated than A5=1 or A1=A2. To design a visualization that would help one easily find such rules is a challenge.

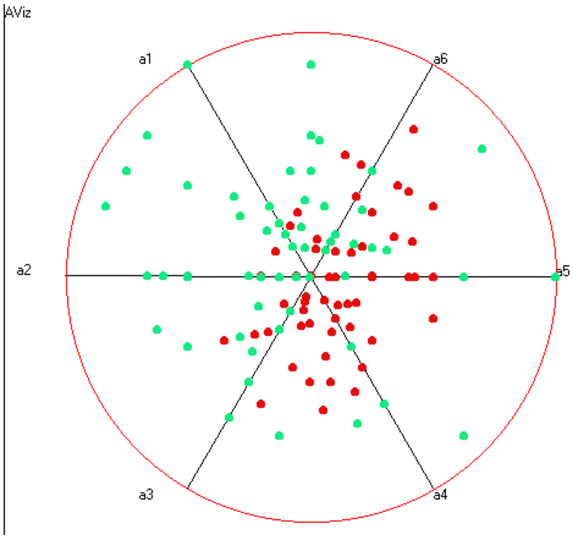


Figure 26 Monks Training Set 1 - Radviz

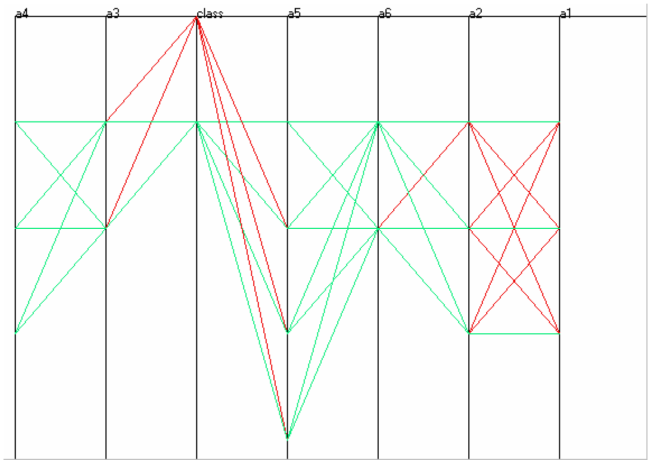


Figure 27 Monks Training Set 1 - Parallel Coordinates

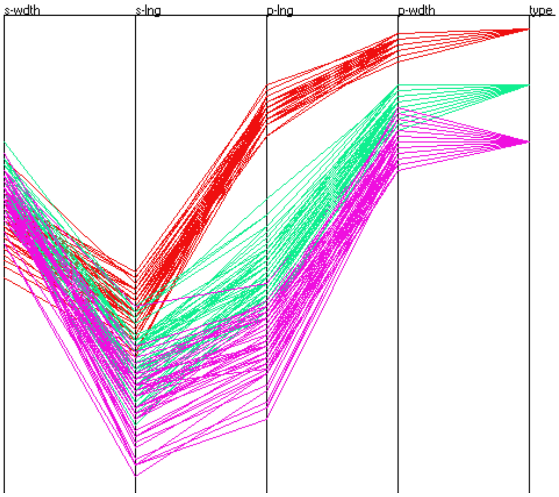


Figure 28 Iris Flowers - Parallel Coordinates

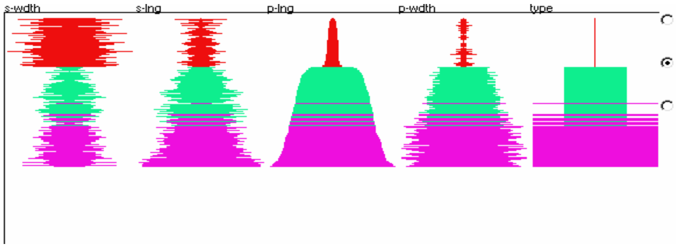


Figure 29 Iris Flowers - Survey Plot

4.3.6 Iris Plant (Fischer 1936 - Flowers) Database

The Iris Database is perhaps the most often used data set in pattern recognition, statistics, data analysis, and machine learning. The task is to predict the class of the flower based on the 4 physical attribute measurements. There are 150 instances, 3 classes, and 4 numeric attributes. The data set is #46 from the UCI collection. The dimensions are: class (Setosa, Versicolour, Virginica); sepal-length; sepal-width; petal length; and petal-width. The cardinality is 35, 23, 43, 22, 3, and 150 (cases) respectively, the PoC is equal to 342688500, and the log of PoC equals 19.65. One class is linearly separable from the other two, but the other two are not linearly separable from each other.

The features this data set can illustrate are:

- Cluster Detection
- Outlier Detection
- Important Feature Detection
- Find Class Clusters

In most of the visualizations, one can see the three clusters of flower types, and in many of them (see Figures 28 and 29), it can be seen that petal-length and petal-width are very good discriminators of the three classes. Several points could be considered "outliers" and show up clearly in several visualizations, as in the Scatter Plot Matrix of Figure 3.

4.3.7 Congressional Voting Records (Republican or Democrat)

This is a data set, which many people can relate to easily. The data set is the voting record of Democrats and Republicans on 16 issues in 1984. There are 435 instances, 1 class and 16 nominal (categorical) attributes. The data set is #106 from the UCI collection.

The dimensions are:
1. Class Name: (Democrat, Republican)
(values for 2 through 17 are y, n, absent)
2. handicapped-infants
3. water-project-cost-sharing
4. adoption-of-the-budget-resolution
5. physician-fee-freeze
6. El-Salvador-aid
7. religious-groups-in-schools
8. anti-satellite-test-ban
9. aid-to-Nicaraguan-contras
10. mx-missile
11. immigration
12. synfuels-corporation-cutback
13. education-spending
14. superfund-right-to-sue
15. crime
16. duty-free-exports
17. export-administration-act-South-Africa

The cardinality is respectively 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3 and 435 (cases) with the PoC equal to 37450647270 and the log of PoC equals 24.35.

The features this data set can illustrate are:

- Cluster Detection
- Outlier Detection

- Important Feature Detection (for class distinction)
- Find Class Clusters (specific clusters that separate classes)
- Usefulness of Circle Segments
- Difficulties of PC, SP Matrix

Most issues segregate Democrats and Republicans to a certain extent, and this was seen in the visualizations. Radviz points out some interesting outliers based on combinations of issues. The Survey Plot with sorting on each issue can show which individual issues predict political parties best. The Circle Segment Visualization showed which issues went mostly according to party lines. (See Figure 5 - mostly dark or mostly white segments). These type of categorical (y/n/?) dimensions pose difficulties for some types of visualizations (Parallel Coordinates, Scatter Plot Matrix). It is possible that a different encoding method could make them more useful. Clementine algorithms had prediction accuracies greater than 95% and could quickly list the best discriminators. An interesting question is whether it is possible for someone trained in the various visualizations to predict classification accuracies.

4.3.8 Liver Disorders (Bupa Medical Research)

This data set is concerned with factors which may contribute to liver disease. The first five attributes are blood tests which are thought to be sensitive to liver disorders that might arise from excessive alcohol consumption.

There are 345 instances (male patients), 2 classes (1,2), 6 numeric attributes. The data set is #54 from the UCI collection. The dimensions are mcv (mean corpuscular volume), alkphos (alkaline phosphotase), sgpt (alamine aminotransferase), sgot (aspartate aminotransferase), gammagt (gamma-glutamyl transpeptidase), drinks, and type (class). The cardinality is 26, 78, 67, 47, 94, 16, 2, and 345 (cases) respectively, with the PoC equal to 6627313854720 and the log of PoC equals 29.52

The features this data set can illustrate are:

- Outlier Detection
- Difficulties of Visual Data Mining

This seems to be the most difficult data set in which to discern any patterns. The description seems to imply that that the seventh attribute (dimension) was a selector on the data set (liver disease or not). The documentation implies that only “drinks > 5” seems to correlate with anything. Visually, this seems difficult to observe. The Scatter Plot Matrix in Figure 30 seems to show that clustering the red and green points (different types) is difficult. Clementine’s data mining tools seem to be able to discriminate better than 70% (see web page [Hoff98a]), however, the “drinks” attribute does not seem to be a factor based on the neural net sensitivity analysis.

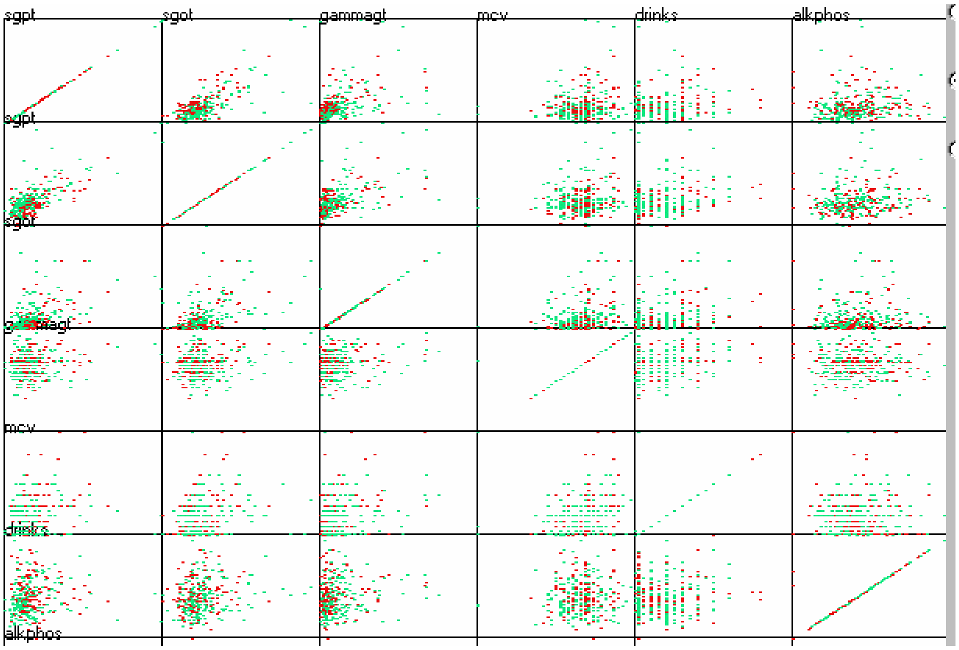


Figure 30 Bupa Liver Disorders - Scatter Plot Matrix

4.3.9 Car Data Set (Auto-Mpg data)

This data set is included in many data mining and visualization packages. It has been modified from the original CMU Statlib Library. Five instances have been taken out because of missing values. The original problem was to predict the miles per gallon for a type of car. The different characteristics of American, European and Japanese cars from 1970 to 1982 are demonstrated in this data set.

There are 393 instances, 7 attributes, 6 numeric attributes, and 1 categorical. The data set is #5 from the UCI collection. The dimensions are MPG, Cylinders, Horsepower, Weight, Acceleration, Year, and Type. The cardinality is 128, 5, 93, 346, 95, 13, 3, and 393 (cases) respectively with the PoC equal to 29986086124800 and the log of PoC equal to 31.03.

The features this data set can illustrate are:

- Outlier Detection
- Cluster Detection
- Class Cluster Detection (type of car)
- Important Feature Detection

In many visualizations one can see the clustering of American cars with increased horsepower, weight, cylinders and acceleration. The Japanese cars have high MPG, low weight, smaller number of cylinders, and lower acceleration. The European cars have more intermediate values, but seem to have the best acceleration. (See Figures 2 and 4.) This is an excellent data set to show a wide range of facts and features using the visualizations. There are several versions of this data set. The one used in Figures 2 and 4 have only 6 dimensions and 1 class attribute. It is interesting that the data mining algorithms cannot seem to classify car type at much better than 70% accuracy for American, Japanese or European.

4.3.10 Wine Recognition Database

Three types of wine are characterized by 13 (continuous) chemical attributes.

There are 178 instances, 3 classes (1,2,and 3), and 13 numeric attributes. The data set is #110 from the UCI collection. The dimensions are class, and 13 unknown continuous variables. The cardinality is 3, 126, 133, 79, 63, 53, 97, 132, 39, 101, 132, 78, 122, 121, and 178 (cases) respectively, with the PoC equal to 1.80e+028, and the log of PoC equal to 65.07.

The features this data set can illustrate are:

- Outlier Detection
- Cluster Detection
- Class Cluster Detection (type of wine)
- Important Feature Detection (which help predict type of wine)

Several visualizations (Scatter Plot, Parallel Coordinates and Radviz) show that many of the 13 dimensions can approximately separate the classes of wines. Some features can discriminate all 3, and some just 2 of the three types of wine. Circle Segments again quickly shows which features are good discriminators (Figure 31) of the 2 and 3 types of wine. From the data set description the three wine types are 100 % separable, but this is not easy to show using standard visualizations (Radviz, Figure 6). Possibly a projection (Radviz or Grand tour) could show a better linear separation. The Neural Net in Clementine had predicted accuracy of 100%. However, the C4.5 algorithm only had 94.4%, using cross validation. This data set demonstrates that statistical and classification algorithms are needed for a full data mining analysis.

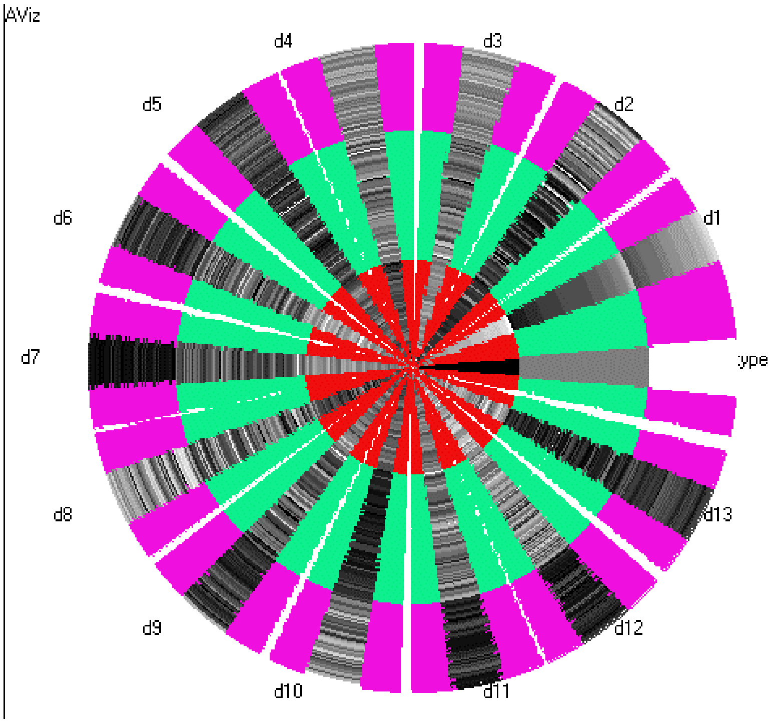


Figure 31 Wine Data Set - Circle Segments

4.4 Summary of Results

In Tables 4 through 8, we provide a summary of the results of the visualization comparisons. Each table represents one of the visualization techniques evaluated on 9 of the data sets (2 of which where flattened for an additional 2 sets). Each column represents one of the 7 features or “tasks”. A “Y” in a column signifies that yes, the visualization can be used to detect that feature satisfactorily. A blank signifies a “no” or “Not Applicable”. The data is a rather fascinating overview of strengths and weaknesses across an interesting set of visualization, data sets and tasks. For example the Survey Plot is clearly superior to the other visualizations in finding the exact rule or model. Circle Segments is rather specialized for finding important features. The charts provide not only a powerful way of comparing this particular set of visualization techniques, but also, and much more important for the point of this paper, they illustrate a potentially very powerful general tool. That general tool could help researchers to gain broad insights regarding strengths and weaknesses of different types and classes of visualization techniques. It can form a basis for developing new techniques and models, and for guiding the evolution and improvement of visualization technology.

Table 4 Parallel Coordinates

TASK DATA SET	See Outliers	See Clusters	Find Class Clusters	See All Important Features	See Some Important Features	See Possible Rule/Model	See Exact Rule/Model
Balloons							
Balloons-flattened							
Lenses							
Lenses-flattened							
O-rings	Y	Y	Y	Y	Y	Y	
Monks1-training		Y		Y	Y	Y	Y
Iris	Y	Y	Y	Y	Y	Y	
Congress							
Liver	Y	Y					
Cars	Y	Y	Y		Y	Y	
Wine	Y	Y	Y		Y	Y	

Table 5 Radviz

TASK DATA SET	See Outliers	See Clusters	Find Class Clusters	See All Important Features	See Some Important Features	See Possible Rule/Model	See Exact Rule/Model
Balloons		Y	Y			Y	
Balloons-flattened		Y	Y		Y	Y	
Lenses		Y	Y		Y		
Lenses-flattened		Y	Y		Y		
O-rings	Y	Y	Y	Y	Y	Y	
Monks1-training		Y					
Iris	Y	Y	Y			Y	
Congress	Y	Y	Y		Y		
Liver	Y	Y					
Cars	Y	Y	Y		Y	Y	
Wine	Y	Y	Y		Y	Y	

Table 6 Survey Plot

TASK DATA SET	See Outliers	See Clusters	Find Class Clusters	See All Important Features	See Some Important Features	See Possible Rule/Model	See Exact Rule/Model
Balloons				Y	Y		Y
Balloons-flattened				Y	Y		Y
Lenses					Y	Y	
Lenses-flattened					Y	Y	
O-rings	Y			Y	Y	Y	
Monks1-training				Y	Y	Y	Y
Iris	Y	Y	Y	Y	Y	Y	
Congress				Y	Y		
Liver							
Cars				Y	Y	Y	
Wine				Y	Y	Y	

Table 7 Circle Segments

TASK DATA SET	See Outliers	See Clusters	Find Class Clusters	See All Important Features	See Some Important Features	See Possible Rule/Model	See Exact Rule/Model
Balloons							
Balloons-flattened							
Lenses							
Lenses-flattened							
O-rings							
Monks1-training							
Iris				Y	Y		
Congress				Y	Y		
Liver							
Cars				Y	Y	Y	
Wine				Y	Y	Y	

Table 8 Scatter Plot Matrix

TASK DATA SET	See Outliers	See Clusters	Find Class Clusters	See All Important Features	See Some Important Features	See Possible Rule/Model	See Exact Rule/Model
Balloons							
Balloons-flattened							
Lenses							
Lenses-flattened							
O-rings	Y	Y	Y	Y	Y	Y	
Monks1-training							
Iris	Y	Y	Y		Y	Y	
Congress							
Liver	Y	Y					
Cars	Y	Y	Y		Y	Y	
Wine	Y	Y	Y		Y	Y	

5 Future Work

We believe that test and evaluation methods can contribute significantly to the development of the next generation of information exploration and KDD tools. We hope that the feedback from researchers and developers who study this paper will serve to guide us in future work to expand and enrich a much needed environment that we have only been able to illustrate here. With help from the research community and industry we would like to develop the taxonomy of visualizations and some benchmark data sets.

We have designed an architecture to support task/feature-based benchmarking. A system is evaluated on a set of tasks and data sets, based on KDD/visualization process tasks and representative data. For each system, a capability matrix can be formed. As evaluations are performed for many systems, a technology matrix can be created, charting algorithms vs. features.

Acknowledgments

This paper is the result of discussions by the working group on Requirements for Benchmarking that was formed during the workshops on Issues in the Integration of Data Mining and Data Visualization held at the 1997 Conference on Knowledge Discovery and Data Mining and at the 1997 Visualization Conference. The authors would like thank those who participated in the discussions held at the workshops on this topic: Mark Levenson, Mohammed Shahbaz Anwar, Ken Collier, and Donald Sautter for their insightful comments.

6 References

[Anke96] Ankerst, M. , Keim, D. A., Kriegel, H. P. Circle Segments: A Technique for Visually Exploring Large Multidimensional Data Sets, IEEE Visualizationí96 Proceedings, Hot Topic, San Francisco, CA, 1996.

[Chai66] Chaitin, Gregory J.. On the length of programs for computing finite binary sequences. Journal of the ACM, 13(4):547-569, October 1966.

[Clem98] Clementine. http://www.isl.co.uk/clem.html

[Fayy97] Fayyad, U., Editorial, Data Mining and Knowledge Discovery, Vol. 1, Issue 3, pp. 237-239, 1997.

[Gibs69] Gibson, E. J., *Principles of Perceptual Learning,* New York, Appleton-Century-Crofts, 1969.

[Grins97] Grinstein, G., Laskowski, S., Rogowitz, B., and Wills, G., Panel: Information Exploration Shootout Project and Benchmark Data Sets, IEEE Visualization '97 Proceedings, pages 511-513, Phoenix, AZ, 1997 [http://www.cs.uml.edu/~phoffman]

[Grins98] Grinstein, G., Inselberg A., and Laskowski, Panel: Key Problems and Thorny Issues in Multidimensional Visualization, IEEE Visualization '98 Proceedings, Research Triangle Park, NC, 1998

[Hoff97] Hoffman, P., Grinstein G., Marx, K., Grosse, I., Stanley, E., "DNA Visual and Analytic Data Mining", IEEE Visualization '97 Proceedings, pages 437-441, Phoenix, AZ, 1997 [http://www.cs.uml.edu/~phoffman/viz]

[Hoff98] http://www.cs.uml.edu/~phoffman.

[Hoff98a] http://ivpr1.cs.uml.edu/shootout/vizdatasets/

[IESH97] http://iris.cs.uml.edu:8080

[Inse85] Inselberg A., "The plane with parallel co-ordinates", The Visual Computer 1, pp. 69-91, 1985.

[Kdnu98] http://www.kdnuggets.com/data sets.html

[Kolm65] Kolmogorov, A.N. (1965). Three approaches to the quantitive definition of information. Problems of Information Transmission 1:1--17.

[Lohn94] Lohninger H.: "INSPECT, a program system to visualize and interpret chemical data.",Chemomet. Intell. Lab. Syst. 22 (1994) 147-153

[Mart96] Martin, J., “Beyond pie charts and spreadsheets”, *Computer World*, May 27, 1996.

[NIST97] http://www.nist.gov/itl

[NIST98a] http://www.nist.gov/itl/div894/894.01/slp.htm

[NIST98b] http://www.nist.gov/itl/div894.03/vip.html

[Olse93] Olsen, K.A., Korfhage, R.R., Sochats, K.M., Spring, M.B. and Williams, J.G. Visualisation of a Document Collection: The VIBE System, Information Processing and Management, Vol. 29, No. 1, pp. 69-81, Pergamon Press Ltd, 1993.

[Pick95] Pickett, R. M., Grinstein,G. G., Levkowitz, H., and Smith, S., Harnessing Preattentive Perceptual Processes in Visualization, in *Perceptual Issues in Visualization*, Grinstein,G. and Levkowitz, H. Eds., New York, Springer-Verlag, 1995.

[Rao94] Rao R. and Card S.K., "The Table Lens: Merging Graphical and Symbolic Representations in an Interactive Focus + Context Visualization for Tabular Information", Proceedings of CHI'94, Boston, ACM Press, pp. 318-322, 1994.

[Schi96] Schiffman, H. R., *Sensation and Perception*, New York, John Wiley and Sons, 1996.

[Solo64] Solomonoff R.J. (1964). A formal theory of inductive inference. Information and Control 7:122,224--54.

[Sted97] Stedman, C., “Users want data mining tools to scale up”, *Computer World*, Sept. 1, 1997.

[Swet82] Swets, J. A. and Pickett, R. M., *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*, New York, Academic Press, 1982.

[UCI97] [http://www.ics.uci.edu/AI/ML/MLDBRepository.html]

[Voor97] Voorhees, E. M. and Harman, D. K.*, The Fifth Text Retrieval Conference (TREC-5)*, NIST Special Publication 500-238, National Institute of Standards and Technology, Gaithersburg, MD, 1997.

[TREC97] http://trec.nist.gov

[Wong97] Wong, W., "Study: Data mining market at $8.4B by 2000*", Computer World,* Oct. 28, 1997.

[Yras96] Yrastorza, T., "The big picture", *Computer World*, Aug. 1, 1996.