Fast Robust Visual Data Mining

Ted Mihalisin

Temple University Department of Physics
Broad and Montgomery Streets
Philadelphia, PA, 19122
tmihal@bellatlantic.net

John Timlin

Mihalisin Associates, Inc.
PO Box 3183
Maple Glen, PA, 19002
johnt@egs.phys.temple.edu

Abstract

TempleMVV a system for visually mining very large high dimensional datasets is presented. The system first developed at Temple University's Department of Physics is based on U.S. Patent No. 5,228,119 and utilizes nested dimensions and hierarchical graphics. The system achieves very high performance which is independent of the size of the dataset by utilizing discrete recursive computing to the maximum degree possible. Data involving any mix of continuous or discrete numeric variables and nominal or ordinal categorical string variables can be mined. In this paper we will try to convey some of the types of knowledge that can be mined utilizing human pattern recognition skills when suitable graphic data representations are chosen. These include but are not limited to 2, 3, ..., 10 way interactions; complex correlations which may be linear or non-linear, marginal or highly constrained, over an entire range or any sub-range of one or more variables; anomalously large and statistically significant frequencies for multi-dimensionally non-contiguous cells ("nuggets"); clustering and discriminate function analysis in up to ten dimensions. Recent enhancements to the system allow one to deal with datasets involving thousands of variables (e.g. "marketbasket" data). The system is superior to neural nets, CART, CHAID and clustering algorithms in several respects.

Introduction to Visual Methods

The types of data analysis and knowledge discovery that are included under the banner of "data mining" continues to grow and now includes a host of techniques from totally automated methods to ones requiring human pattern recognition skills i.e., visualization products. Numerous articles have appeared in the last few years which highlighted how data mining can be used for a wide range of endeavors in fields such as genetics, drug discovery, marketing, finance, insurance, telecommunications etc.

TempleMVV is a data visualization and visual data analysis system that is well suited to the task of information discovery in massive high dimensional datasets. As we have noted in prior articles (Mihalisin et al. 1995 and references therein) TempleMVV differs radically from other visualization techniques including scatter-plot matrices (Cleveland and McGill 1988), parallel axes (Inselberg and Dimsdale 1990), trellis displays (Becker, Cleveland and Shyu 1996) and mosaic plots (Johnson and Shneiderman 1991).

These other approaches suffer from a variety of drawbacks when applied to very large high dimensional datasets. Some suffer from extremely slow performance when the number of records reaches ten thousand e.g. scatter-plot matrices, parallel axes and trellis displays. Some namely traditional 3d scientific visualization systems involve intrinsically low dimensional graphs. Some e.g. parallel axes and mosaics plots only allow for human pattern recognition of a limited number of multidimensional structures. Some are limited in their interactive capabilities in that an inordinate number of operations must be performed and graphs remembered in order to look at multidimensional data i.e. scatter-plot matrices. Although all of the techniques listed above are of value for limited data and analysis regimes, TempleMVV enjoys compelling advantages over other visual techniques. In particular TempleMVV is much faster. For very large datasets it can be literally thousands of times faster. Moreover, TempleMVV can be used to find highly conditional correlations and determine complex multivariate models that are far beyond the capabilities of the other methods.

TempleMVV does, however, suffer from a serious drawback. TempleMVV, which is based on U.S. Patent No. 5,228,119 is a new and very different way of thinking about data analysis and visualization. Analysts must learn how to pattern recognize important results such as the presence of a highly conditional correlation, a fourth order interaction etc. TempleMVV contains a rich tool set that allows one to modify the TempleMVV graph in ways that make these patterns visually obvious, but the analyst must learn how to use the tool set and that requires a true understanding of multiple dimensions. Only a small fraction of analysts qualify at this time. This is true whether the analyst's background is marketing or statistics. Statisticians are sometimes the most difficult users to train since they are constantly attempting to relate the technique to standard multivariate statistical methods many of which are seriously flawed by assumptions such as normality or rule out possibilities. For example an overall (for the entire dataset) correlation matrix for a set of variables rules out the detection of highly conditional correlations. Numeric techniques such as cluster and discriminate function analysis are well entrenched even though it is well known that such procedures have severe limitations even in 2 dimensions that are

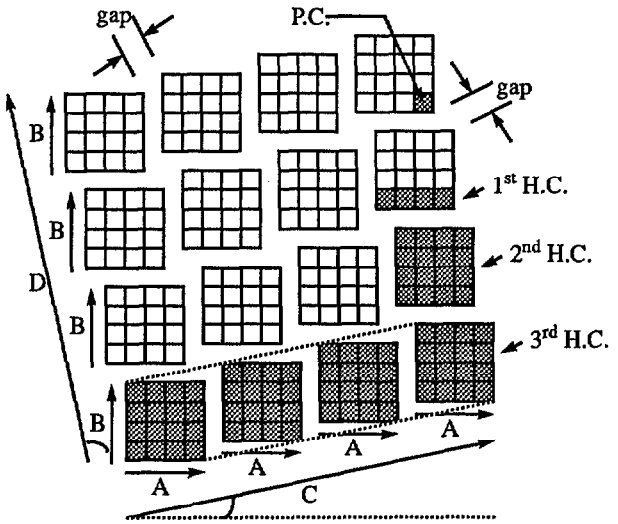

Fig. 1 - A simple 4d space. In general A, B, C, D can have different numbers of bins/values.

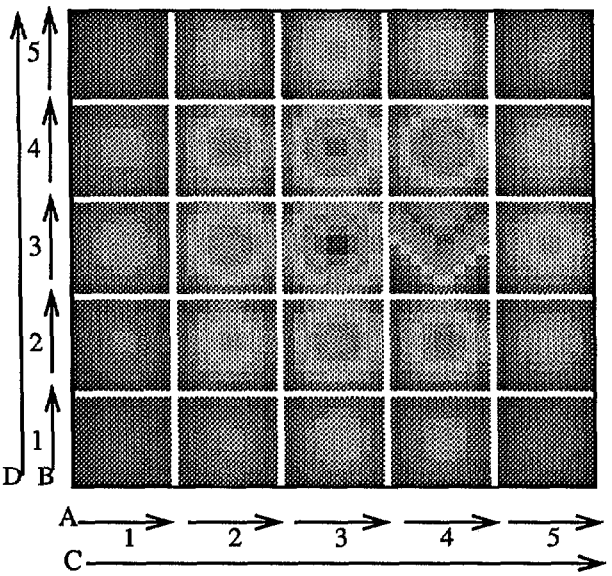

Fig. 2 - Visual detection of a non-linear conditional correlation in 4D.

easily and instantaneously overcome by simply looking at a 2 dimensional graph. The same is true for higher dimensionality.

TempleMVV Basics

TempleMVV has 2 generically different types of computation phases. Namely there are "pre-process" phases which are unattended and have computing times that scale with the number of records. It is during the pre-process phase that all of the cell statistics described in section II) below are computed. And there is the interactive visual data analysis phase wherein all computations required for a new data view are done recursively on the cell statistics. Hence all operations in this "human in the loop" visual data analysis phase are extremely fast. Below we give a 3 step procedure for setting up the TempleMVV system:

I) Form a Discrete Space of Independent Variables or IVs (explanatory variables or dimensions) and Hierarchically Arranged Cells

TempleMVV can handle any mix of categorical, ordinal and continuous variables. But if a continuous variable is chosen to be an independent variable (explanatory variable or dimension) then it must be made discrete by binning. Categorical and ordinal variables are of course discrete at the outset. If there are n_A bins/values for variable A, n_B for B etc. then the total number of primitive cells is $n_A \times n_B \times n_C \times \ldots$ The size and/or color of a symbol such as a vertical bar is going to represent a scalar namely a statistic that represents all the records in the primitive cell.

II) Compute Dependent Variable (response variable or measure) Statistics for All Records in Each Primitive Cell

Once the nature of the primitive cells is defined compute a set of statistics for each dependent variable or DV (response variable or measure) for all records in each primitive cell. These sets of statistics include minimum value, maximum value, sum of values, and sum of squared values. In addition find N_R the number of records (data points) in the primitive cell. These statistics allow TempleMVV to recursively compute means, standard deviation, etc. for collections of primitive cells consisting of cells with several or all values of one or more independent variables (explanatory variables or dimensions) and ultimately are responsible for TempleMVV's extraordinary speed. Fig. 1 shows variable A nested inside variable C and variable B nested inside variable D. Fig. 1 also shows a primitive cell (P.C.) as well as first, second and third hierarchical cells (H.C.).

III) Provide Tools to Alter the Multidimensional Space

These tools allow one to select subsets of bins, combine bins and change the nesting order of IVs (explanatory variables or dimensions). All resulting recomputations are done recursively.

Conditional Non-Linear Correlation

Consider 4 variables A, B, C and D. Shown in Fig. 2 is an MVV graph in which the dependent variable is number or count i.e. the number of data points. There are 9 values for A and 9 for B, 5 for C and 5 for D. A color map (here a gray scale map) is used to indicate the value of N. As N increases the color evolves from red for small N to blue for large N just as in a rainbow. For the gray scale version red is a fairly dark gray (see the region where C = 1 and D = 1 in the graph) then lightens as N increases to intermediate values and then darkens again for high N values (see the center of C = 3, D = 3 region of the graph where N is greatest). Each of the graphs of N(A, B), (i.e. one for each of the 5x5 = 25 values of the C, D doublet) has N maximum in the center and falling off symmetrically away from the center

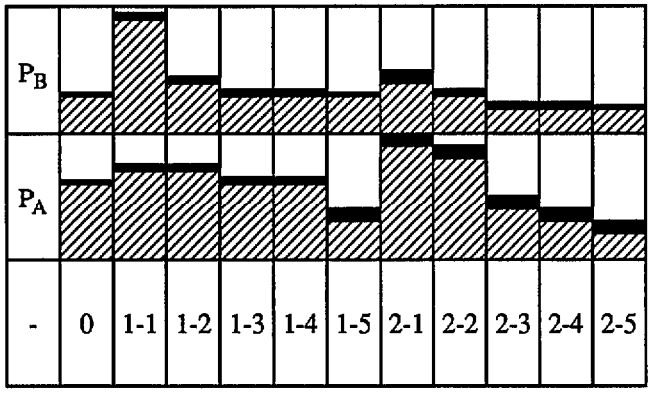

Fig.3 - Probability of purchasing A, B with no constraints and subject to the purchase of 1 or 2 other products.

except for one. For C = 4 and D = 3 there is an easily detectable non-linear (in fact quadratic) correlation of A and B. Visually the non-linear and conditional correlation of A and B is quite obvious and yet a correlation matrix for this data would return zero for all matrix elements. Moreover, even if we took a subset of the data namely the C = 4, D = 3 subset and then evaluated the A, B element of the correlation matrix we still would get a zero result due to the non-linear nature of the correlation. This is one of many examples where conventional multivariate statistical procedures produce invalid results. On the other hand CHAID based products would detect the effect if and only if they go to the appropriate depth i.e. checking for χ^2 in an AB Space for C and D constrained.

Unfortunately space does not allow us to demonstrate how TempleMVV allows one to visually discover clusters that are missed by clustering algorithms.

Extending TempleMVV to 100 DVs

TempleMVV allows one to display multiple DVs as either stacked bars or as individual symbols (bars, circles, boxes or color maps) in subcells. As stated above each DV other than count (number of records) requires that certain statistics for all records in each primitive cell be computed and stored in RAM. This requires 40 Bytes per primitive cell per DV. Hence for 100 DVs, 4KBytes of RAM per primitive cell are needed. In addition other overhead takes 110 bytes per cell for a total of 4.11KBytes per cell for 100 DVs. Assuming say 10,000 primitive cells a system with at least 41.1 MBytes of RAM would be required. Thus PCs and UNIX workstations can be used even when there are hundreds of DVs (measures or response variables) if the number of primitive cells (given by the product of IV bins/values) does not exceed 10,000. There are of course situations in which the number of primitive cells is considerable less than 10,000 which would allow for a correspondingly larger number of DVs.

Extending TempleMVV toThousands of DVs

Certain datasets involve literally thousands of variables. An important example of this is the "marketbasket" problem where one is interested in knowing which products are purchased together. If one is only interested in a handful of general product categories (say 10) then this problem can be addressed by the standard TempleMVV computational engine. In fact only one DV (response variable or measure) is required namely the percent of customers and the IV (explanatory variable or dimension) space would consist of 10 two state variables of form A or $\tilde{A}$ meaning bought A or didn't. The resulting number of cells would be $2^{10} = 1024$.

On the other hand this "IV approach" quickly becomes impossible as the number of products grows. A typical grocery store or auto supply house etc. might sell 10,000 different products and $2^{10,000} \sim 10^{3000}$ would be a very large number of primitive cells. If instead one adopts a "multiple DV" approach one could generate a set of DVs (response variables or measures) for single items, a second set for item pairs etc. However even going to depth 2 i.e. finding the probability or frequency of buying A given that one has purchased B and comparing it to the marginal frequency would imply on the order of $10{,}000^2 = 10^8$ cases.

Thus it would appear that for 10,000 items going beyond depth 3 (i.e. 2 constraints) is not feasible. However, there are in fact practical considerations that greatly simplify the problem and allow one to go to greater depth. The first simplification arises because one is usually not interested in analyzing all 10,000 items only a small subset. Moreover, one is not interested in the leveraging of product A by another B if B is rarely bought. Usually the subset of leveraging items consists of the top n selling products.

Consider the case where there are 100 leveraging items and we are interested in analyzing 10 products. It would appear that one would need to consider $(100)^2$ pairs as constraints for a depth 3 analysis, $(100)^3$ triplets for depth 4 etc. for each of the 10 products. But the number of pairs, triplets etc. are far smaller if one uses the same threshold concept i.e. that only pairs, triplets etc. with purchase probabilities above the threshold should be considered. In fact the number of each cannot exceed 100 and can be considerably less. Moreover, extremely fast algorithms can be employed to locate the 100 or fewer pairs, triplets etc. Hence one has a fast robust system for finding all of the important nuggets!

TempleMVV has been modified for the marketbasket problem. Non-product variables such as the buyer's age, income, gender, education etc. or store related variables become MVV IVs (explanatory variables or dimensions) and the resulting primitive cells define record subsets. Shown in Fig. 3 is a simple example of multiple DVs (response variables or measures) that would go into one primitive cell. They are the probabilities of buying product A and of buying B with and without constraints and their binomial uncertainties which are indicated by the thickness of the horizontal line at the top of each vertical bar. In the first column labeled 0 for 0 constraints the probabilities are computed with no constraints. In the next 5 columns labeled 1-1 to 1-5 are the top 5 probabilities when there is one constraint i.e. that one particular product has also been purchased. In the last 5 columns labeled 2-1 to 2-5 are the

top 5 probabilities when there are 2 constraints i.e. that 2 particular products have also been purchased. Of course which product corresponds to 1-1 or 1-2 etc. differs in general for product A and product B. One can read which products correspond to 1-1 through 1-5 and 2-1 through 2-5 for product A or for product B by clicking on the appropriate row. The 3 most important "nuggets" of information are the high values of P_B^{1-1}, P_A^{2-1} and P_A^{2-2}.For visual clarity only 2 individual products namely A and B are shown in Fig. 3 and only 5 single product and 5 product pair constraints. These numbers can be increased and the constraints can include triplets etc. Remember that Fig. 3 shows the probability of purchasing A with no constraints and with 1 or 2 other products for one particular primitive cell of TempleMVV, that is, for one age group-gender-income bracket etc. Clearly changes in the size and decay rates from 1-1 to 1-5 and from 2-1 to 2-5 of the vertical bars (the probabilities) as well as changes (if any) of the identities of products 1-1 etc. as one moves from one primitive cell to the next would reveal important marketing information.

TempleMVV vs. Other Methods

When comparing data mining tools several issues should be addressed. First, what are the types of knowledge that can be discovered by the tools? Second, is there any guarantee that the tool in the hands of a knowledgable user will discover all relevant information of a given type or at least the most important information? Third, does the tool provide a means for evaluating the importance of the discovery in the context of competing discoveries of the same genre. Fourth, since combinatoric complexity prevents (for non-trivial datasets) the completion of exhaustive search algorithms (in a human lifetime), does the tool lend itself to interactive analysis in which answers lead to new questions which can be addressed by the same tool? Fifth, is the tool fast enough for truly interactive analysis? Sixth, to what depth does the tool allow one to obtain results? This is a vital issue since analysis involving no constraints or the constraint of only a few variables (i.e. "shallow" analysis) can yield very misleading results. Seventh, what are the limitations of the tool in so far as the size and complexity of the data that can be analyzed? In particular can the tool be applied to all the data or must one select a relatively small subset of say 10,000 to 100,000 points and hence severely limit the depth of the analysis? Eighth, can the tool lead to accurate models of the data even in the limit of "complete" complexity as in a fully saturated log-linear model?

On the basis of these issues TempleMVV has decisive advantages over neural nets, CART, CHAID and clustering algorithms. TempleMVV can be utilized to perform all of the types of analysis performed by these tools. Moreover, utilizing human pattern recognition capabilities it can discover data patterns that elude the other tools. Patterns which can be found by one or more of these tools can be discovered far faster using TempleMVV. TempleMVV's fully recursive computing allows it to deal with far more data and hence pursue an analysis of greater depth than the other tools. TempleMVV not only allows one to find the most important information of a particular type but also allows one to view the information in context. TempleMVV's graphical views of the data and sub-second response allow for truly interactive exploratory data analysis and knowledge discovery. Finally, TempleMVV's nested hierarchical graphics and built-in fitting routines allow one to model complex information.

Space does not allow us to detail all of the shortcomings of the other tools. However, some should be highlighted. Neural nets cannot guarantee that one has found the most important nuggets of information and do not provide context or an interactive exploring analysis environment. CART suffers from a lack of depth in that it only looks at one or at most two variables ahead and can lead to very deceptive results. CHAID has four important drawbacks. First, for computational reasons it utilizes limited depth. Second it involves a procedure which aggregates squares of deviations over a variable's entire domain and hence can undervalue striking deviations that occur over sub-domains. Third it does not provide a framework for understanding/modeling data. Fourth it does not provide an interactive analysis environment. Finally, Clustering algorithms are extremely limited in function and in terms of the number of data points they can handle. In addition, they cannot detect inter-penetrating filamentary structures. Moreover, the commonly held view that the use of a complete set of neural net, CART, CHAID, clustering and standard visualization tools can overcome these limitations can be shown to be unrealistic. We believe that these are compelling reasons for analysts to invest the time required to learn the new data analysis methods provided by the TempleMVV system.

References

Becker, R.A.; Cleveland, W.S.; and Shyu, M.J. 1996. The Visual Design and Control of Trellis Display. *Journal of Computational and Graphical Statistics* 5: 123-155.

Cleveland, W.S., and McGill, M.E. eds. 1988. *Dynamic Graphics for Statistics.* Belmont, Calif.: Wadsworth and Brooks/Cole.

Inselberg, A.; and Dimsdale, B., 1990. Parallel Coordinates: A Tool for Visualizing Multi-dimensional Geometry. In Proceedings of the IEEE Conference on Visualization, 361-375. San Francisco, Calif.: IEE Press.

Johnson, B.; and Shneiderman, B.,1991. Tree-Maps: A Space-Filling Approach to the Visualization of Hierarchical Information Structures. In Proceedings of the IEEE Conference on Visualization, 284-291. San Diego, Calif.: IEE Press.

Manly, B.F.J. 1986. *Multivariate Statistical Methods.* London, England: Chapman and Hall.

Mihalisin, T.; Timlin, J.; Schwegler, J.; Gawlinski, E.; and Mihalisin, J. 1995. Visual Analysis of Very Large Multivariate Databases. In Proceedings of the Section on Statistical Graphics, 18-27. Orlando, Florida: American Statistical Association.