05.31.97 TAC.doc

Vladimir Yancher, Alexander Genkin

Visualization of modeling results in Database Marketing

Introduction

Analyzing increasing amounts of data to detect hidden patterns is a major problem facing organizations today. So, there is a surge of interest in data mining and data visualization for business applications. Almost any organization has a large amount of historical data that is available for analysis in machine-readable form. Getting a visual grasp of useful facts or data patterns hidden in this data from different perspectives is a primarily goal of any data visualization technique [1]. In the essence, visualization techniques can be categorized into two basic classes: user guided or system guided. The first one represents a wide range of graphical or spreadsheet data presentation tools with no pattern discovery involved that enable the user to explore data on his/her own. The second one, the system guided visualization, involves results of data modeling as a guide within a data map leading the user directly to the point or the area of his/her objective. Under different circumstances, both approaches can be employed. The following work represents an approach to visualization of modeled marketing data with different level of detail and system supervision.

Problem domain

In the business world the most successful and widespread application of data-mining and visualization technique is DataBase Marketing (DBM). One of DBM major goals is analyzing customer data to identify customer purchase patterns among existing preferences. As a result, marketing strategies created based on these patterns become more effective. As a practical example to apply visualization techniques, we consider a customer response analysis where a maximization of a number of responders is a goal of analysis. CrossZ Software has been developing data visualization approaches belonging to both: user guided or system guided classes. The first, user guided one is based on unique multidimensional data querying technology with the ability to abstract very large data sets and create customized data marts called Query Objects (**QO**) which are small enough to fit on a PC. The user of QO is enabled to confirm or refute his hypothesis in the form of multidimensional queries (defined as a card deck metaphor) in real time (see Figure. 1).

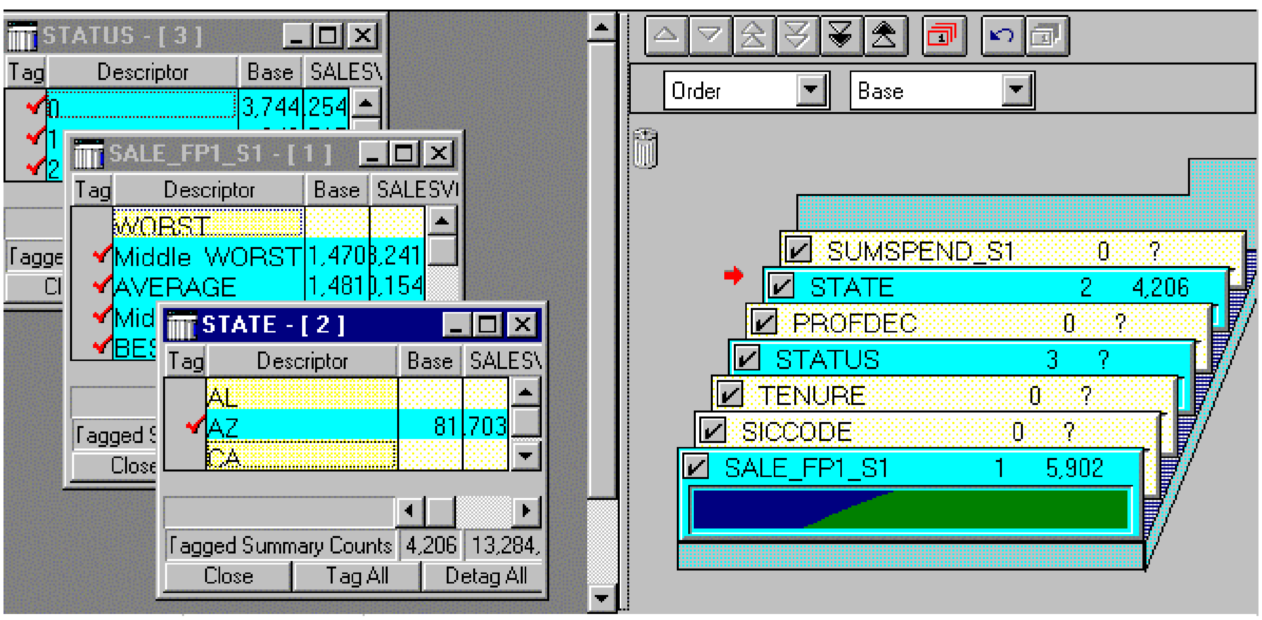

Figure 1. Multidimensional Query Browser

In particular, this QO product was chosen by Allstate Insurance Company, one of the leading insurers of Hispanics. The QO system enabled Allstate to identify how many Hispanic surname customers they insure, to size the market potential and to make a founded decision of how to reach them and target specific segments of them. However, the bottleneck for query analysis is the shortage of hypothesis on the part of the analyst. In the end of this paper we show how to leverage the potential of QO browser with results of data modeling, how to guide the user in his attempt to turn operational data into meaningful information and how to build and visualize multidimensional cubes covering those areas and eventually use the most interesting cubes for querying in attribute space.

Model guided data visualization

To help the user explore very large data sets the CrossZ employs various data modeling techniques such as a linear regression or tree based models such as CHAID or CART [2]. The high-value information patterns uncovered by models are delivered in the form of regression coefficients or easy-to-grasp decision trees. Those patterns are identified based on historical customer records and used to predict customer behavior. Patterns found by a model guide the user through data exploration and visualization process in the following system guided visualization steps.

Cumulative Gains-Charts. The most general form of result visualization (see Figure 2) is a cumulative gains-chart that is aimed to compare modeling techniques and to target customer segments. After building a marketing response model, data patterns are used to rank customers according to their likelihood of response. Subsequently the ranked records are used to build cumulative gains-charts of response vs. base counts. Although the response rate in many cases is often low (e.g. 3%), and it is difficult to predict the class of respondents with high accuracy for all records, but a good model can concentrate the likely responders near the top of the list. The best selected model (see Figure 2) of those used in our analysis yields considerable improvement in the lift of response rate. E.g. the top 5% of the sorted by model score records may contain over 30% of all responders.

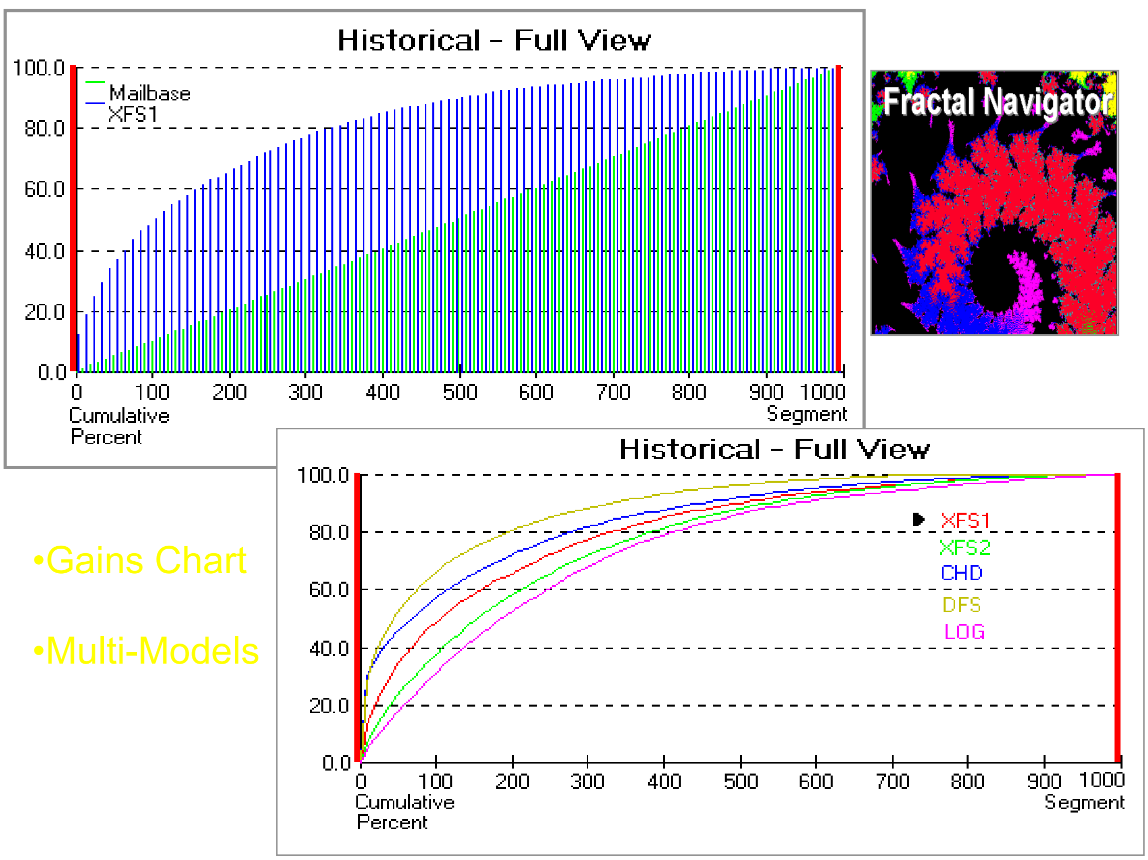

Figure 2. Modeling Cumulative Gains-Charts

To build cumulative Gains Charts shown above we used the Fractal Navigator program developed at CrossZ for business applications to present results of predictive modeling. Fractal Navigator enables the user to pick a desired data segment directly from the graph by adjusting high or low bars and to make an ultimate back-end-selection of customers for the mail list. However, a drawback in this graphical visualization technique is absence of individual characteristics of records to profile people.

Tree Viewing. For the decision tree based models only, direct access and visualization of record patterns in the form of tree nodes are enabled through the CrossZ Tree Browsing utility. The idea is to present a "perception" map that plots data clusters falling into selected terminal nodes in different color based on a response rate. Every terminal node can be considered as a pattern shaped into multidimensional cube. The user can see record attribute values and their positions within a tree branches or as 3-D cubes displayed for records having certain modeling score values (see Figure 3). The drawback of this approach is its inability to display or approximate boundaries of a target segment explicitly to see an overall picture of the target area in the attribute space.

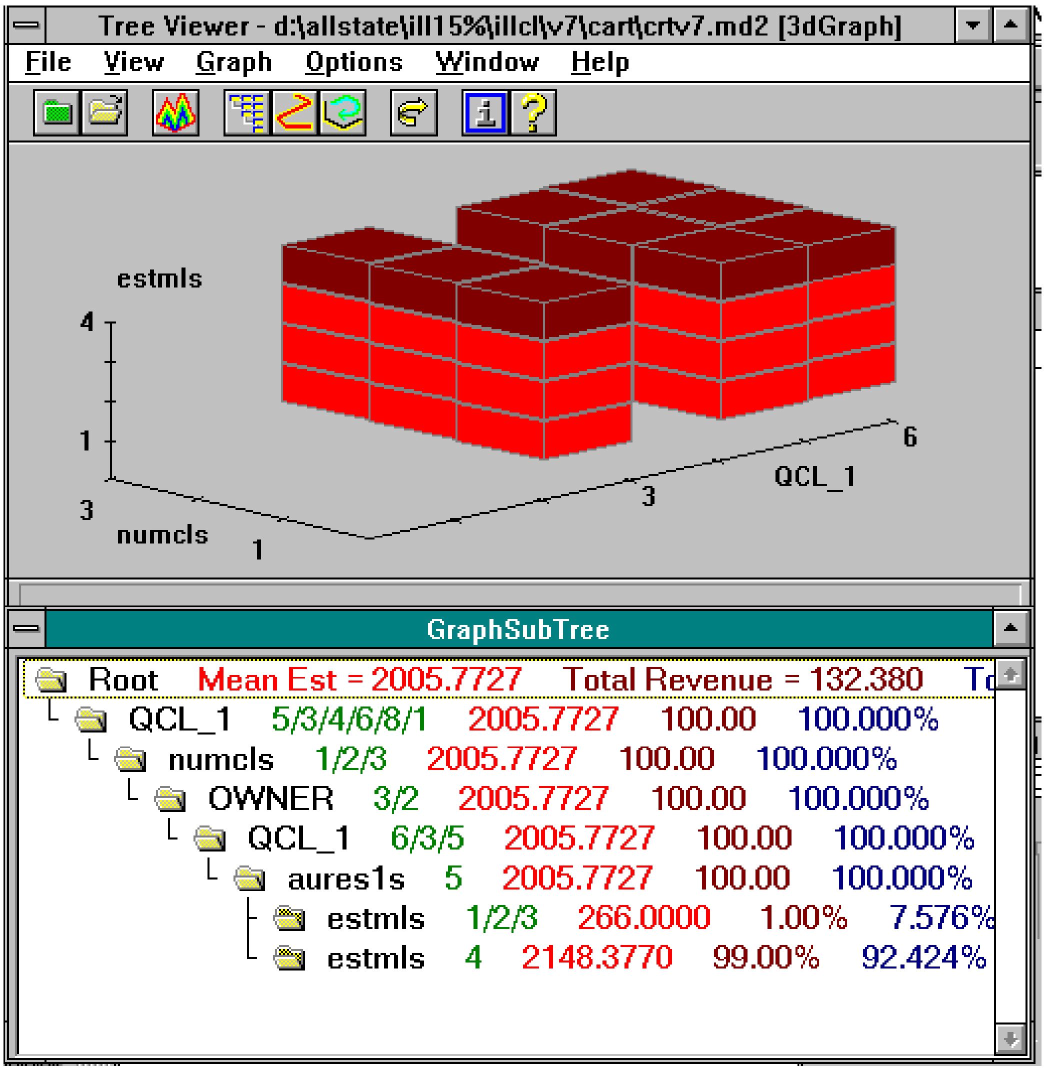

Figure 3. Tree Viewer Browser

Visualization of target segments identified by a model

Detecting and visualizing a target area in the attribute (factors effecting a goal) space can be crucial for making an optimal decision. That is true for any industry dealing with databases. A visual presentation of the target area is usually based on mathematical data modeling having parametric or non-parametric form. Parametric modeling is the most acceptable for engineering analysis way of describing a phenomenon. For example, one can see the spatial area of critical conditions for an earth dam during earthquake. The spatial surface covering the area [3] is drawn based on parametric modeling (non-linear differential equations) of dam behavior during an earthquake in an attribute space containing seismic, constructional and hydraulic factors. However, non-parametric models are a basic form of analysis in business applications.

Many developers are now working on modeling strategies that present results in the form convenient for human comprehension. From the other hand, a lot of existing modeling strategies proved to be very good, and their advantages should not be lost. That's why we are looking for a way to generate presentation for any model, whatever it's mathematical nature is.

"Target Area Cubes" (TAC) approach. The presented below innovative TAC approach to presentation and visual profiling of target business segments is based on approximation of those segments by multidimensional cubes. In general, we assume that any practically valid predictive model of unknown nature can generate a list of scores for all file records based on a selected goal. Thus, this list of scores along with data records contains enough information for a system (model) guided data visualization buildup.

Target Area Cubes are defined in two steps. First, given an arbitrary predictor variable X, define an interval of its values: $[x_1,x_2]$. We call the subset of data elements (both history and prospect) defined by this interval as *primitive selection*. From the logical point, primitive selection is a predicate. The predicate value is true for all data elements (cases or observations) where the variable X value is between x_1 and x_2, false otherwise. If the variable X is categorical, then the interval concept does not apply to it; primitive selection in this case is defined by the subset of states of the variable.

For the second step, assume primitive selections $S_1,...,S_k$ are prepared for several predictors $X_1,...,X_k$. Define the selection in logical terms as a conjunction: $S=S_1\&...\&S_k$. From the geometric point, if all the variables $X_1,...,X_k$ are numeric, the selection S defines the parallelepiped in the subspace defined by these variables; we roughly call it a cube. Thus, defined selection is easily understandable for a business user who is required only to understand the variables involved. This framework serves as a basis for the TAC presentation.

Criteria of presentation quality:

- *readability*: presentation language and other means like visual must be convenient for user comprehension
- *observability*: presentation must be of limited size for user to comprehend with limited time and effort
- *precision*: generally, for a non-trivial model, presentation cannot be perfectly precise and exhaustive if we take into account the two above criteria; the precision criterion limits the deviation of presentation from the real model

Applying these criteria to the TAC presentation, obtain the following. The TAC presentation is *readable* due to its clear logical nature. To make TAC presentation *easy to observe* two conditions must be satisfied:

- limited dimensionality of each cube
- limited number of cubes

The *precision* of TAC presentation is inverse proportional to discrepancies of two types:

- number of targeted objects that belong to none of the cubes
- number of not targeted (alien) objects belonging to some of the cubes, and the difference of the model score value for these objects from the targeted interval

The problem of building the TAC presentation is the trade-off between the observability and precision. It is easy to generate a set of cubes of perfect precision when the number of cubes grows without limitation as well as dimensionality of cubes. It is also important to limit overlap of cubes in the presentation. Overlapping cubes may confuse the user, because it is not possible to sum up metric values assigned to such cubes, like number of objects or summary score value. It gets even more dangerous in case of multiple overlaps.

Solution: Decision Tree

The problem is a kind of discrimination: we are to discriminate between the objects with model score value within the given range and all other objects. Decision trees provide powerful discrimination technique, similar to that of CHAID or CART [2,4,5].

The root node of the tree corresponds to the total data set. Subsequent nodes correspond to segments of the data set. Starting from the root, branching at each node is made as a split of one of predictor variables. Terminal nodes must define segments of types:

- segments that contain (almost) only targeted objects
- segments that contain (almost) no targeted objects

As a result, each terminal node can be easily represented as a defined above selection. It is sufficient to reduce the splits of the same variable, if they occur on a branch, to the single interval. The terminal nodes of the first type constitute the Target Area Cubes presentation. By the way, decision tree methodology produces non-overlapping terminal nodes.

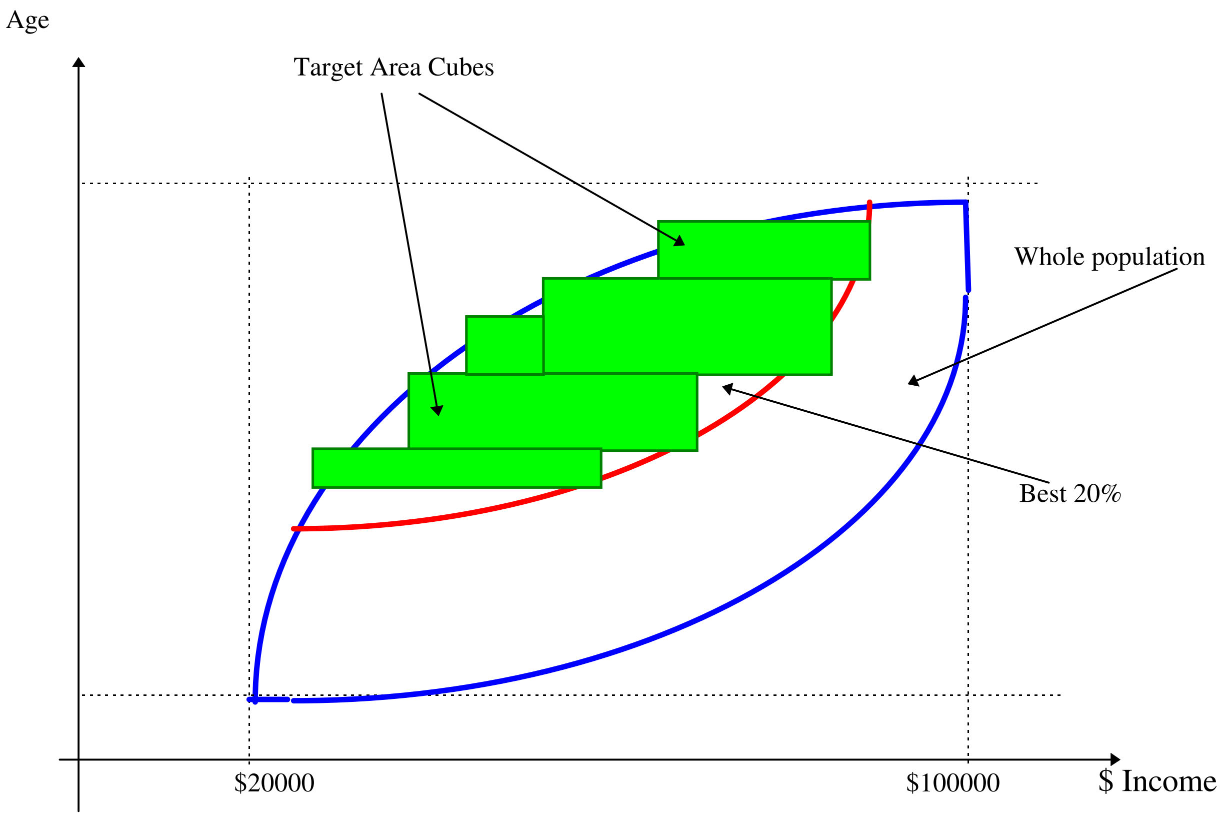

Figure 4. Area Approximation by Multi-Dimensional Cubes

Process of generating the decision tree is supervised by the following rules.

Branching criterion. The segment splitting condition is determined for the current node by the chi-square statistic in a form:

$$C = \frac{(t_1 a_2 + t_2 a_1)^2 \cdot (T + A)}{(t_1 + a_1)(t_2 + a_2) \cdot T \cdot A}$$

where: t_i -- number of targeted objects in the *i*-th split subsegment, *i=1,2*. a_i -- number of objects in the *i*-th split subsegment, *i=1,2*. T, A -- total numbers of targeted and alien objects in the segment

Termination condition. Branching is usually terminated when the segment is homogeneous. In the ideal case, a homogeneous segment contains only targeted or only alien objects. However, it is not realistic to count on pure targeted segments, since cubes can get too small and their number can be too large. That's why we introduce the tolerance function that evaluates how close is the alien object to the target. Denote: *l, h* – low and high limit of the model score, that define the user's target and *L, H* – low and high limit of the model score value.
The tolerance function for a given model score value *s* is defined by the equation:

$$\tau(s) = \begin{cases} 1 & \text{if } l \le s \le h \\ (l-s)/(l-L) & \text{if } s < l \\ (s-h)/(H-h) & \text{if } s > h \end{cases}$$

Then the node will be considered terminal if and only if the mean tolerance in the segment is close enough to 1:

$$1/N \sum \tau(s) \ge 1-\varepsilon$$, where *N* is the number of objects in the segment; ε -- parameter of the algorithm.

The described approach has the following advantages: an ability to visualize results of modeling of any nature; an ability to depict the area and its boundaries as multidimensional figures; an ability to generate a selection in a simple form of cubes covering the area of desire. The resulting picture (Figure 4) illustrates the TAC concept of area approximation by cubes. Each cube that participates in an approximation can be examined in detail with CrossZ QO browser (Figure 1). This program provides the user with "Drill-down" capabilities to analyze goal-affecting variables along with their metrics.

Conclusion

The presented approach to visualization of predictive modeling offers several ways of using model resulting images for decision making. The most advanced and model independent is the TAC approach to visualization. Any examined "slice" of the cube is *readable* due to its logical nature. The user needs nothing but to understand the variables involved in the model. Approximation by cubes is *easy to observe. But two* conditions must be satisfied: limited dimensionally of each cube and limited number of cubes. To maintain satisfactory *precision for approximation* a tolerance criteria must be tuned. This approach is not aimed to substitute the modeling strategies that yield good yet transparent presentation by themselves, neither the specific visualization procedures developed for any particular model types. Instead, it should provide the quick and easy way of visualizing results for non-transparent models with no alternative ways of visualization. Thus, the implementation of TAC approach to data segments visualization adds a new level to understanding insights of data patterns. It enables human ability of visual perception of potentially useful facts. However, in many practical situations Gains Charts and Tree Viewing are more preferable due to their simplicity and easy to use.

Bibliography

1. Grinstein G., John P. Lee., Describing Visual Interactions to the Database: Closing the Loop Between Users and Data, Proceedings of the second SPIE'96 Visual Data Exploration and Analysis Conference, San Jose, p. 93-103

2. Magidson J., Improved Statistical Techniques for Response Modeling. Journal of Direct Marketing. 1988, 2(4), p. 6-18

3. Yancher V. B., Ivaschenko I. N. Quantitative Estimation of Reliability of Earth Dams Built in Seismically Active Areas. Problems of Engineering Seismology. "GIDROPROECT" research work collection. 1988, Moscow, v. 130, p. 103-112

4. Morgan J.N., Sonquist J.A. Problems in the analysis of survey data, and a proposal. J. Amer. Stat. Assoc., 58: 415-434. 1963

5. Breiman L., Friedman J.H., Olshen R.A., Stone Ch.J. Classification and Regression Trees. Wadsworth. 1984