# Case Study: Visualizing Customer Segmentations Produced by Self Organizing Maps

Holly Rushmeier, Richard Lawrence, and George Almasi*

IBM Thomas J. Watson Research Center

## Abstract

We describe a set of visualization programs developed for understanding segmentations of customer records produced by a self organizing map (SOM) algorithm. A SOM produces segments of similar customer records that can then be used as the basis of a marketing campaign. Since the characteristics that each segment will have in common are not specified a priori, visualization is essential to understanding the segment to design specific marketing strategies. Two different styles of visualizations were found to be useful for the two types of observers of the data. Abstract overviews of the entire segmentation were designed for analysts applying the SOM algorithm. Detailed scatterplots of individual records were designed for communicating the results to decision makers specifying marketing strategy.

## 1 Overview

Data mining is the process of obtaining previously unknown information from very large data bases and using it to make effective business decisions [9] . Segmentation is a specific data mining operation used to identify groups of records that are similar based on attributes (or fields) in these records. For example, each record may represent a customer account, with fields such as purchases , demographic data, and other data which characterize this customer. From a business perspective, the development of a single marketing campaign across an entire database may not be effective since different customers will respond differently to a single marketing campaign. A more effective approach is to develop specific messages targeted at subgroups of customers identified by advanced segmentation models. Visualization is an essential aid in understanding the characteristics of newly discovered customer segments to determine the appropriate marketing approach. In this paper we present a series of visualization techniques for understanding the segments produced by the neural network algorithm known as self organizing maps.

Conventional approaches to this problem are "verification-driven" in the sense that the analysis proceeds based on a priori hypothesis. For example, a marketing expert might decide that a campaign combining baby care products and video tapes would be effective. A data base search could be made for customers with high spending on these two types of products. The records retrieved from this search would be the recipients of marketing literature. Two types of opportunities are missed with this approach: there may be more useful trends in spending that the expert has overlooked, and there may be customers who would respond to this campaign who do not *currently* have high spending in these areas.

True data mining, on the other hand, is "discovery-driven." No a priori assumptions are made about the data. The objective is to discover customer segments using a selected set of contributing attributes. The segmentation algorithm finds clusters of records that are similar over all of the spending fields, and finds them in a computationally efficient manner. Discovering segments in this way overcomes the two shortcomings just listed. First, trends in spending patterns that were not hypothesized beforehand are discovered. For example, it may be found that there are some segments with strong spending on video tapes, dairy products and household cleaners that were not hypothesized. Such a pattern would have been computationally expensive to discover using traditional statistical correlations. Second, segments are not exclusively composed of records with high spending in specific areas. In the group that has strong spending in video tapes, dairy products and cleaners there may be records that do not have strong spending in these areas. However, because these other customers have a lot in common with customers that *do* have these spending patterns, they represent a *potentially* profitable untapped market for these products.

There are many possible unsupervised segmentation techniques. In this application we used a neural net technique called self organizing maps (SOMs), originated by Kohonen [5]. SOM's have been used in many applications including organizing textual information and color quantization [8]. The SOM can be viewed as a nonlinear projection from a high-dimensional input space onto a low-order (typically two-dimensional) regular lattice of cells. Such a mapping is often useful in detecting and visualizing characteristic features of the input data, and ultimately in identifying clusters of records that are similar in the original N-dimensional input space.

In the basic SOM algorithm, an N-dimensional reference vector is associated with each cell in the two-dimensional lattice or "map". After random initialization, the reference vectors are updated during a training phase by making repeated passes over the input data set. As each input record is encountered, the reference vector with the smallest Euclidean distance (ie. the reference vector most similar to current input vector) is allowed to adjust or "learn" such that it more closely represents the input vector. A key aspect of the SOM algorithm is that cells near to this "winning" cell on the two-dimensional map are also adjusted in response to this input record. This aspect of the SOM produces a useful topology in that records with similar attributes are assigned to adjacent cells on the feature map. Once the reference vectors have converged, input vectors used in the training, as well as newly available input records, are assigned to segments on the basis of minimum Euclidean distance. Records which fall in the same cell are similar and records in adjacent cells on the map also retain some similarity. The converged reference vector can be interpreted as a prototype vector capturing the characteristics of the records in the segment. More detail on the SOM algorithm and its applications can be found in [5]. In particular we use a new parallelized version of SOM, described in detail in [7].

The difficulty with unsupervised approaches is interpreting the results – examining whether a useful segmentation has been found, and determining what actions should be taken. Simply looking at the numbers is not enough – particularly when there are dozens of segments produced based on potentially hundreds of spending fields. Visualization clearly is useful in organizing and interpreting the results.

Designing an appropriate visualization is complicated by the fact that two *different* types of people are involved in producing a successful outcome – data mining analysts and decision makers. The

*P.O. Box 704, Yorktown Heights, NY 10598, {holly,lawrence,almasi}@watson.ibm.com

0-8186-8262-0/97 $10.00 Copyright 1997 IEEE.

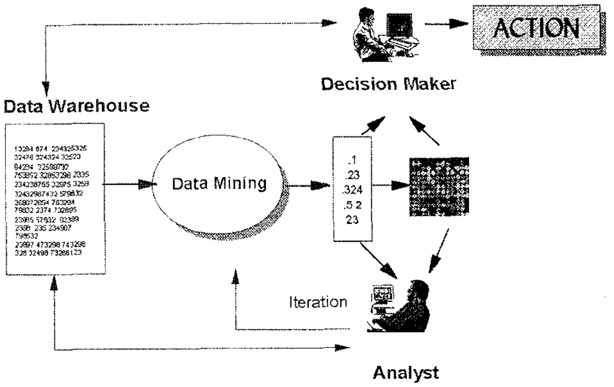

Figure 1: The data mining process involves an analyst who *iteratively* runs data mining algorithms, and a decision maker who takes *action* based on the results.

role of each type of person in the process is diagrammed in Figure 1.

The data mining analyst takes the large quantity of data from a warehouse, selects a mining technique, and organizes the data for input. In the case of SOM, the analyst must decide what fields to include in the analysis, how to scale the values in the field to have appropriate influence on the solution, and how many segments to use. An analyst may go through several iterations to find a good segmentation.

Questions that the analyst asks when looking at a solution are:

**A1: Do the segments have distinctive characteristics?** If all of the segments have similar distributions of the various fields, perhaps too many segments have been specified.

**A2: Is the segmentation unduly influenced by individual fields?** For example, a marketing segmentation might have been strongly influenced by the inclusion of gender as an input field – putting all the men and the women in separate segments with otherwise similar spending characteristics. In this case, the segmentation is being too influenced by this field, and the analyst may choose to rerun the analysis without it.

After a good segmentation has been found, the results need to be interpreted by a decision maker, who will determine the appropriate marketing message and the customers to be contacted.

The decision maker asks the following questions:

**D1: What are the preferences of each customer segment? Are there useful groupings of the segments for a particular marketing strategy?** The decision maker may not want to design a marketing strategy for the top three spending areas of each segment. Instead, she/he may be interest in finding two spending areas of interest across three or four segments.

**D2: How does this segmentation information relate to other data mining results against the same database?** Several data mining operations may be performed on the same database. For example, a telephone promotion may be the marketing method to be used for this project. A separate predictive model may have been built for the purpose of predicting the likelihood that a customer will respond positively to telemarketing. Visualization can be used to combine the results of the predictive model with a segmentation operation.

**D3: How reliable are the models that have been generated by the data mining?** If a predictive model has been built, the decision maker needs some assessment of how effective this model is when used to predict historical (or training) data with a known outcome (e.g. did they respond to a previous campaign?). One way of doing this is to present these records with known outcome in a form that shows how well the model did, and, in particular, how well it did in each of the segments identified via the segmentation model.

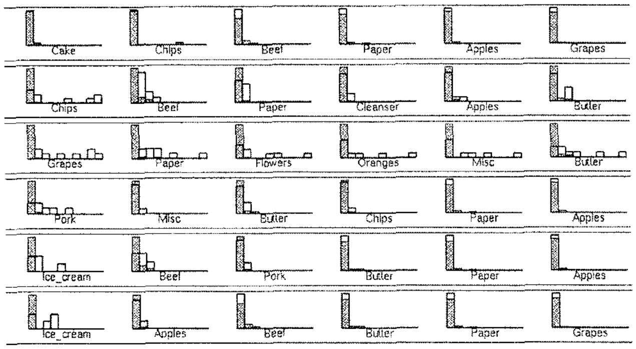

Figure 2: Previous visualizations of segmentations showed each segment as a row of histograms for the most important fields in that segment.

**D4: How can the results of the segmentation be recorded in terms of action?** Unlike scientific visualization, the end result of this visualization is not insight, but an action. In the course of using the visualization the decision maker wants to produce a data file that can be used to take a real world action.

In this application we designed two different styles of visualization to address the different questions posed by analysts (A1 and A2) and by decision makers (D1 - D4.) The application was developed for a specific customer, but the results are disguised here to preserve customer anonymity.

## 2 Visualizations

An existing visualization for segmentations was available for this project from the IBM Intelligent Miner. A typical output is shown in Figure 2. The figure shows a row of histograms for the most significant fields for each segment. The figure is a useful starting point for both analysts and decision makers to understand the characteristics of segments. The display has a degree of interactivity, but it does not have the techniques built into it to address all of the questions discussed above. In particular, because the display was developed for a more general class of segmentation algorithms, it does not show any of the topolological features of the SOM map. The display also does not allow access to individual records.

For analysts, new high level representations showing an overview of the entire segmentation were developed. For decision makers, a representation showing the results of segmentation and a predictive model for the original records was developed. The new programs were developed using IBM Visualization Data Explorer.

### 2.1 Analysts

Since SOM's project multi-dimensional data into 2-D space, an overview of the entire segmentation is logically a 2-D map. SOM visualizations in the form of 2-D maps have been developed for search applications [3]. Major search topics are plotted in 2-D, with a third or height dimension representing N-dimensional distance between the cells on the map. These maps are used to navigate towards articles in a topic area of interest. SOM visualizations in 2-D have also been developed for organizing relatively small data sets (e.g. 100 records, [1].) In this case, the number of records is small enough that each record can be represented by a text glyph. The goal of the visualization is to examine how alike individual objects are to one another. The "PLANES" code in the publicly available

SOM_PAK [6] produces 2D grey-coded maps of any selected field, but this capability is difficult to use to analyze problems with large numbers of attributes.

These previous visualizations are not useful in our application, since we are attempting to characterize groups of records, rather than seek out individual records. However, we use the 2-D SOM segment layout as a starting point, as shown in Color Plates a,b, and c. The squares in the grids represent the cells in the SOM, in this instance 36 cells laid out as a 6x6 rectangular map. We then represent fields within each segment with glyphs that are colored and/or sized to represent average data values for that field in that segment.(In this context a field is a spending category – e.g. butter, cleanser, beer, etc.) This is similar to the pixel-oriented techniques for examining large numbers of items for many attributes described in [4]. However, while the glyphs may be adjusted to be pixel sized to get an overview of the segmentation, they can also be enlarged so that the analyst can zoom in on a segment and select individual glyphs to get more information for segments and fields of interest.

Basic variables that were computed for each field for each segment were the *ratio* of spending to average spending in the population (*ratio1* for the average of all records, *ratio2* for the average over non-zero records only), the *penetration* of spending – the percentage of customers in that segment that spent in that field, the neural *weight* (the weight is the component for this field in the SOM reference vector), the *value* of total spending and the *count* of the number of records. Summary variables were computed for each segment – the *count* records in the segment and the *value* of spending in the segment. Summary information was linked to the main 2-D view by highlighting segments and fields selected in the main view in the summary view.

Three variations of the 2-D map were developed. In Color Plate a, each glyph is colored by one user-selected variable, and its opacity is adjusted by a second. (With a black background, adjusting opacity is equivalent to adjusting the color value. With arbitrarily selected backgrounds, opacity serves to emphasize glyphs with high values and de-emphasize glyphs with low values.) The glyph for each field is in the same relative position in each box for each segment. Changing the variables determining color and opacity reveal different aspects of the segmentation.

In Color Plate a, opacity is determined by penetration and color by the spending ratio2 for 36 segments and 272 fields. The visualization addresses question A1. The segments do have distinctively different spending patterns. The segment in the upper right has one field with a high ratio and high penetration. Nearly everyone in this segment spends in this field, and their total spending is much higher than that of the population average. The upper left shows a segment in which there is above average spending in many areas, and most of the customers in in this group shop in these areas. The other segments show spending less specialized than the upper right, and less uniform than the upper left. Since the red spot in the upper right stands out, the analyst can select it and get the name and statistics for the field. The selection also results in a display of how total spending in that segment compares with other segments, and how the actual dollars spent on the field compares to dollars spent in other fields in that segment.

In Color Plate b, color again is determined by a user-selected variable. The second variable is thresholded, rather than determining opacity. This provides the user the precise control necessary to answer question A2. For example, by using the neural weight as a color and using a penetration of 90 as a threshold, Color Plate b shows the influence of penetration in the determination of neural weights. Relatively high values of penetration resulted in high neural weights – nearly all have a weight above .75. By adjusting the threshold,a fine sense of how this variable affected the result can be obtained.

To further address question A2, the specific fields that were important in a particular segment need to be examined. In the variation shown in Color Plate c the top ten fields for each segment according to a selected variable are shown. In Color Plate c the top fields for ratio2 are shown. In overview mode, colored and sized glyphs only are shown. One large glyph and nine small ones indicates that one field is far more important than the others in the segment. In zoom mode the names of fields are displayed. The user can adjust the size of the displayed labels based on the portion of the map occupying the current window. The display of these fields also helps the analyst to gain an understanding of the segmentation to communicate to the decision maker.

## 2.2 Decision Makers

Unlike the analyst, the end decision maker is not interested in the topology of the SOM or the influence of parameters on variables such as neural weight. Visualizations are needed that lead to a plan of action.

The visualization developed for decision makers is shown in Color Plate d. The visualization is designed to address questions D1 to D4. In the figure is a colored histogram showing the population of the 9 segments that were ultimately used in this case. The colors are assigned by segment number. The large window shows a 3-D scatter plot for a selected segment. There is a report, in the same color as used for that segment in the histogram, of the top five spending fields for this segment. This list is a first step in answering question D1, what are the customer preferences. The numbers by each field are ratio and penetration.

The scatter plot shows all records in the segment, with two of the axes representing spending in two fields. Because absolute value of dollars spent can vary widely, the spending is scaled so that 1.0 equals the average spending in the population for that field, and values are clamped to a user specified multiple of average spending (i.e. all valuess above the specified maximum value are set to the maximum value.) In Figure 3 the left and center columns of images show the values clamped to two different values of average spending. Scaling to the average and clamping prevents outliers from compressing all of the data into the corner of the plot.

To continue to pursue question D1 other segments can be displayed with the same fields. The two rows of images in Figure 3 show the results for two different segments. In this case, these two segments have radically different spending patterns.

Looking for common spending interests alone could be accomplished by a 2-D array of 2-D scatter plots. However, we use a third dimension to address question D2, how do these segmentation results relate to the results of a predictive model for the same database. In this case, the vertical position is score for probability to respond to a marketing campaign. (In the plot, the score for probability is denoted RBF_score, since it was computed using a radial basis function (RBF) technique.)Similar to the other axes, the probability is scaled so that the average probability ($\approx$ .3 here) is scaled to one, hence the values that are larger than one on the vertical axes. The observer can now simultaneously examine preferences for a segment, and their likelihood to respond.

Using three dimensions for displaying probability helps alleviate the "overstrike" problem in looking at the probability scores. If just probability were plotted, the result would be a single line with many points falling on top of one another. By spreading out the values with the plots on the other two axes, the points have been effectively jittered off the line (see [2], p. 106).

Even with the spread over three dimensions, there are still overstrike problems, particularly at the clamped surfaces and edges. In a sense the clamped surfaces and edges are positive, in that they enhance the visibility of the 3 dimensional distribution. To deal with overstrikes, the points have adjustable opacity. By reducing the opacity areas of high density emerge. The images in the right

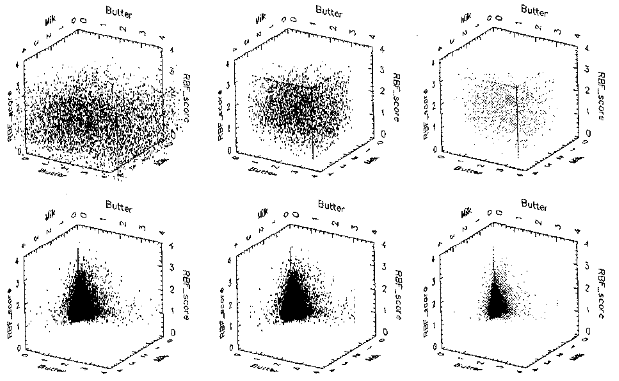

Figure 3: A 3D distribution of points as originally displayed (left), with clamping to a maximum value (center) and with opacity less than one (right) to show the density of points. The two rows show results for two different segments with very different spending patterns.

column of Figure 3 show a clusters of 3-D points with a decreased opacity demonstrating this effect.

The 3-D visualization in can also be used to address question D3, how reliable are the models that have been generated. In the case of the prediction model, historical data of customer response was available. Records were coded red if they had responded in the past, and blue if they had not. An ideal predictor would have all red at the top of the plot, and all blue at the bottom. The display allows the user to inspect the validity of the predictor model segment by segment. In Color Plate d, the display shows that the predictor does a reasonable job for this segment, with most of the red points above the blue points. The 3-D visualization could also be used to test the segmentation model. In practice, a set of records which was not used in training the SOM model would be passed through the net to see how records were distributed into segments. We would expect the segment by segment portraits produced by the 3-D visualization to look similar with new data; significant differences might suggest new trends to customer response that should be analyzed.

Finally, a set of cutting planes for the x,y, and z axes were implemented to address question D4 – how to record the results in terms of an action. The user can adjust the cutting planes to desirable values (e.g. above-average spending and probability of response) and see how many records in the segment fall into the octant defined by the planes. By writing out the field names, cutting plane values and segment id's of interest, the data could be recorded that is needed for a data base extraction to produce a list of customers for a marketing campaign.

## 3 Results and Discussion

There are results of three types from this study – the results of the specific project, insights into visualization techniques, and some insights into broader issues in visualization.

*Results of the Specific Data Mining Project* The visualization programs developed for analysts were found to be successful, and provided insights into tuning the segmentations. The programs are currently being extended to deal with categorical, as well as numerical variables.

The images generated by the visualization program developed for decision makers were successful in communicating results. The interactive program itself though was found to be too complex to appeal to non-quantitative decision makers. We found that spreadsheet level graphics define the level of program complexity decision makers are willing to use. We also developed a spreadsheet application to allow decision makers to examine text-based and graphical summaries of the results.

*Lessons Learned: Visualization Techniques* Two new observations can be made about visualizing 3-D scatter plots from this work:

1.) Comprehending the shape formed by a set of 3-D points is difficult. Allowing the user to adjust the ceiling to which values are clamped can help give the sense of the cloud shape. 2.) Overstrikes are always a problem in scatter plots. Letting the user adjust opacity can help give the sense of where points are very dense and multiple overstrikes are occuring. Both clamping and opacity take some time to explain to users. They are not as intuitive as the use of motion parallax to get a sense of 3-D distribution.

*Lessons Learned: Broad Issues* 1.) For the same problem and the same data, very different representations are required for different observers. In this case quantitatively oriented analysts found abstract representations more useful than the more qualitatively oriented decision makers. 2.) No single picture is adequate. Each display raises new questions. Techniques such as pixel-oriented plots or the Intelligent Miner visualization tool are helpful, but they need to be modified to be interactive to answer further user questions. 3.) Business applications have decidedly different goals from scientific applications. The goal of visualization in business is action, not insight. Pictures need to be embedded in a system that allows the observer to take an action based on what is seen.

# References

[1] BACK, B., IRJALA, M., SERE, K., AND VANHARANTA, H. Managing complexity in large data bases using self-organizing maps. *Turku Centre for Computer Science, Technical Report No 48* (September 1996).

[2] CHAMBERS, J., CLEVELAND, W., KLEINER, B., AND TUKEY, P. *Graphical Methods for Data Analysis*. Wadsworth, 1983.

[3] HONKELA, T., KASKI, S., LAGUS, K., AND KOHONEN, T. Newsgroup exploration with websom method and browsing interface. *Helsinki University of Technology Report A32* (January 1996), 1–13.

[4] KEIM, D., AND KRIEGEL, H.-P. Visualization techniques for mining large databases: A comparison. *IEEE Transactions on Knowledge and Data Engineering 8*, 6 (December 1996), 923–938.

[5] KOHONEN, T. *Self-Organizing Maps*. Springer, 1995.

[6] KOHONEN, T., HYNNINEN, J., KANGAS, J., LAAKSONEN, J. *SOMPAK: The Self-Organizing Map Program Package*, Helsinki University of Technology, http://nucleus.hut.fi/nnrc/som_pak.

[7] LAWRENCE, R., AND ALMASI, G. *In preparation.*

[8] MYKLEBUST, G., AND SOLHEIM, J. Parallel self-organizing maps for actual applications. *1995 IEEE International Conference on Neural Networks* (1995).

[9] SIMOUDIS, E. Reality check for data mining. *IEEE Expert: Intelligent Systems and their Applications* (October 1996), 26–33.