Reprinted from IFPress (2018) 21(4) 3,4

Characterizing Sensory Segmentation using Machine Learning

William J. Russ and John M. Ennis

Background: A revolution is happening in business worldwide as companies increasingly leverage computational and algorithmic power to solve problems that previously might never have even been considered. For example, who would have expected ten years ago that large numbers of people would routinely awaken to countertop virtual assistants turning on their lights and greeting them with music, telling them the news and weather for the day, checking their calendars and messages, and eventually summoning people they've never met to ferry them to their destinations? Yet such actions are now commonplace. Similarly, in sensory and consumer science, advances in computation are reaching an inflection point of supporting behaviors that were previously inconceivable. For example, just a few years ago it would have required a supercomputer to run a full TURF[2] analysis on a dataset of even moderate size. But now such analyses can be run instantly - it is now possible to evaluate the entire space of possible TURF solutions on a standard desktop computer and to find the "best" solution according to a variety of criteria[3]. From a computational perspective, several of our previous technical reports have discussed the ubiquitous sensory and consumer science problem of finding best combinations - this problem may appear in the guise of finding the best combinations of features, benefits, ingredient, images, and/or products in a portfolio[4,5,6].

In this report we turn our attention to a different combinatorial problem, the problem of characterizing consumer segments using the best combination of descriptors for the segments. These descriptors may be demographic, behavioral, psychographic, or could encapsulate any other information known about the respondents. Just as with other combinatorial problems, an explicit search for solutions is intractable. But, just as with other combinatorial problems, recent advances in computing and algorithmic power now allow us to find solutions, this time using a method known as machine learning.

Scenario: You work for a corporation that has recently acquired a company with a brand of boutique barbecue sauces, and you want to offer a wide-reaching and well-positioned portfolio of sauces to the national market. Because the recently acquired company had only conducted a minimal level of consumer research, your understanding of the national market is sparse and you plan to conduct a nationwide category appraisal. You realize that, even with well-defined hedonic segments that allow you to choose your portfolio, you will still need to know to whom to market each product. Historically, you have been restricted to one of two approaches: 1) Consider various pre-defined consumer groups, and look for differences in hedonic responses, or 2) Examine hedonic clusters one consumer variable at a time and see if the distribution of consumers with respect to that variable differs from cluster to cluster. You resolve to find a more comprehensive solution.

Machine Learning: Machine learning is any process whereby software programs, when given more data, improve their ability to perform tasks. Despite the science-fiction sounding nature of the definition, most consumer scientists have already used machine learning in some form — linear regression is one such technique. While regression is a powerful tool, it can be limited by its assumptions[7]. Other common methods for machine learning include support vector machines (SVMs) and artificial neural networks (ANNs). For the purposes of understanding consumer segments another machine learning technique known as "decision trees" is especially well-suited because of the interpretability of its results[8].

Decision Trees: Decision tree algorithms create a flowchart to guide model predictions. These flowcharts are often easy to understand and can handle highly correlated predictor variables and large datasets. While there are several variations of the exact algorithm, the basic approach of creating a tree uses recursive binary partitioning. The algorithm starts with all of the data and then tests each predictor variable for a value that will split the data according to various internal metrics. For categorical variables, the algorithm divides the categories into two subsets. For continuous variables, it searches for a threshold for splitting into two groups. This splitting continues until the algorithm reaches a pre-specified stopping criterion. Random forest methods, which were created to handle stability issues with decision trees, combine results from many trees to create a final prediction which is more stable and usually more accurate, but at the expense of interpretability[9].

Barbeque Sauce Segmentation: You design a central location test (CLT) in which acceptors of barbeque sauce provide a range of attitudinal and psychographic data after evaluating a variety of well-chosen barbeque sauces. With the rise of computerized data collection, having respondents complete a fairly extensive questionnaire after completing product evaluations is feasible. In order to use machine learning, you follow a three-step process to arrive at your consumer characterization.

Step 1 - Create Hedonic Space: You first identify consumer segments based on which products the consumers like and dislike. Landscape Segmentation Analysis (LSA)[10] is a method that creates a hedonic map where consumers are

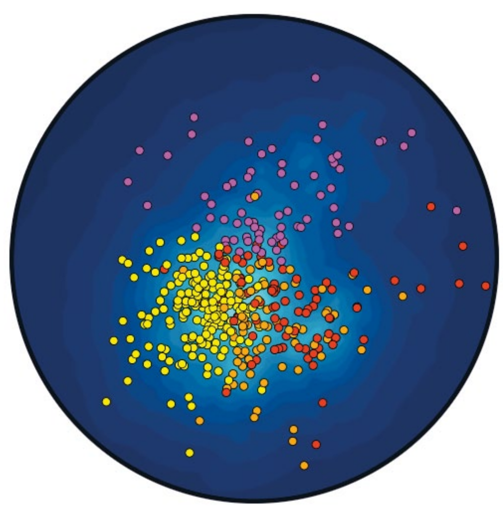

Figure 1. 2D projection of LSA map with subject ideals color coded according to segments from Step 2.

near products they like and far from those they dislike. By having the consumers rate a range of products that cover the market, you create a map which represents the market product space as shown in Figure 1. You then use this map to study the relationships between the consumers, the products, and their respective properties.

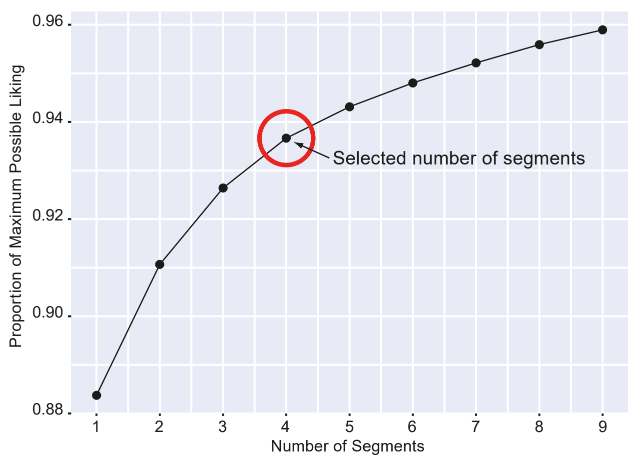

Figure 2. Team liking predicted from different segments identified by cluster analysis of ideal points from LSA.

Step 2 - Clustering of Ideal Points: Next you determine how many segments to model. Using clustering methods, you group the ideal points into segments of individuals with similar hedonic patterns. To determine how many segments to consider, a scree plot is a good visual tool for identifying points of diminishing returns, which will appear as "elbows" in the plot (See Figure 2). In this case, you see that there is an "elbow" at four segments. Machine learning tools perform best when there are roughly equal numbers of subjects in each cluster, but as long as there are not too few subjects, we can proceed. Here, we have sizes of 75, 94, 162, and 92 (18%, 22%, 38%, and 22%).

Step 3 - Segmentation via Machine Learning: You apply machine learning using demographic, behavioral, and psychographic datasets, with respondents labeled by segment, to obtain the decision tree as shown in Figure 3.

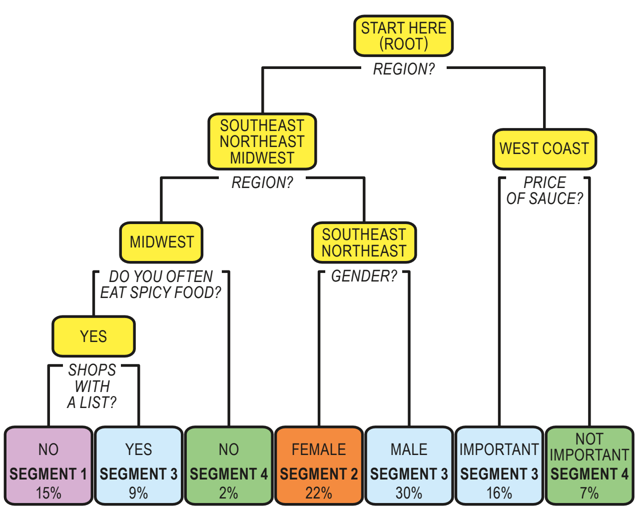

Figure 3. A machine learning solution in the form of a decision tree characterizing consumer segmentation.

You examine your decision tree to describe the segments. From the root of the tree, you find paths that lead to each of the segments and notice that two segments, Segments 3 and 4, lie at the end of several paths. Segment 1 is characterized as consumers from the Midwest who often eat spicy food and don't always shop with a list. Segment 2 is characterized as female consumers from the Northeast and Southeast (so from the East Coast). Segment 3 is characterized as consumers from the Midwest who often eat spicy food and almost always shop with a list together with male consumers from the East Coast. A third portion of Segment 3 is characterized as consumers from the West Coast who are price sensitive. Segment 4 has consumers from the Midwest who do not often eat spicy food as well as consumers from the West Coast who are not very price sensitive. From your tree you also obtain the percentage of the total that each leaf contains; you decide to ignore one of the characterizations of Segment 4 since it only contains 2% of the population.

These characterizations are not the only kinds of respondents in each segment, but they represent groups that are likely to be found in their respective segments. This information can be used for developing advertising or deciding which areas of the market to develop. Since you can determine which products are closest to each segment within the LSA map, you also know which products to select for inclusion in your portfolio and which products to market to each segment.

Conclusion: Machine learning provides new tools for consumer research. Specifically, it is now possible to find multiple characterizations for consumer segments in terms of psychographic, demographic, and behavioral data and their interactions, providing more specific descriptions of the respondents. This information can be used to guide marketing, product development, and future testing.

References and Notes

1. IFPress references are available at both www.ifpress.com and within D.M. Ennis, J.M.Ennis, and B. Rousseau (Eds.). *Tools and Applications of Sensory and Consumer Science.* Richmond, VA: The Institute for Perception.
2. Miaoulis, G., Free, V., and Parsons, H. (1990). Turf: A new planning approach for product line extensions. *Marketing Research,* March.
3. Ennis, J. M., Fayle, C. M., and Ennis, D. M. (2012). eTURF: A competitive TURF algorithm for large datasets. *Food Quality and Preference,* **23**(1), 44-48.
4. Ennis, J. M. and Russ, W. J. (2016). eTURF 2.0: From astronomical numbers of portfolios to a single optimum. *IFPress,* **19**(2) 3-4.
5. Ennis, D. M. and Ennis, J. M. (2013). How to find optimal combinations of brand components. *IFPress,* **16**(4) 3-4.
6. Ennis, J. M. and Russ, W. J. (2017). Large TURF problems: Finding custom solutions. *IFPress,* **20**(3) 3-4.
7. Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The Elements of Statistical Learning* (Vol. 1, No. 10). New York, NY, USA: Springer series in statistics.
8. See our recent IFPost, "An Overview of Machine Learning Methods", available at www.IFPress.com, for more information.
9. Kuhn, M. and Johnson, K. (2013). *Applied Predictive Modeling* (Vol. 26). New York: Springer.
10. Ennis, D. M. (2014). Unfolding. *IFPress,* **17**(3) 3-4.