Data Mining: Qualitative Analysis with Health Informatics Data

Brian Castellani and John Castellani

Qual Health Res 2003; 13; 1005
DOI: 10.1177/1049732303253523

Data Mining: Qualitative Analysis With Health Informatics Data

Brian Castellani
John Castellani

The new computational algorithms emerging in the data mining literature—in particular, the self-organizing map (SOM) and decision tree analysis (DTA)—offer qualitative researchers a unique set of tools for analyzing health informatics data. The uniqueness of these tools is that although they can be used to find meaningful patterns in large, complex quantitative databases, they are qualitative in orientation. To illustrate the utility of these tools, the authors review the two most popular: the SOM and DTA. They provide a basic definition of health informatics, focusing on how data mining assists this field, and apply the SOM and DTA to a hypothetical example to demonstrate what these tools are and how qualitative researchers can use them.

Keywords: *qualitative method; data mining; neural networking; decision tree analysis; complexity theory*

In her keynote address to the Seventh Qualitative Health Research Conference, Janice Morse (2002) argued that it is time for qualitative health researchers to broaden their methodological toolboxes. Too much emphasis, she explained, is being placed on interview data at the expense of other methods and data sources. Morse stated, "I am concerned that the preponderance of interview methods—in particular in grounded theory—is narrowing the contribution of qualitative inquiry and, therefore, our perspectives on and understanding of health and illness" (p. 116).

To remedy this limitation, Morse suggested that qualitative health researchers adopt larger programs of research to include not only multiple methods but also multiple forms of data. This broader approach, she believes, will help researchers provide a more comprehensive picture of their phenomena of study. She stated, "I am making the case for researchers to move beyond single studies using single methods toward developing multiple-method research programs that consider phenomena comprehensively, in all of their complexity" (pp. 116-117).

Inspired by Morse's argument, we will attempt in this article to expand the methodological toolbox of qualitative inquiry by introducing researchers to the field of data mining, which Han and Kamber (2001) have defined as the methodological process of "extracting, or 'mining,' knowledge from large amounts of [quantitative] data" to find meaningful patterns and rules concerning a given

QUALITATIVE HEALTH RESEARCH, Vol. 13 No. 7, September 2003 1005-1018
DOI: 10.1177/1049732303253523

research question (p. 4). The utility of data mining to qualitative health research is twofold. First, although computational, the algorithms of data mining can be used as qualitative tools. Second, as qualitative tools, these algorithms allow health researchers to analyze quantitative data, particularly the large, complex databases being created by the health informatics community (Young, 2000). This is an important contribution to qualitative inquiry given the fact that the field of health informatics is currently restructuring the organization and usage of information in health care (Englebardt, 2002; Hebda, 2001).

Because of the potential contribution of data mining to qualitative health research, we begin with a quick review of health informatics, focusing on how data mining assists in this field. Second, we review the data mining techniques most useful for qualitative analysis: the self-organizing map (SOM) and decision tree analysis (DTA). For our review, we apply the SOM and DTA to a hypothetical example to explain what these tools are and how qualitative researchers can use them.

HEALTH INFORMATICS

With little more than a decade to its name, the field of health informatics has changed the face of health care throughout many parts of the world (Cios, 2001; Tan, 2001; Young, 2000). It has done so by changing not only the way in which information is collected and stored but also the relevance this information has to the organization, delivery, and payment of health care (Hedba, 2001). Examples include hand-held personal data assistants, electronic hospital charts, and cost and utilization databases (Young, 2000).

In terms of a definition, health informatics is the application of computer science, communications technology, and database management to the organization, delivery, and analysis of any and all information relevant to health care (Cios, 2001). The complex databases created by the health informatics community are called *data warehouses*. A data warehouse is "a vast database that stores information, as a data repository does, but goes a step further, allowing users to access data to perform research-oriented analyses" (Young, 2000, p. 264). The types of data stored in these warehouses range from quantitative to analog to qualitative data. The format of these warehouses varies as well, such as relational databases (e.g., electronic patient charts) and time series databases (e.g., insurance and accounting records).

What all of this data warehousing amounts to is an information explosion within the health care field. The problem, however, is finding the right methodological tools to "mine" this new data given its enormous variety, size, and complexity. As Klose, Nurnberger, Nauck, and Kruse (2001) have stated, "Exploiting the information contained in these archives in an intelligent way turns out to be difficult" (p. 1). This "difficulty" is a challenge to both quantitative and qualitative method.

In terms of quantitative method, the challenges are three. First, quantitative method is far too linear, reductionistic, and "homogenizing" in its assumptions to engage the nonlinearity, diversity, and complexity endemic to most health informatics databases (Kosko, 1993). It is because of this limitation that Lloyd-Williams (1999) has stated, "It is often the case that large collections of data, however well structured, conceal implicit patterns of information which cannot readily be detected by conventional analysis techniques" (p. 139). Second, quantitative

method is limited by its rigid theory requirements. Because the goal of quantitative method is to test a set of hypotheses—verify a theory—it cannot freely explore health informatics data warehouses. In fact, to do so is considered a breach in method. For example, the terms quantitative researchers use to describe exploratory analysis, such as *data drugging*, *data snooping*, and *data fishing*, to give a few examples, are pejorative. Theory should guide investigation, not the other way around. In terms of health informatics, this is a major limitation because mining knowledge from these data warehouses often requires an inductive, exploratory approach (Ragin, 2000). The final limitation has to do with the messiness common to health informatics data. Health informatics data are practical: They are collected by and for health professionals. As such, variables are often poorly defined, data are missing or not easily transformed into analyzable information, and many fields of information have non-normal distributions. Consider, for example, the patient chart. Think of the variety of ways in which diagnostic and treatment codes are used, or the variability in billing procedures and utilization rates, not to mention differences in charting. For quantitative researchers, this "messiness" amounts to a major violation of the positivistic paradigm, which, quite often, leads to a breakdown in quantitative method (Lloyd-Williams, 1999).

In terms of qualitative method, again, the main issue is the data. Although the advantage of qualitative method is its freedom from the limitations of quantitative method, by definition it is not the best method for analyzing quantitative data (Glaser & Strauss, 1967). This limitation is particularly evident when analyzing large data warehouses: They are too complex and often not worth the time necessary to study them. As Ragin (2000) explained, qualitative strategies "are easy to implement when the number of cases is small—the usual situation in qualitative inquiry. However, they are rarely used when *N*s are large because of analytic difficulties" (p. 5). It is because of these limitations that the health informatics community has turned to data mining and its new algorithms. The strength of these algorithms is that although they can "crunch" tremendously large and complex quantitative databases, they have the sensitivity of traditional qualitative methods.

DATA MINING DEFINED

Depending on your field of reference, data mining—otherwise known as *knowledge discovery in data* (KDD)—refers to the latest methodological advances in ecology (Capra, 1996), business (Berry & Linoff, 2000), education (Tsantis & Castellani, 2001), and health care (Cios, 2001). It is in the field of artificial intelligence and machine learning, however, where these terms and algorithms were first created (see Cilliers, 1998; Garson, 1998; Han & Kamber, 2000).

Despite disciplinary differences and variant histories, the basic definition of data mining across these fields is the same: Data mining is a data-driven, exploratory process of knowledge discovery where the focus is on finding and extracting useful patterns of information from large, complex databases (Berry & Linoff, 2000). However, data mining is more than just a method. It is an entire way of thinking about the organization and analysis of data, with the emphasis on an active program of data management. The goal is to use the algorithms of data mining to create and develop an intelligent and active database that researchers can use to generate

important and timely information about an ongoing area of inquiry. This is why, for example, data mining is so appealing to the business community (Berry & Linoff, 2000).

In grocery sales, for example, the more knowledgeable you are about your shoppers, the better able you are to develop the right inventory. To do this, however, you need to maintain an ongoing data warehouse of information, which you "mine" continually to update yourself on changing customer preferences. That is why most grocery stores have a "preferred shopper card." These cards not only provide customers with discounts but also allow store managers to know what you and customers like you might buy in the future based on your current shopping preferences. Following the logic of data mining, these grocery store data warehouses are very large and complex (e.g., thousands of shoppers with long grocery lists), the process of knowledge discovery is exploratory and data driven (e.g., what shoppers buy or will purchase is not known ahead of time), and the goal is to find complex, qualitative patterns and rules within the data. For example, a store manager might find out that for those shoppers who buy peanut butter, there has been an increase in the purchase of organic bread. This suggests that if organic peanut butter was shelved next to the organic bread, these "peanut butter" people might buy it. This is intelligent data mining. As this database is developed, it becomes "smarter" about the preferences and choices of shoppers, and managers are therefore able to make better decisions about what shoppers might like in the future or how their preferences might change.

Another example is health informatics (e.g., Cios, 2001; Richards, Rayward-Smith, Sönksen, Carey, & Weng, 2001; Stefanelli, 2001). Suppose, for example, we were interested in the quality of treatment Hispanic Americans with type 2 diabetes receive in a local hospital. By creating an intelligent database on these patients, we could develop a typology of the different ways in which they experience their health care—this might include issues of discrimination, language barriers, cultural differences, difficulty accessing care, and provider differences. With this initial research in place, we could then begin to refine and develop our typology over time. In terms of the logic of data mining, we would do this through an ongoing program of research: data collection, exploration, knowledge discovery, and refinement. In true qualitative fashion, however, "refinement" in this case would include revising both the knowledge "discovered" and the data warehouse on which this "knowledge" depends.

The idea of data mining as an active and intelligent program of research fits with Morse's (2002) agenda. Morse stated, "We urgently need to develop large projects—research programs—to study significant phenomena from multiple dimensions and at multiple levels to obtain the sense of the whole" (p. 128). Given these similarities, we turn to the SOM and DTA to define what they are and demonstrate how they can be used to further Morse's agenda.

THE TOOLS OF DATA MINING

There are several computational algorithms found in the data mining literature. Two of the most popular and powerful are the self-organizing map and decision tree analysis (Han & Kamber, 2001). Neither is, however, a panacea. Like anything else, these algorithms are built for specific tasks and have their weaknesses (Garson,

1998). In terms of qualitative method, for example, they cannot replace the important information learned through interviews or direct observation, and they certainly cannot be used on their own. They work only as one set of tools among several. This is, however, what makes these algorithms so important: They enhance the qualitative process by allowing researchers to analyze quantitative data, yet their insights are useless unless grounded in other qualitative methods. It is to the SOM and DTA, then, that we now turn our attention.

The Kohonen Self-Organizing Map

In the data mining literature, the SOM is considered a neural networking technique, the latest in artificial intelligence (Kohonen, 2001). Although the SOM in no way matches the abilities of the human brain, it is similar in design. Like the brain, the SOM is composed of a dense web of artificial neurons (also called nodes), each of which contains (a) a mathematical cell body and (b) a set of dendrites and axons, which connect the artificial neuron to the other neurons in the net. In traditional research terms, each artificial neuron represents a variable within a data warehouse, such as gender (neuron 1), ethnicity (neuron 2), age (neuron 3), and so on, to the *n*th variable (Garson, 1998).

Even though the SOM represents the latest in artificial intelligence, it is, at its most basic, a clustering technique (Kohonen, 1999, 2001). This means that its goal is to cluster people across a list of relevant variables (Garson, 1998). It accomplishes this task by taking a complex, nonlinear database and reducing it to a smaller, simpler space on which it maps the clusters it finds. As Kohonen (1999) has stated, the SOM is a clustering algorithm that "converts complex, nonlinear statistical relationships between high-dimensional data elements into simple geometric relationships between points on a low-dimensional display" (p. 59). In other words, as shown in Figure 1, the SOM is a data compression (clustering) technique composed of two basic layers: (a) the input layer, which is the data in its original form—the high-dimensional data warehouse—and (b) the output layer, which is the two-dimensional Euclidean grid onto which the clusters are mapped. The beauty of this output grid for qualitative researchers is that it functions as a conceptual topology that can be mined for important patterns and rules (Figure 2).

Applying the SOM to a Hypothetical Example

To understand the SOM more fully, let us return to our example on the quality of treatment Hispanic Americans with type 2 diabetes receive in a local hospital. As we stated earlier, we know that the quality of treatment Hispanic Americans, particularly those in the working class, receive is influenced by several major factors: access to treatment, discrimination, language barriers, and treatment compliance (Department of Health and Human Services, 2000). What we do not know, however, is how these factors influence the care Hispanic American patients receive at our hospital of study. To begin, therefore, we need to collect our data. Fortunately, our hospital has an excellent data warehouse that includes not only basic demographics (e.g., gender, occupation) but also insurance type, diagnosis, treatment, provider, and outcome. In total, our data warehouse is composed of 200 variables and 1,200 patients.

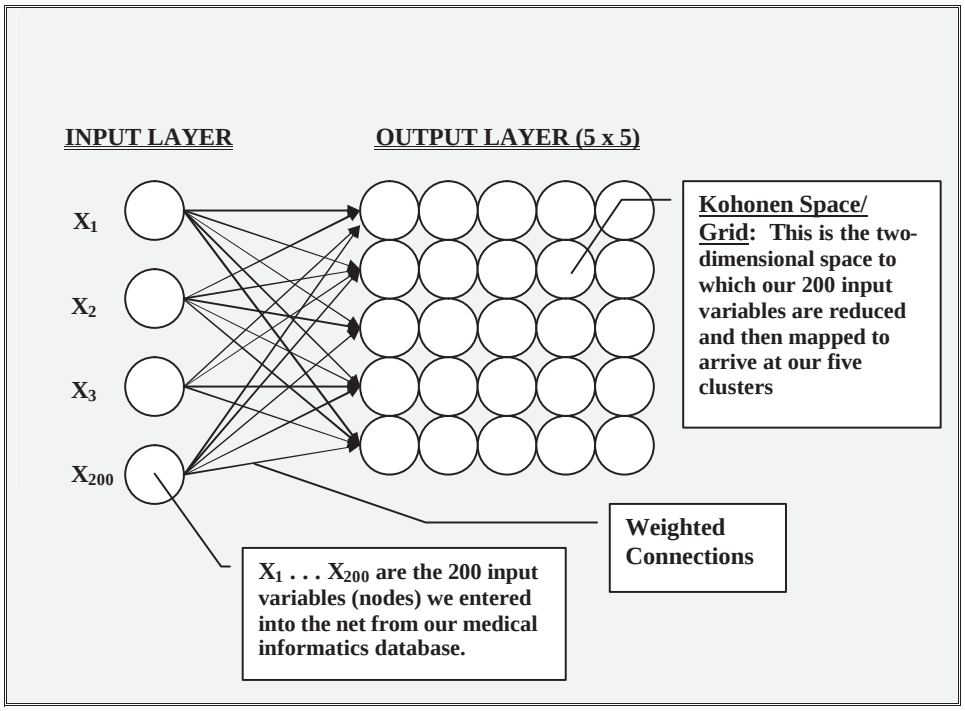


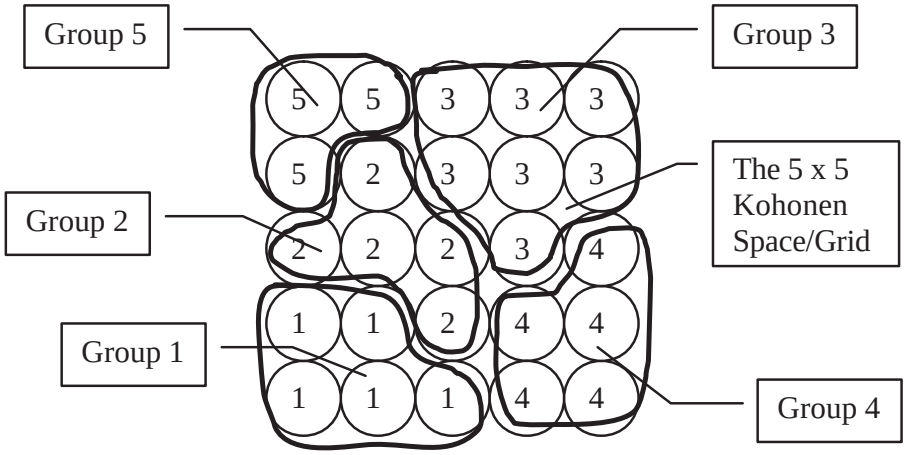
FIGURE 1: Diagram of a Kohonen Self-Organizing Map

Because we are interested in an exploratory analysis of between-group differences, our technique of choice is the SOM. We want to see if Hispanic American patients can be clustered into a set of meaningful groups as a function of the complex interactions that exist among the 200 variables in our data warehouse. In this case, the 200 variables represent the input layer of our SOM; the smaller space represents the clusters we want to find. Accomplishing our analysis requires three steps: (a) organizing the input space, (b) running the SOM, and (c) interpreting our results.

Organizing the Input Space

Once our data warehouse is identified, the first and most difficult step is to organize and connect all of the variables (artificial neurons) in the data warehouse to create our input space. Typical to most input spaces, the data are large, nonlinear, multidimensional, and complex. It is because of this complexity that the input space is referred to as a multidimensional space. Once this initial space has been organized, cleaned, and connected, it takes the form of a data warehouse.

Preparing the data is the least glamorous aspect of data mining, but it is the most important. As we pointed out earlier, a main reason why traditional quantitative techniques collapse in the presence of health informatics data is that most data warehouses are messy: they have non-normal distributions, missing data, and crudely defined variables. Being a data mining technique, however, the SOM can handle these issues—and this is particularly true when the SOM is used as a



Patients in Each Cluster
Group 1, (*n* = 400)
Group 2, (*n* = 200)
Group 3, (*n* = 400)
Group 4, (*n* = 100)
Group 5, (*n* = 100)
Total (*N* = 1,200)

FIGURE 2: The Five-Patient Clusters on the (5 x 5) Output Layer of a Self-Organizing Map

qualitative technique. As Garson (1998) has pointed out, one of the major strengths of the SOM is its ability to handle messy data. Nevertheless, as Lloyd-Williams (1999) has stated, the SOM has its limits. If the data are too messy, it is hard to trust the SOM's results. This is an important point.

In their groundbreaking work, *The Discovery of Grounded Theory* (1967), Glaser and Strauss made it clear that when qualitative researchers are using quantitative data, messiness is not a major issue. Qualitative researchers are not—contra statisticians—trying to verify hypotheses or measure the exact quantity of a relationship between two or more variables. Instead, the goal is to outline and describe general patterns and rules of difference. In such cases, as Glaser and Strauss explained, " 'crude' or 'general duty' indices . . . suffice to indicate the concepts of the theory and to establish general relationships between them" (p. 190). This is not to suggest, however, that anything goes. If the data are too messy, the SOM's results can be just as misleading as those produced by any other method. This is particularly problematic, as in the case of our current example, if, as a result of overly messy data, a recommendation for treatment is made that ends up hurting Hispanic American patients. Therefore, even when used qualitatively, the success of the SOM, as explained by Lloyd-Williams (1999), "is determined to a large extent prior to the actual data-mining activity, i.e., during the activities performed leading up to the

production of the cleaned data" (p. 157). Therefore, having gone through our database at our local hospital, we will make sure we have a reasonably well-organized data warehouse.

Analyzing the Data

As a nonlinear clustering technique, the process by which the SOM conducts its analyses is called *unsupervised learning*. Unlike traditional statistics, the SOM requires no predefined hypotheses or theoretical model to guide its analysis of the data. Instead, in true qualitative fashion, it depends on the self-organization of the data to inductively arrive at the best set of clusters representing the data (Kohonen, 2001). The fact that the SOM can engage in unsupervised learning is why it is considered a form of artificial intelligence (Kosko, 1993). Although, as stated earlier, the SOM is in no way at the level of a human brain and there is no conscious intent, the fact is that the SOM learns. It learns by returning to the input connections repeatedly, until it arrives at the best set of weighted relationships for the two-dimensional output space. This learning is entirely mathematical, but it is, nonetheless, learning. In the case of our hypothetical study, for example, the SOM tries to figure out which set of clusters best represents the differences between patients across our 200-variable input space. The SOM ends its analysis when—similar to the process of saturation—successive passes through the data no longer reveal new information. It is at this point that the SOM has completed its analyses, and it is at this point that we turn to our output space to interpret our clusters.

Interpreting Our Results

In terms of qualitative research, the output grid is the most important aspect of the SOM. On completion of its analyses, what the SOM produces is a two-dimensional graphic representation of the data as clustered into its most basic components. The utility of this grid, as shown in Figure 2, is that it can be mined for important qualitative patterns and rules. Using this output space, the qualitative researcher can look at the output clusters, draw circles around them, and compare them to one another. In fact, because the output space is Euclidean, the researcher can even measure the distance between clusters and assume that the clusters closer to one another are more alike than those far apart.

In the case of our hypothetical study, for example, as shown in Figure 2, we can draw circles around what appear to be five clusters. Then, having outlined these clusters, we can make the initial interpretation that Clusters 1 and 3, and 2 and 4 are the farthest apart from one another and that Cluster 2 shares characteristics with the other four clusters.

Applying Traditional Qualitative Methods

Although the SOM provides an excellent graphic representation of the clusters in our hypothetical study, its major limitation is that it does not tell us which variables account for our five group differences. In other words, we know that Hispanic American patients with Type 2 diabetes can be clustered into five different experiences with treatment, but we do not know why. We do, however, have a hint. Because we can classify our patients according to the five clusters to which they

belong, we can go back to our data warehouse to determine which variables account for these group differences. In other words, we can group patients according to their cluster membership and begin examining what variables seem to account for the differences between them.

To generate our five clusters of patients, we do as follows. First, we go back to our database and create a new variable called a *cluster*. Second, we assign each patient a number between 1 and 5 according to the cluster to which he or she belongs. Finally, we generate a report for each cluster, which would include, initially, (a) the patients who belong to the cluster and (b) descriptive information about their scores on the 200 variables. Once we have generated our five groups and basic statistics for each cluster, the next step is to decide how to mine our data and input space (map) for important patterns and trends. Given the qualitative methods available, we could approach this analysis in any number of ways.

We could, for example, apply grounded theory method to each cluster (Castellani, Castellani, & Spray, in press). Using the techniques of coding and note taking, we could comb through the 200 factors to determine which set of variables best explains each cluster. Slowly, using the comparative method of analysis, we would arrive at a grounded theory for the different experiences our Hispanic American patients have with treatment. To further our understanding of the five clusters, we could also apply the tools of ethnography and clinical interviewing. For example, we could interview a theoretical sample of patients from each cluster to see how they, as patients, conceptualize their different treatment experiences. We could then, using the tools of grounded theory, see if their explanations match the set of variables we found to be unique to each cluster. We could even observe a theoretical sample of patients from each cluster during their treatment encounters to see what their different experiences with treatment are. Again, this would follow Morse's argument for a multimethod, long-term research project.

Still, despite the qualitative approach chosen, our qualitative analyses will eventually run into a major problem. Because our data warehouse is large, complex, and quantitative, it will take a tremendous amount of time using qualitative methods alone to figure out which combination of variables account for our study's five clusters. Consider, for example, the possible combinations of our 200-variable matrix. For this reason, even when using other qualitative methods, we recommend the assistance of DTA.

Decision Tree Analysis

Whereas the SOM is a clustering technique, DTA is a method of classification. Its basic goal, at least when used with the SOM, is to figure out which set of variables best explains why people belong to one cluster or another (Han & Kamber, 2001). Breiman, Friedman, Olshen, and Stone (1984) stated this goal as a classification rule: "Given a set of measurements on a case or object, find a systematic way of predicting what class it is in " (p. 3). The "systematic" process of classification is known as *recursive partitioning* (Berry & Linoff, 2000). The graphic output of this classification rule is the decision tree (Figure 3).

Recursive partitioning works as follows: DTA combs through a database to find which variables can be partitioned in such a way that people can be classified according to them. As Berry and Linoff (2000) have stated, "Each branch of a

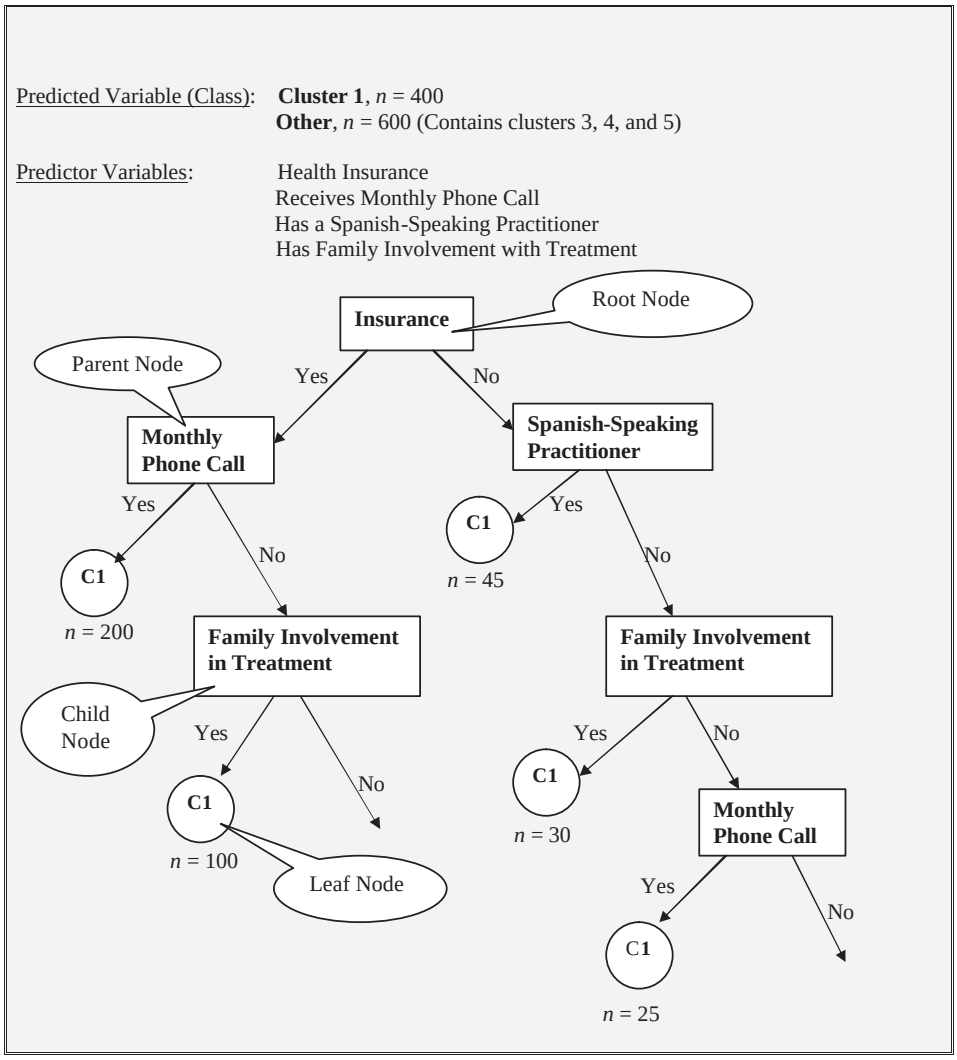


FIGURE 3: Diagram of a Decision Tree for Cluster 1 Versus Clusters 3, 4, and 5

decision tree is a test on a single variable that cuts the space into two or more pieces" (p. 112). This basic split is then used to correctly classify people. For example, in a study on depression, it might be the case that age is an important variable for determining when people are most at risk, but only when age is split into three groups: (a) below age 25, (b) between 25 and 55, and (c) over 55. In terms of nonlinearity—and this is one of the reasons why DTA is so helpful—it may be that, in this particular study, people most at risk for depression are under the age of 25 or over the age of 55 but not between 25 and 55, providing a nonlinear, U-shaped distribution. More important, this three-section split allows DTA to correctly classify this study's two clusters of patients: Cluster 1 (depressed patients) and Cluster 2 (nondepressed patients). The patients from Cluster 1 would be primarily in the low

and high age groups, and patients from Cluster 2 would be primarily in the middle group.

Applying DTA to Our Hypothetical Example

To further clarify how the DTA works, let us return to our hypothetical study. As you recall, our goal in using the DTA is to determine why the SOM clustered people the way it did. To begin, we already know what our clustering variable is (the quality of treatment received), and we know the five clusters that we are trying to classify. All we need to do, then, is run DTA.

Following the logic of recursive partitioning, DTA begins by selecting the first variable in the tree that will explain our five clusters. This first variable is called the root node. The root node is chosen first because, in comparison to the other 199 variables in our study, it does the best job of distinguishing our five clusters from one another. For example, DTA might choose glucose level. Choosing glucose level as the root node means that DTA was able to split this variable into five different levels, one for each cluster. For example, it might be the case that Clusters 1 and 2 had good glucose levels, whereas Clusters 3, 4, and 5 had middle to poor glucose levels.

Once the root node is determined, DTA finds the next "best" variable for each of the major splits. For example, for Cluster 1, which has the best glucose level, the next most important variable might be having insurance. As Figure 3 shows, it is this hierarchical fashion of layering variables that gives DTA its name. Actually, the more accurate description would be "root analysis," because, starting with the root node, the branches of the tree for each subsequent node (variable) go downward, like roots. Finally, each branch or set of branches ends with what is called a leaf node, which is the point at which any path along the tree ends (Berry & Linoff, 2000).

To make the structure of a decision tree clearer, let us consider running DTA on just Cluster 1 in comparison to the rest of the clusters. We already know from the first DTA that patients in Cluster 1 (C1) have good glucose levels. Our question, then, is Why? What is it about the treatment experience of these patients that led them to have such good clinical outcomes?

For our DTA, our sample would be broken into two classes: those patients belonging to C1 ($n = 400$) and those in the other clusters ($n = 600$). As a side note, we will drop Cluster 2 (C2) from this DTA because C2 patients also have good glucose levels and might therefore represent another group who had a good experience with treatment.

As we have shown in Figure 3, running our DTA on C1 versus the other three clusters, the root variable chosen is *health insurance*: Patients with insurance go to the left; the rest go to the right. In terms of those with insurance, the next variable is *reception of a monthly phone call* ($n = 200$), followed by *family involvement in treatment* ($n = 100$). Again, those with a monthly phone call go the left, and the rest go to the right; the same is true for family involvement in treatment.

Interpreting Our Results

In terms of interpreting our results, what this first major branch of our tree suggests is that a prime reason for membership in C1 is having insurance (75%). We also

know that the next major variables are receiving a phone call or having family involved in treatment. These three variables, therefore, seem to explain reasonably well why the SOM clustered the patients in C1 together. This is not, however, where our analyses end. Although the majority of patients in C1 have insurance, roughly 25% do not. How, then, do we explain why these patients are in C1?

As we look at our decision tree in Figure 3, it appears that although having a Spanish-speaking practitioner accounts for 11% of C1 patients, overall, the two key factors for our uninsured C1 patients are family involvement in treatment and receipt of a monthly phone call. In fact, if insurance is taken out of the DTA—even though it accounts for 75% of all the cases—it seems as if the main predictor variables around which C1 patients cluster is receiving a monthly phone call or having family involved in treatment. What "having insurance" might suggest, therefore, is that for the majority of C1 patients, the reason they received monthly phone calls and had family involvement in treatment is that they can afford such services. This tentative theory is reinforced by the fact that C1 is located on the opposite side of Clusters 3, 4, and 5—whose members did not have insurance—on our SOM output map. In other words, the style of treatment received is ultimately important, but patients have to be able to afford such services. However, this is only a guess. We still have to explain, for example, why C1 and C2 are so close to one another on our SOM map.

We dropped C2 from our second DTA because the patients in this group, like those in C1, had good glucose levels. It would be very informative, then, if we found out that although these patients do not have insurance—which makes them different from the majority of patients in C1—the variables accounting for C2's good glucose levels were the same as those for C1. It would also be interesting if, after running several more DTAs, the patients on the edge of C1, closest to C2, were those who did not have insurance. Conversely, our tentative theory would be further strengthened if it turned out that members of Clusters 3, 4, and 5 did not receive the same types of services as did the patients in C1 and C2 and had poor clinical outcomes. These findings would begin to suggest that C2 is in the middle of our SOM output matrix because, like Clusters 3, 4, and 5, these patients do not have insurance, but, like the patients in C1, the C2 patients received effective services.

However, again, all of this is tentative. Given the complexity of the variables involved, we would still need to explore these clusters further. We may decide, for example, to run DTA on only C1 and C2 to see what distinguishes them from one another. These differences may turn out to be a function of more than just insurance; perhaps a variable exists that we have yet to consider. We could also run a DTA comparing C2 to Clusters 3, 4, and 5 to see what makes this cluster similar and yet different from these other clusters. We could also further examine why the SOM distinguished Clusters 3, 4, and 5 from one another. This is equally important because cluster 3 is the largest cluster on our SOM output map. Therefore, although we have arrived at some initially interesting results, we have yet to really mine the data for its underlying patterns of complexity. However, we have made a start. This, then, provides a brief but useful demonstration of the value of the DTA and the SOM for the qualitative analysis of quantitative health informatics data warehouses.

CONCLUSION

As Morse (2002) has made clear in her keynote speech, to address the increasing complexities of health and illness in today's society, qualitative health researchers need to broaden their methodological toolboxes. They need to embrace multiple methods with multiple data types, and they need to engage in larger programs of research. This is particularly true for qualitative health researchers who wish to make use of the new data warehouses being created by the health informatics community.

To "mine" these data warehouses, however, qualitative researchers need help from the computational algorithms emerging within the data mining literature, for two reasons. First, the new data mining computational algorithms—particularly the SOM and DTA—can be used as qualitative tools. Second, as qualitative tools, these algorithms allow researchers to engage in data-driven, exploratory analysis of quantitative health informatics data. It is for these reasons that we recommend the SOM and DTA as useful tools for qualitative researchers.

REFERENCES

Berry, M., & Linoff, S. (2000). *Mastering data mining: The art and science of customer relationship management*. New York: Wiley.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Boca Raton, FL: Chapman & Hall/CRC.

Capra, F. (1996). *The web of life*. New York: Anchor Doubleday.

Castellani, B., Castellani, J., & Spray, S. L. (in press). Grounded neural networking: Modeling complex quantitative data. *Symbolic Interaction*.

Cilliers, P. (1998). *Complexity and postmodernism: Understanding complex systems*. New York: Routledge.

Cios, K. J. (2001). *Medical data mining and knowledge discovery*. Denver, CO: Springer-Verlag.

Department of Health and Human Services. (2000). *Healthy people 2010: With understanding and improving health and objectives for improving health* (2nd ed., 2 vols.). Washington, DC: Government Printing Office.

Englebardt, S. (2002). *Health care informatics: An interdisciplinary approach*. St. Louis, MO: C. V. Mosby.

Garson, D. (1998). *Neural networks: An introductory guide for social scientists*. Thousand Oaks, CA: Sage.

Glaser, B., & Strauss, A. (1967). *The discovery of grounded theory*. New York: Aldine De Gruyter.

Han, J., & Kamber, M. (2001). *Data mining: Concepts and techniques*. San Francisco: Morgan Kaufmann.

Hebda, T, (2001). *Handbook of informatics for nurses and health care professionals*. Upper Saddle River, NJ: Prentice Hall.

Klose, A., Nurnberger, A., Nauck D., & Kruse, R. (2001). Data mining with neuro-fuzzy models. In A. Kandel, M. Last, & H. Bunke (Eds.), *Data mining and computational intelligence* (pp. 1-36). Denver, CO: Springer-Verlag.

Kohonen, T. (1999). Spotting relevant information in extremely large document collections. In B. Reusch (Ed.), *Computational intelligence: Theory and applications* (pp. 59-61). New York: Springer.

Kohonen, T. (2001). *Self-organizing maps* (3rd ed.). New York: Springer.

Kosko, B. (1993). *Fuzzy thinking: The new science of fuzzy logic*. New York: Hyperion.

Lloyd-Williams, M. (1999). Empirical studies of the knowledge discovery approach to health-information analysis. In M. A. Bramer (Ed.), *Knowledge discovery and data mining*. London: Institution of Electrical Engineers.

Morse, J. (2002). Qualitative health research: Challenges for the 21st century. *Qualitative Health Research*, *12*, 116-129.

Ragin, C. (2000). *Fuzzy-set social science*. Chicago: University of Chicago Press.

Richards, G., Rayward-Smith, V., Sönksen, P., Carey, S., & Weng, C. (2001). Data mining for indicators of early mortality in a database of clinical records. *Artificial Intelligence in Medicine*, *22*, 215-231.

Stefanelli, M. (2001). The socio-organizational age of artificial intelligence in medicine. *Artificial Intelligence in Medicine*, *23*, 25-47.

Tan, J. (2001). *Health management information systems: Method and practical applications*. Gaithersburg, MD: Aspen.

Tsantis, L., & Castellani, J. (2001). Enhancing learning environments through solution-based knowledge discovery tools: Forecasting for self-perpetuating systemic reform. *Journal of Special Education Technology*, *16*(4), 39-52.

Young, K. (2000). *Informatics for healthcare professionals*. Philadelphia: F. A. Davis.

Brian Castellani, Ph.D., is an assistant professor of sociology at Kent State University, Ashtabula, Ohio.

John Castellani, Ph.D., is an assistant professor at the Center for Technology in Education, Johns Hopkins University, Columbia, Maryland.