TEACHING DATA MINING AND OTHER DATABASE RESEARCH METHODS IN MARKETING RESEARCH CLASSES: AN INTERIM ASSESSMENT

Joseph L. Orsini
College of Business Administration California State University, Sacramento
Sacramento, CA 95819-6088 916-278-6992

ABSTRACT

Continuous growth in computer storage capacity has led to the development of methods to investigate large databases directly, rather than taking samples of the database population. However, while industry is heavily involved in using databases for marketing purposes, the academic component of the marketing research discipline has not yet caught up with industry practice. Literature review and the result of an industry survey form the basis for assessing the impact of this phenomenon on teaching marketing research.

BACKGROUND

My first introduction to Data Mining [DM] was several years ago as a reviewer for a University research grant program when one of my colleagues in Management Information Systems proposed a grant to use DM. In my mind, DM had the pejorative meaning as noted by Berry and Linoff (1997): "selectively trying to find data that will support a particular hypothesis". Out of ignorance, since the colleague did not define what DM was, I was not very favorable to his application.

Time passed, and I began to see the Data Mining term mentioned in marketing contexts, but never fully defined. After two decades of teaching research methods, my feeling was that any method that was "just another exotic methodology" was of little relevance to my undergraduate and MBA research classes. In these classes, it is rare that a student can even explain how to do a single sample t-test at the start of the semester, so there is more than enough material to cover during the semester without getting into the more exotic methods.

However, changes in computer technology have allowed the accumulation of large quantities of data in digital format, and business (and other) organizations have taken advantage of it. At the midpoint of the 20th century no company had more than the equivalent of 30 or 40 megabytes of data [in paper format] in its ledgers, books and file cabinets. By the end of the 20th century, however, things have dramatically changed; UPS, for example, has 17 million megabytes [17 terabytes] in the database of package level detail it uses to track shipments. This is about the same amount of storage as contained in the books of the Library of Congress, the largest repository of information in the

world (Berry and Linoff 2000). Changes in methods of data analysis induced by storage of this order of magnitude led me to take another look at this "Data Mining" methodology.

LITERATURE REVIEW

Journals

Thinking that perhaps I had been remiss in my currency, I turned to the leading marketing, marketing research, and marketing education journals for help in explaining to me what I had missed. While there were faculty admonitions to "stay in the front of technological change and take a true leadership role in helping students develop business skills necessary for success" (Smart, et al. 1999, p. 206), and to "enhance the relevance of research inquiries" (Day and Montgomery 1999, p. 12), very little was found to accomplish this "true leadership role" relative to DM.

First I reviewed the 1997 through 1999 Journal of Marketing Research and Journal of Business Research, the journals primarily related to the courses I teach. Only one article on DM was found, by Ying, Platt and Platt (1999), which was not on a marketing topic (it investigated reasons for business bankruptcy. The article compared the discriminating power of neural networks (one of several DM methods) to discriminant analysis, finding slightly better results with the latter. However, their entire dataset consisted of only 122 cases, so the results are not surprising, as DM methods are designed to be utilized on very large databases.

The small dataset used in this study is exacerbated by the requirements of neural networks, typical of several DM methods, which require more than one set of data for development. For neural networks, three datasets are needed: a training set, a validation set, and a test set (Berry and Linoff 1997, Yang, et al. 1999). These 122 records thus had to be divided into three, thereby seeming not to provide a suitable venue for a comparison test. It would be interesting to replicate this investigation with a more suitable set of data.

Further investigation of journals carrying relevant articles was carried out. A search of the *Journal of Marketing's* "Marketing Literature Review" for the three volumes in 2000 yielded only two articles on DM, and those in *Stores*, an industry journal. They were not pursued, as the point had been made that there was insufficient literature to develop a solid grounding in database analysis and DM.

Textbooks

An examination of several current editions of marketing textbooks for application examples indicates that many of the books, but not all, have begun to include mention of DM. For example, Berkowitz, et al. (2000) do not mention Data Mining, while Solomon and Stuart (2000) do. Those marketing textbooks which do discuss DM (e.g. Czinkota, et al. 2000) typically discuss it in conjunction with Direct Marketing and Database Marketing, at most devoting a paragraph or two to DM specifically. The heavy involvement of Direct Marketing in utilization of DM is understandable, as Hughes (1996) notes that this form of retailing began in the 19th century with catalogs, and has continually adopted new approaches as technology has changed.

The current editions of marketing research textbooks, and business research textbooks, have typically begun to contain discussions of databases and DM in their sections on decision support systems [DSS], or secondary data. For example, Churchill (2001) vs. Churchill (1996); Aaker, Kumar and Day (2001) vs. their 1998 edition; Cooper and Schindler (1998) vs. Cooper and Schindler (1995); and Zikmund (2000) vs Zikmund (1997) all have DM discussions in their new editions, but not in the previous ones. However, there is typically not enough discussion included to give an instructor confidence in answering student questions on the topic.

INDUSTRY SURVEY

Still a bit uncertain about how much database research utilizing DM methods was actually taking place in the marketing research industry, it was decided to ask some research practitioners. A survey of the "Honomichl 50" marketing research organizations was undertaken to find the level of interest and utilization of DM applications among marketing research consultants. The Honomichl listing in the *Marketing News* (1999) was used as the sample frame, with the assumption that the larger of the research organizations would be the ones most likely to adopt the new methods. Given the small sample (including three questionnaires returned as "undeliverable"), the response rate of 23% makes projections of the results to the entire marketing research industry somewhat tenuous.

The most important finding of the survey was that some marketing research organizations (45% of the respondents) were using DM in their consulting activities, and had been doing so for several years. Others were in the process of developing DM for future use (27%), while still others were not using DM, and had no immediate plans for doing so (27%). All of the organizations that were using DM were utilizing in-house expertise; none of the respondents were contracting out those services.

The survey indicated that the primary applications of DM methods were in the areas of customer relationship management and market segmentation, followed by product configuration development. Those organizations not using or developing DM cited lack of customer demand for the methods as the main reason for not doing so, rather than either expense or a belief that DM is a fad.

The Data Mining family of methods includes a number of mathematical techniques that investigate the entire set of data in a large dataset, rather than using samples as traditional statistical methods do. The DM tools most widely used by the survey respondents were cluster detection and decision trees, followed by market basket analysis, neural networks, and rule induction.

Respondents to the survey were not significantly different from the entire Honomichl 50 population with respect to organization size and proportion having revenues from outside the United States. Somewhat surprising was the finding that those organizations using or developing DM methods were independent of organization size. That is, the smaller of the research organizations were proportionally just as likely to be using DM methods as the larger organizations in the Honomichl 50 list.

WHERE TO FIND MORE INFORMATION

Professional Seminars

The paucity of literature on marketing applications of DM methods may be changing. Within the last year, DM has begun to appear in seminars primarily aimed at professionals. For example, AMA's (2000) Advanced Research Techniques (ART) Forum includes sessions on DM, which the 1999 ART did not; the 2000 Applied Research Methods Conference and School of Marketing Research also have sessions or tutorials on DM. Non-AMA research seminars also have begun to include DM in their list of offerings; e.g., Neilsen-Burke (*Marketing News* 2000). Unfortunately, the cost of attending these seminars, and the relative brevity of the presentation and the complexity of the topic, do not hold forth much promise of developing a level of expertise comparable to what most professors have in traditional statistical research methodology.

Professional Books

This category of DM information currently appears to be the most useful. Several professional books which directly address DM, its database context, and marketing applications, are available. The original Berry and Linoff (1997), their updated version (2000), and Groh (1998) are all very helpful (the original Berry and Linoff contains some useful material not in the later version). There are also professional books more oriented toward applications and less on methodology (e.g., Hughes 1996) that may provide some useful material.

SOME TECHNICAL DETAIL

With a new and evolving discipline, definitions are not firmly established, and DM is very much in accordance with this model. For this discussion, DM will be defined as: "the process of exploration and analysis of large quantities of data in order to discover meaningful patterns and rules" (Berry and Linoff 2000, p. 12).

Database Data

Hair, et al. (2000) note that the second half of the 20th Century has seen two shifts in the "fundamental character of data analysis" (p. 671). The first was the mid-1970s, when mainframe and (later) personal computers allowed multivariable methods to gain widespread acceptance and use. The second was in the 1990s, with the development of large-scale databases, and the development of new methods to analyze the "information avalanche".

The preparation of data warehouses (large sets of data accumulated from different sources) for DM analysis is no simple task. In fact, it can be the most time consuming part of any DM analysis, as noted by Groh (1998):

> "The biggest challenge business analysts face in using data mining is how to extract, integrate, cleanse, and prepare data to solve their most pressing business problems. This issue is a formidable one, and can take the bulk of the time in the data mining process."

Insofar as a data warehouse may not even exist in any given research application, one may have to be developed using Information System methods to merge sets of data from a variety of sources, or query tools to create files from existing databases (Groh 1998).

Data quality is also an important issue, so much so that it is advised to "beware the consultant or data-mining product that does not address (or downplays) how to get the data ready for analysis." (Pettit 2000).). In addition to redundant data, and incorrect or inconsistent data, there is an issue of data format. Data from operations tends to be either continuous or categorical (ratio or nominal, in typical marketing research terminology), and some of the DM methodologies do not handle continuous data; they must be made categorical (ordinal or nominal) to be used. All this is assuming that the existing database is sufficient to use without adding any additional data from other sources (e.g. consumer credit card information).

Database Analytical Tools

The following is a brief listing of analytical tools which may be applied after a suitable database has been prepared.

Standard Statistics: a sample of a database may be obtained and inserted into spreadsheet format; therefore, all of the "traditional" statistical techniques are potentially available for use (Hair, et al., 1998). Since there is access to the entire "population", very large samples, to the limits of the capacity of the statistical package employed, may be obtained. As with all samples, however, the existence of very small "segments" may still be too small to yield useful information, even though they may be profitable in the marketplace. Further, the concept of statistical significance becomes less useful with large samples, as findings are increasingly likely to be "significant", even if they are of little managerial importance.

Queries: queries are investigative questions relating to specific items in the database, and may be in the form of Structured Query Language (SQL) for use in On Line Analytical Processing (OLAP) methods (Berry and Linoff 2000). A simple query may look at the entire database for questions such as "how many times has Brand X been purchased concurrently with Brand Y?" A more complex query may first require the formation of a new and smaller dataset consisting only of the variables of interest, thus allowing faster examination of relationships, and the use of a variety of DM tools to examine the relationships.

Data Visualization: methods of displaying data for visual review have been found to be useful, and have been increasingly developed for assisting managerial decisions (Hair, et al., 1998). Visualization methods range from simple graphs and charts to quite complex multidimensional structures, which are combinations of art and mathematics. Shape, color, line and artistic graphics are all used to convey what is hopefully informative material. Query software may use visualization methods to answer the question asked.

Data Mining Tools: Data Mining, as contrasted with queries and visualization, includes the following methods: Market Basket Analysis, Memory Based Reasoning, Genetic Algorithms, Cluster Detection, Link Analysis, Decision Trees, and Neural Networks (Berry and Linoff 1997). The tasks performed by these methods may be classified as Directed methodologies (which are generally referred to in statistical analysis as dependence methods), and Undirected methodologies (interdependence methods). The former include classification, estimation and prediction tasks, and the latter include affinity grouping or association rules, clustering, and description (Berry and Linoff 2000). Each of the DM methodologies indicated above may be useful in performing both undirected and directed tasks.

CONCLUSIONS

The last two decades have seen an explosion in the capacity of business organizations to retrieve and store operational data. When combined with customer data, these massive databases constitute a valuable source of information for marketing purposes. Businesses are increasingly using this information in their marketing efforts, particularly in promotion.

Development of methods of researching these databases has paralleled the expansion of storage capacity and access. While it is possible to obtain samples of the databases and use traditional statistical methods, the newly developed database investigation methods have some unique advantages. In particular, they have the capacity to discover very small segments that may not exist in sufficient sizes in a sample to be uncovered, but which may still be profitable to management to pursue.

Developing the expertise necessary to teach basic DM concepts in a marketing research classroom context is increasingly necessary. However, there are some important barriers to developing this expertise, including:

- Lack of marketing literature information
- Complexity of dataset preparation
- Complexity of the investigation methods
- Costs of software

Some of these barriers may be addressed with minimal difficulty; others are less tractable. This paper is titled "An Interim Assessment" because, at its writing, the author had not yet surmounted these barriers.

The usual academic information sources of journal articles and academic conferences do not seem to be very helpful at the present time. The acquisition of appropriate textbooks is also currently a problem. The few DM marketing discipline-oriented books available are written more for a managerial or professional audience than an academic one.

Data issues are far more complex in database investigation than they are in data used for statistical analysis. While databases appropriate for analysis are becoming more common in business data repositories, there still may be a substantial amount of data preparation necessary prior to any investigation using either traditional statistics or the newer data mining methods.

The net result is that, even when the investigator has a good grasp of the function of the various tools, their application for research purposes will require some additional expertise. The data complexity will require either that the researcher be well versed in information technology, or work closely with an IT person. It is anticipated that this situation will continue to exist for some time.

Fortunately, this aspect of the task for the marketing research instructor is potentially not quite so difficult. "Groomed" datasets may be developed for classroom purposes, so that the more technical problems involved could have already been rectified, and only research decision issues be unresolved (e.g. where to define category limits in order to make a continuous variable into a discrete variable).

Developing some level of expertise in the tools still remains a barrier to classroom inclusion, due to both the information shortages and the substantial complexity of the investigative tools used. Insofar as the roots of the tools lie in different disciplines, it is more difficult to find an "expert" in all the tools than in statistical methods. However, these tools are increasingly becoming more "user friendly". Software for classroom purposes is also a problem, as the commercial versions of available software are quite expensive. "Student editions" of data mining software seem to be only in the initial stages of development.

The American Marketing Association has, since its inception, been an organization that has recognized that the cooperation of both professionals and academics is necessary for the development of a strong discipline. This necessity appears to be especially true in the area of database research methodologies. It is to industry's benefit to have marketing graduates who are aware of, and familiar with, database research methods. To achieve this, those academics who teach the students need to be brought "up to speed" in this area. It is hoped that the AMA leadership will work toward methods to achieve this goal, e.g., substantial faculty discounts for professional training programs in database research methods, and "student editions" of software and prepared databases.

REFERENCES

Aaker, David A., V. Kumar and George S. Day (2001), *Marketing Research*, New York: John Wiley & Sons, Inc.

Aaker, David A., V. Kumar and George S. Day (1998), *Marketing Research*, New York: John Wiley & Sons, Inc.

American Marketing Association (2000), ttp://www.ama.org/conf, October 1.

Berkowitz, Eric N. Roger A. Kerin, Steven W. Hartley and William Rudelius (2000), *Marketing*, Boston: Irwin McGraw-Hill.

Berry, Michael J. A., and Gordon Linoff (1997), *Data Mining Techniques*, New York: John Wiley & Sons.

Berry, Michael J. A., and Gordon S. Linoff (2000), *Mastering Data Mining*, New York: John Wiley & Sons, Inc.

Burns, Alvin C. and Ronald F. Bush (2000), *Marketing Research*, Upper Saddle River, N.J.: Prentice-Hall.

Churchill, Gilbert A. Jr. (2001), *Basic Marketing Research*, Fort Worth: The Dryden Press.

Churchill, Gilbert A. Jr. (1996), *Basic Marketing Research*, Fort Worth: The Dryden Press.

Cooper, Donald R. and Pamela S. Schindler (1998), *Business Research Methods*, New York: Irwin/McGraw-Hill.

Cooper, Donald R. and Pamela S. Schindler (1995), *Business Research Methods*, New York: Irwin/McGraw-Hill.

Czinkota, Michael R., et al., (2000), *Marketing: Best Practices*, Fort Worth: The Dryden Press.

Day, George S. and David B. Montgomery (1999), "Charting New Directions for Marketing", *Journal of Marketing* 62, Special Issue, p. 3-13.

Groh, Robert (1998), *Data Mining*, Upper Saddle River, N.J.: Prentice-Hall.

Hair, Joseph F. Jr., Ralph E. Anderson, Ronald L. Tatham, and William C. Black (1998), *Multivariate Data Analysis*, Upper Saddle River, N.J.: Prentice-Hall.

Hair, Joseph F. Jr., Robert P. Bush and David J. Ortinau (2000), *Marketing Research*, Boston: Irwin McGraw-Hill.

Hughes, Arthur M. (1996), *The Complete Database Marketer*, New York: McGraw-Hill.

Hess, Michael, and Robert Mayer (2000), "Integrate behavioral and survey research", *Marketing News*, January 3, p. 22.

Journal of Marketing (2000) "Marketing Literature Review", January, p. 102.

Marketing News (1999), June 7, H1 - H40.

McDaniel, Carl Jr., and Ronald Gates (1999), *Contemporary Marketing Research*, Cincinnati: South-Western College Publishing.

Pettit, Raymond C. (2000), "Data mining: race for mission-critical info", *Marketing News* January 3, 18.

Smart, Denise T., Craig A. Kelley and Jeffery S. Conant (1999), "Marketing Education in the Year 2000: Changes Observed and Challenges Anticipated", *Journal of Marketing Education*, 21,3, December, 206-216.

Solomon, Michael R. and Elnora W. Stuart (2000), *Marketing*, Upper Saddle River, N.J.: Prentice-Hall.

Yang, Z. R., Michele Platt and Harold Platt (1999), "Probabilistic Neural Networks in Bankruptcy Prediction", *Journal of Business Research*, 67-74.

Zikmund, William G. (2000), *Business Research Methods*, Fort Worth: The Dryden Press.

Zikmund, William G. (1997), *Business Research Methods*, Fort Worth: The Dryden Press.