Breaking Out of the Black-Box: Research Challenges in Data Mining

Padhraic Smyth
Information and Computer Science
University of California, Irvine
CA 92697–3425
smyth@ics.uci.edu

1 Introduction

Database researchers, statisticians, and "data owners" often have quite different views of data. In a database context the traditional goal has been to provide a general and flexible data management framework, with less concern about the content of the data. Statisticians on the other hand have traditionally focused primarily on issues of data modeling and inference with relatively little concern for where the data physically reside or how the data will be accessed. Data owners, in turn, tend to be more focused on using the data as a means to an end: business data owners want to increase revenue by developing better predictive models, and scientific data owners typically want to develop insight into the phenomena generating the data. In this brief position paper we take a broad-scale view of "what people do with their data" (the data owner's perspective) and use this viewpoint to identify current opportunities and challenges for data mining research.

2 The Process of Data Mining

The data mining process is often characterized as a multi-stage iterative process involving data selection, data cleaning, application of data mining algorithms, evaluation, and so forth. Here we adopt a somewhat different process-oriented view and break it down into five basic steps:

1. **Exploring and Preprocessing:** the initial steps of exploring, visualizing, and querying the data, to gain insight into the data in an interactive manner. Preprocessing steps such variable selection, data focusing, and data validation can also be included in these initial steps.

2. **Modeling:** the steps involved in (a) selecting the model representations that we seek to fit to the data (e.g., a tree, a linear function, a probability density model, etc.), (b) selecting the score functions that score different models with respect to the data, and (c) specifying the computational methods and algorithms to optimize the score function (e.g., greedy local search). These "components" combined together specify the data mining algorithm to be used. The components may be "precompiled" into a specific algorithm (e.g., CART or C4.5 decision tree implementations) or may be integrated in a "customized" manner for a specific application (much more common in the sciences).

3. **Mining:** the step (often repeated) of actually running a particular data mining algorithm on a particular data set.

4. **Evaluating:** the step (often ignored) of critically evaluating the quality of the output of the data mining algorithm from step 3, both the predictions of the model and the interpretation of the fitted model itself.

5. **Deploying:** the step (rarely achieved) of putting a model from a data mining algorithm into routine predictive use, e.g., using the model continuously in real-time for scoring customers visiting an ecommerce Web site. A challenging (and under-appreciated) technical issue in this context is how and when models should be updated for such “continuous data stream” applications.

3 Two Extremes of Data Mining

Given the vast numbers of different users of data analysis tools, across different application disciplines, it is clearly dangerous to make broad generalizations about data analysis and data mining! Nonetheless, we will now consider two “prototype” users of data mining tools in the general context of the 5-step data mining provess above. These two prototypes are in some respects at opposite ends of a hypothetical spectrum of approaches to data mining.

3.1 The Business Data Miner

The first prototype data miner is “The Business-Person,” or “BP” for short. BP typically deals with large numbers of customers, for example in retail consumer or financial services environments. BP’s main goal is to use data mining techniques for prediction of customer behavior to gain competitive advantage, e.g., using decision trees to try to identify promising customers for a marketing campaign. BP’s approach to data mining is often strictly constrained by a number of factors: cost factors (the cost of data mining development and deployment should not be less than anticipated gains in revenue), integration factors (the deployed predictive model may need be integrated into an existing transaction-oriented environment), and time constraints (any potential competitive advantage may depend critically on being able to develop and deploy a predictive model quickly).

In terms of the 5-step process from the last section, the BP may not have much time for exploration and modeling, but instead will often use “off-the-shelf” tools such as decision trees to quickly generate results, i.e., they may jump quickly to Step 3 of data mining. Step 4, evaluation, is becoming increasingly important in practical applications, where evaluation may go well beyond the use of simple validation data sets. For example, ecommerce retailers (such as Amazon.com) and financial services companies (such as CapitalOne) conduct designed experiments on random subsets of consumers to get more precise and realistic evaluation of new predictive methods.

If one asks the BP what their most pressing problems are, they will almost certainly *not* say that they need a slightly more accurate decision tree algorithm, or a slightly faster association rule algorithm. Instead their most pressing problem is that of managing the overall process: how should features be defined for time-dependent data? which models are most appropriate? how much data is needed for training? will a random sample of 100k customers from a database of 10 million be “good enough”? what kind of a system can be put in place to update the models as new data arrives? and so on.

3.2 The Science Data Miner

The other prototype we consider is the “Science Person” (SP for short) which is in some sense at the other end of a hypothetical spectrum of data miners. The SP might be an atmospheric scientist, investigating global climate change patterns via analysis of spatio-temporal gridded observations of temperature, pressure, wind-speed, taken on a global grid on a daily basis for the past 30 years.

Or the SP might be working in computational biology, exploring a large gene expression data set and its relation to cancer.

The SP's data mining approach is quite different to that of the BP. SP typically spends significant time in Steps 1 and 2: exploring, visualizing, defining alternative models, and so forth. For example, in atmospheric science, the data can be "shaped" in different ways via numerous preprocessing techniques such as principal components analysis, time-series smoothing, and so forth. A single model may take 6 months or a year to develop and only be evaluated once on a particular data set. SP's primary goal is to explore model space in such a manner as to better understand the phenomena generating the data: generating better predictions is merely a stepping stone on the path to identifying better models. Thus, being able to "get inside the black box" is absolutely critical to the SP. A decision tree may provide useful predictions, but ultimately the SP will want to understand precisely what is making the tree work. In this context, it is not surprising that traditional statistical models, "generative models" for the data, have tended to be more widely accepted by SPs than "black box" algorithms such as neural networks, trees, and so forth. For example, hidden Markov models (HMMs) have been very successful in protein sequence alignment because they have taken a generic model structure (the HMM) and integrated domain-specific knowledge into the model to provide a scientifically plausible model-based approach for sequence alignment, clustering, and so forth.

The SP is typically much less constrained (in terms of time, cost justifications, etc) than the BP. However, while the BP can use aggregate population metrics (such as classification accuracy, squared error, lift, etc.) to evaluate models, the evaluation process for the SP is typically much more subjective in the sense that (a) it is not only the quality of the predictions that matter but also the structure of the learned model itself, and (b) the knowledge captured by the learned model must be evaluated relative to what is already known to the SP and to the SP's research community. The effect of this is to make the overall data mining process much more interactive and human-centered: models are carefully constructed and teased apart in a continuous labor-intensive cycle of model being proposed, evaluated, discarded, revised, verified, and so on.

4 Research Challenges in Data Mining

In the context of the discussion above, there are several challenges that appear to be worthy of attention for data mining in the coming years.

4.1 A General Grand Challenge for Data Mining

A grand vision for data mining is the development of general-purpose *data mining software environments* that assist the user in the overall process of data mining (such as the 5 steps described earlier). The software would ideally help the data miner navigate through the space of possible exploratory steps, modeling steps, algorithm choices, evaluation metrics, and deployment options. The current state of affairs is that for many applications (from ecommerce to climate modeling) the branching factor in terms of selecting specific methods is so high that most novice users are bewildered by the space of possible choices that they can make in the data mining process. The conventional solution to date (e.g., in commercial data mining packages) is typically to support a few standard methods and algorithms at each step. Clearly this can severely constrain how we model our data and in the extreme may be entirely inappropriate for the scientific data miner (where time and space are often important enough that they must be explicitly accounted for in any model).

Development of such a software environment is clearly quite a challenging problem. Statisticians have been thinking about such approaches for quite some time, i.e., general purpose environments for "programming with data" (e.g., Chambers, 1998) as well as graphical model environments that provide flexible and general-purpose high-level languages for model construction (e.g., Gilks, Thomas, and Spiegelhalter, 1994). However, these tools are primarily intended for use by statisticians. To get BP and SP domain experts to use statistical algorithms on a routine basis we need to develop a "next-generation" of interactive user-centered data exploration tools. If we don't, the current situation will continue where only a very small set of algorithms and models are widely-used, and the broader spectrum of modeling and algorithmic techniques are accessible to only a small subset of data miners skilled in these techniques.

4.2 Challenges for Business Applications

4.2.1 Grand Business Challenges

- **Self-Tuning Data Mining Algorithms:** "turn-key" data mining tools that require minimal intervention and tuning from data mining experts. A problem with current data mining algorithms is that they can often require a team of experienced Phd-level researchers to "baby sit" a data mining algorithm to get reasonable results in practice. For data mining solutions to be economically effective in large-scale operational business environments they will need a degree of autonomy beyond what we currently have available. Of course it is not clear that it is even possible to achieve such autonomy, but clearly there are important resource-management issues that need to be considered in this context. For example, a decision-theoretic autonomous agent framework for data mining could be very useful. Such a data mining algorithm could use utilities/probabilities to autonomously decide how much historical data to store, when to update the model, how much data to use in training, which models to use, how to validate and test the models, and so forth. This is of course a rather challenging assignment, given the real-time nature of the environment, the scale of data that is typically involved, and the uncertainties that abound.

4.2.2 Specific Business Challenges

- **Modeling Time:** predictive learning algorithms for time-dependent streams of customer data. For example with clickstream data the current approaches in practice largely involve converting clickstreams into feature vectors so that vector-based algorithms such as decision trees can be utilized. While this is a good engineering approach that takes advantage of existing tools, it cannot take account of many critical aspects of temporal data such as periodicity, seasonality, non-stationarity, and so forth. Models and algorithms that can incorporate these time aspects of customer behavior will tend to be more useful and valuable in the long run. (For an example of such models for predictive modeling with clickstream ecommerce data see Moe and Fader (2000)).

- **Personalization from Sparse Data:** integration of ideas from Bayesian statistics into customer profiling and prediction, e.g., the use of hierarchical Bayesian models for borrowing strength given huge numbers of customers but with very little data for each on average. Such modeling approaches are routinely used in statistics but are virtually unknown in data mining at present.

4.3 Challenges for Science Applications

4.3.1 Grand Scientific Challenges

- **Scalable Exploration:** The importance of the exploration step is easy to underestimate: in many scientific applications this is where the bulk of time on data analysis is actually spent, on tasks such as visualization and clustering, leading to basic theory formation and hypothesis generation. Notions such as outliers, unusual patterns, and trends can play a particularly important role in scientific discovery. Work in database research can play an important role here: intelligent caching for efficient visualization, novel techniques for querying spatio-temporal data, multi-resolution data structures for efficient access to data, and so forth. See Critchlow and Musick (1999) for a general discussion of the role of data management in this context.

As a specific example, in the atmospheric sciences significant research effort is currently expended on constructing and interpreting complex general circulation models (GCMs) of the Earth's atmosphere, oceans, and landmass. These are very complex physical models (including wind and ocean current dynamics, atmospheric chemistry, polar ice-cap models, and so forth) that are used to simulate multivariate gridded measurements over the Earth at roughly daily intervals for a 100-year time-period. The returned data is a huge spatio-temporal multivariate field. Scientists believe these models are quite accurate (at least in terms of long-term climatic time-scales). The models are widely used to investigate basic hypotheses about global warming phenomena (for example): if carbon emissions were increased at rate x how might this propagate through to polar ice-caps, and in turn would this have any effect on numbers and intensities of winter storms?

Despite the success in improving the quality of the underlying GCMs, the methods available for exploring the resultant simulation data tend to be rather primitive. There is no equivalent of CART or C4.5 for easy modeling and exploration. Scientists typically investigate the data in a quite manual fashion by plotting various grids at certain times, and various summary statistics over time. There are clear research opportunities for data miners. However, the problems are quite challenging in that the raw underlying grid data are not the phenomena that the scientists are interested in, but rather the evolution and characteristics of "coherent structures" (such as storms, eddies, and so forth). Furthermore, once detected, these objects have varying spatial and temporal extent, making clustering (for example) a non-trivial problem.

Nonetheless, despite these challenges, there are significant opportunities for data mining research in a broad sense: coupling ideas from pattern recognition and computer vision (for object detection and tracking), from data mining algorithms (quantifying novelty), and from data access and management (how to carry this out in an efficient manner given the data sizes).

4.3.2 Specific Scientific Challenges

- **Models and Languages for Spatio-Temporal Data:** Much scientific data involves spatial and temporal data (and sometimes both). "Non-vector" data sets are inherently more demanding to work with than multivariate vector data from a modeling viewpoint because the branching factor in modeling choice is very high. For example, the modeler must make many decisions about how much memory (in a temporal problem) should be modeled, what representations to choose, etc. While more conventional multivariate data also can have a high branching factor (variable selection, variable pre-processing, etc), this factor can be amplified when time, space, and hierarchies are introduced. Thus, there is a general need for "intelligent assistants" and high-level languages that support scientific data miners as they navigate through large spaces of models and algorithms.

- **Pattern-Finding and Prior Knowledge:** techniques and algorithms that can search massive databases to find unusual structures that are both novel in the context of what is already generally known to the scientist and are useful in some sense. A key word here is "structure": simple associations (such as produced by association rule algorithms) tend to be of limited value for many scientific data sets. Instead, computational biologists are often interested in local subsequences in DNA sequences (motifs) that are significantly different from the background DNA distribution (e.g., Pevzner and Sze, 2000). Similarly, astronomers who study galaxy formation are often particularly interested in finding objects in radio-images of the deep-sky that are morphologically different to known stars and galaxies. While many data mining algorithms focus on searching for *global models* that describe a data set in its entirety, searching for *local structure* (such as motif-finding) is quite common in scientific applications, where only a small portion of the data is of interest. The research challenges are significant: how does one incorporate the prior knowledge of the scientist in an effective manner? what is the appropriate score function for patterns? how does one solve the search problem when potentially looking for a "needle-in-a-haystack"? Is there a general theory for such pattern-finding algorithms or is each application relatively unique?

5 Conclusions

Research in data mining as currently practiced is good at developing specific "black box" algorithms (such as learning algorithms for decision trees, naive Bayes, support vector machines, and so forth). But the "black box" algorithms are only a part of the overall landscape of data mining practice. We also need to be aware of how our tools are used in real applications. Ideally data mining research should focus more on what happens traditionally both before (exploration and modeling) and after (evaluation and deployment) the actual execution of a specific modeling or pattern-finding algorithm. These steps in the process often involve inherently hard problems but also present interesting research opportunities that have significant potential scientific and economic impact.

Acknowledgements

Writing of this paper was supported in part by research awards from the following organizations: NSF (IRI-9703120), NIST Advanced Technology Program, KLA-Tencor, Microsoft Research, and Lawrence Livermore National Laboratories.

References

Chambers, J. M. (1998) *Programming with Data: A Guide to the S Language*, Seattle, WA: Mathsoft.

Gilks, W. R., Thomas, A., and Spiegelhalter, D. J. (1994) A language and program for complex Bayesian modeling, *The Statistician*, 43, 169–178.

Moe, W. W., and Fader, P. S. (2000) Capturing evolving visit behavior in clickstream data, Working Paper Number 00–003 Wharton School of Business, University of Pennsylvania.

Musick, R., Critchlow, T. (1999) Practical lessons in supporting large-scale computational science, *SIGMOD Record*, 28(4), 49–57.

Pevzner, P., and Sze, S.-H. (2000) Combinatorial approaches to finding subtle signals in DNA sequences, *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, Menlo Park, CA: AAAI Press, pp. 269–278.