Speed-up Mining or "Why is data mining iterative?"

David B. Skalak
Senior Data Mining Analyst
IBM Decision Sciences Practice Group
skalak@us.ibm.com
Cornell Theory Center, Rhodes Hall 606
Cornell University
Ithaca, NY 14853

Abstract

Speeding up the data mining process is an important economic goal. In pursuit of that goal, we propose several areas where additional basic research and development may spur progress to improve the speed of knowledge discovery. We examine the propriety of research into three areas: automated control of the mining process, facilitated data preparation, and the automatic identification of inadequate models.

Introduction

The primary perspective of this paper is that the data mining process is iterative and exploratory but that it is not iterative and exploratory completely by necessity. A secondary perspective is that it is a reasonable objective to make the process more straightforward and, to the extent possible, to automate it. These perspectives fall within a more general view of how data mining is used by the mass of firms and governmental agencies. Data mining is often branded a special project, a unique project to be undertaken by a specific group. This technology will be mature when it evolves from a one-shot process that is performed by experts to one that is performed without (much) human intervention and to one that is integrated into business processes of a firm. Just as database and OLAP reports are provided daily to apprise executives and others throughout organization of the up-to-the-day state of the business, we anticipate that data mining models will soon be provided on a daily basis to people holding a variety of positions within an organization. For this to happen, it will be desirable for the knowledge discovery process to be straightened as much as possible, to

eliminate unnecessary iterations.

Of course, there are many other considerations in an effort to bring a sophisticated technology to a large, lay populace, just as for any knowledge management effort [Davenport and Prusak, 1998]. Our goal here is merely to examine three agenda items where additional research may help to automate and accelerate the mining process from extraction to execution.

The purpose of this paper is to suggest several research topics that would need to be addressed to make data mining better understood and more efficient. This paper is conceived as a vision paper, but for each research topic we offer high-level technical approaches that might be successfully applied to the problem.

Data Mining Steps

In the commercial and governmental world, there are steps typically taken to go from data to the fielded implementation of a data mining model. Standard steps for a target marketing campaign might be:

1. Extract the data from one or more data sources
2. Prepare the data: place the data in a format required for input to the data mining software, prepare any meta-data to describe the fields in the data records, impute missing values, remove suspect records, etc.
3. Build a predictive propensity model to assign to each record a score that indicates the propensity of the establishment to purchase some item(s) of interest.
4. Rank the establishments according to their predictive scores
5. Determine and extract the top candidates for campaign treatment.
6. Execute the campaign.
7. Measure the results of the campaign.

Berry and Linoff refer to a similar series of steps as the *Virtuous Cycle* [1997]. While we cannot attest to the virtuousness of this approach, the cycles in the process are well recognized. In an outer, business process loop, measuring results identifies new problems and provides new data to kick-start another round of data mining [Berry and Linoff, 1997]. The inner cycle of returning to a previous step, making changes, and retracing the subsequent steps is the one that reflects the difficulty of data mining. If the analyst realizes that she failed to extract a field from a legacy database, she will have to return to step 1 in the list. If she made an error in assigning classes to the dependent variable, then she will have to return to an early data preparation step, step 2. In general, it is easy to envision a

return from any step to any previous one.

The twists of the mining task have been recognized empirically and have been the impetus for work such as Brachman and colleagues' Interactive Marketing Analysis and Classification System (IMACS) [Brachman and Anand, 1997]. This system supports a human-analyst-centered view of the mining process, by helping keep track of files, mining results, etc.

Each step in this process certainly affects later steps. But the degree and the manner to which each step affects later ones are usually unclear. For example, in step 1, the effect of including or omitting a data source on the ultimate results of the campaign is unknown and unknowable, in general. So at the same time that data mining analysts go round in circles, retracing previous steps, the effect of those temporary setbacks is really not known. They can only be surmised, based on experience. Our thesis is that if we possessed a model of the entire mining process, we would know whether it were worthwhile to return to earlier stages. This issue is not just an academic nicety. Data mining practitioners are in short supply, their time is pricey, and they are under pressure to produce top results for demanding, paying clients. So it is important that the steps taken to produce a data mining result are no more time-consuming and expensive than necessary.

The remainder of this paper nominates several research agenda items that may help meet the goals of understanding, accelerating, and automating data mining. In pursuit of our vision of reducing inefficient cycles in the mining process, we suggest three research topics: modeling the data mining process [Kleinberg, et al, 1998], automating data preparation, and mechanically recognizing the inadequacy of models.

Control of the Data Mining Process

Three possible ways to model the data mining process are: to treat it as (1) a dynamic systems simulation problem; (2) a control problem, suitable for the application of planning methods from Artificial Intelligence (AI), and (3) a control problem, to which reinforcement learning techniques can be applied.

The first approach views the data mining process as a dynamic process that can be modeled using simulation techniques [Sterman, 2000]. Vendors such as Ventana Systems provide software for

modeling business problems as time-dependent dynamic processes [Vensim, 2000]. Simulation has already been applied to the more general problem of modeling project management dynamics, including the evaluation of alternative policies to improve performance [Vensim, 2000].

A second way to study the data mining process is to regard it as a planning problem. One standard AI formulation of planning requires (a) an initial state, (b) a goal state and (c) operators [Russell and Norvig, 1995]. For example, the initial state may correspond to a set of databases that are available for extraction and a formalized description of a business problem. The goal state is to achieve some business objective, such as keeping attrition below 5%. Plan operators correspond to the various data mining actions that can be performed. The steps may be arranged in a hierarchical plan. The step of applying an algorithm to data, for example, requires several sub-steps, including designation of the target data, the selection of parameters, storing and indexing the results, recording comments about the run, etc.

A third and related way to represent the data mining process is as a control problem that may be amenable to reinforcement learning and dynamic programming methods [Sutton and Barto, 1998]. A reinforcement learning problem is characterized by an *agent* that takes *actions*, based on representations of an environment's *state*, and receives a scalar *reward* at each step. Reinforcement learning assumes only that the actions taken by the agent are evaluated, rather than instructing the agent as to the correct response [Sutton and Barto, 1998]. The possibly noisy and weak supervision provided to a data mining analyst may be more faithfully modeled through the reinforcement learning framework than through classical supervised learning.

There are, nevertheless, hurdles to the application of reinforcement learning to the control of data mining procedures. For one, reinforcement learning typically has been applied in well-circumscribed domains, e.g., elevator control [Crites and Barto, 1996]. Its application to an area like data mining may be a leap for these techniques, particularly if large of amounts of data are to be generated through simulation or otherwise to support the learning of an agent's policy that maps states to the probability of selecting each action.

Data Preparation

Anecdotally, it is estimated that 70% (plus or minus 20%, say) of the time used in a data mining project is dedicated to data preparation. Data preparation is a lengthy, often tiresome, stage in a mining engagement, and so limiting the iterations of preparation is a particularly attractive goal.

While outlier identification, missing value imputation, discretization, and other cleansing techniques have received much research attention, other standard preparatory tasks have received less [Pyle, 2000]. For example, common is the aggregation or "rolling up" of a customer's transactional data into a single, summary non-transactional record (Cf. [Howe, 2000]). The transactions of a banking customer may be aggregated so that a single record represents the customer, rather than a series of transactions.

A preliminary question is the level to which the transactional data is aggregated. In a business-to-business setting, a transacting corporation may have one or more physical sites or locations, one or more regional headquarters, one or more affiliated corporations: the appropriate level of representation for a business is not obvious. Situations like this may call for applying a predictive (or other) algorithm at various levels of transactional aggregation and determining the accuracy at each level. In an extreme case, various types of search could be applied to find an appropriate business unit for aggregation. Exhaustive search, heuristic search, or blind search may be useful to find the level of roll-up that leads to the greatest predictive accuracy, especially if different segments of the training population may be aggregated to different levels.

Another problem that arises in the aggregation of transactional data is the value that is imputed for a variable that summarizes a set of transactional records. For instance, suppose a bank customer makes various automated teller transactions. One may want to use as a derived variable the mean amount of cash withdrawn, a robust mean, or the maximum and the minimum. Whether to take these (mean, minimum, maximum) or other summaries of the transactional data to the aggregate level is a mining question whose answer may be found through learning or search.

Research into principled ways to choose the right level of aggregation and the most representative aggregated features would help to eliminate inefficient cycles in data preparation.

Automated Recognition of Inadequate Models

One of the drivers of additional knowledge discovery iterations is the recognition that a model is inadequate. To the extent that a model can be automatically recognized as insufficient, the iterations might themselves be automated.

There are several types of models that may be recognized as defective. The first type is the trivial model. A classic example is the one-node decision tree. Of course, it is easy computationally to recognize a decision tree with one node. Further, there may be also a set of prototypical mining responses to a trivial model. In the case of a one-node decision tree, for example, one immediate check would confirm that the records are not all assigned to the same class.

At the other end of the spectrum of inadequate models is the model that is trivial because each record is placed in its own partition. For instance, if a unique key is assigned to each record and it is then used as a clustering variable, then one-element clusters may result. Again, this type of overfitting is easy to recognize, and could give rise to automatic responses, such as eliminating a variable that appears with a unique value in too many records.

Other types of model faults may be identified. An empirical way to determine associated responses to inadequacy may be to observe data mining analysts in practice and to discern the heuristics that they use to respond to various model shortcomings.

Summary

The objective of this paper has been to suggest several research problems and possible solution paths to make the data mining process less costly by eliminating or automating iterations. We have suggested three areas for additional research: modeling of the mining process, facilitated data preparation, and the mechanical recognition of sub-par models. As the data mining process becomes better understood and more straightforward, it may then be more efficient and more closely integrated into the business processes of organizations.

Bibliography

Berry, M.J.A. and Linoff, G. Data Mining Techniques for Marketing, Sales, and Customer Support Wiley, NY, NY 1997

Brachman, R.J. and Anand, T. The Process of Knowledge Discovery in Databases. In Advances in Knowledge Discovery and Data Mining, Fayyad, M et al., Eds., AAAI Press/MIT Press. Cambridge, MA. 1996.

Crites, R. and Barto, A.. "Improving Elevator Performance using Reinforcement Learning". In *Advances in Neural Information Processing Systems 8* (NIPS8), D. S. Touretzky, M. C. Mozer, and M. E. Hasslemo (Eds.), Cambridge, MA: MIT Press, 1996, pp. 1017-1023.

Davenport, T.H. and Prusak, L. Working Knowledge. Harvard Business School Press, Cambridge, MA. 1998.

Fayyad, U.M.; Piatetsky-Shapiro, G.; Smythe, P. and Uthurusamy, R. Editors. Advances in Knowledge Discovery and Data Mining. AAAI press/MIT Press, Cambridge, MA 1996

Howe, N.R. Data as Ensembles of Records: Representation and Comparison. Proceedings of the Seventeenth International Conference on Machine Learning. Morgan Kaufmann, San Mateo, CA 2000.

Kleinberg, J., Papadimitriou, C., Raghavan P. Segmentation Problems: A Micro-Economic View of Data Mining. Proc. 30th ACM Symposium on Theory of Computing, 1998.

Pyle, D. Data Preparation for Data Mining. McGraw-Hill, NY, NY 2000.

Russell, S. and Norvig, P. Artificial Intelligence: A Modern Approach. Prentice-Hall, Englewood Cliffs, NJ. 1995.

Sterman, J.D. Business Dynamics. McGraw-Hill, NY, NY 2000.

Sutton, R. and Barto, A. Reinforcement Learning: An Introduction. MIT Press, Cambridge, MA 1998.

Vensim Professional Dynamic Systems Simulation. Ventana Systems, Harvard, MA. 2000.