

Data Mining - An Industrial Research Perspective

C. Apte

IEEE Computational Science and Engineering
April-June 1997

Data Mining – An Industrial Research Perspective

Chidanand Apté

T.J. Watson Research Center
IBM Research Division
Yorktown Heights, NY 10598
apte@watson.ibm.com

Abstract

Data mining has burst into the limelight recently, thanks to a series of key application successes [2]. Big business and industry has raised the stakes by investing in this nascent technology, and has laid down high expectations for this emerging area. Although the methods and systems for data mining are based upon and influenced by years of classical work in statistics, pattern recognition, information theory, and machine learning, a combination of factors has caused the recent resurgence of interest. These factors include the availability of high volumes of on-line enterprise data, inexpensive access to high performance computational resources, and continuing impressive advances to the underlying data analysis algorithms. This article describes activities that center around this emerging area of technology, with a focus on research in progress, and applications being pursued.

1 Current State of the Art

Just what exactly is data mining? At a broad level, it is the process by which one extracts accurate and previously unknown information from large volumes of data. This information should be in a form that can be understood, acted upon, and used for improving decision processes of the data owning entity. Obviously, with this definition, data mining is a technology that encompasses a broad set of technologies, including data warehouses, database management, data analysis algorithms, and visualization. The crux of the appeal for this new technology lies in the data analysis algorithms, since they provide the automated mechanisms for sifting through these large volumes of data for extracting useful information. The analysis capability of these algorithms, coupled with today's data warehousing and database management technol-

ogy, make it possible to mine and extract useful knowledge from very large business and industrial data.

The data analysis algorithms (or data mining algorithms, as they are more popularly known nowadays) can be divided into three major categories based upon the nature of their information extraction. These categories are as follows; predictive modeling (*aka* classification or supervised learning), clustering (*aka* segmentation or unsupervised learning), and frequent pattern extraction.

The data representation model for all these algorithms is quite straightforward. Data is considered to be a collection of records, where each record is a collection of fields. Using this tabular data model, the data mining algorithms are designed to operate on the contents, under differing assumptions, and delivering results in differing formats.

Predictive modeling is based upon techniques used for classification and regression modeling. One field in the tabular data set is pre-identified as the response or class variable, and these algorithms produce a model for that variables as a function of the other fields in the data set, pre-identified as the features or explanatory variables. If the response variable is discrete valued, then classification modeling is employed. If the response variable is continuous valued, regression modeling is employed. The principal problem being addressed by this family of algorithms is to be able to produce a predictively accurate function approximation for the response variable, by using the data set as examples of the relations between instances of explanatory variables and the response variable, in the presence of noise. Once produced, the model can be used to predict the value of a response variable, given the specifications for the explanatory variables.

This modeling work has its roots in classical statistics [4, 8], although many recent advances have come from other areas, including pattern recognition, information theory, and machine learning. The important

shift in the modeling paradigm that has taken place here is the shift towards non-parametric techniques, where no assumptions are made about any underlying distributions in the data. These methods include techniques such as neural networks, decision trees, and decision rules. Recent references that describe these methods in detail can be found [6, 7, 9]. The two key technical aspects to all predictive modeling algorithms is their ability to generate models in the presence of noise in the data, and their emphasis on producing an accurate error estimate on the model that is produced. Many techniques have been developed for noise handling and error estimation, and provide the foundational basis for most modern predictive modeling methods.

A further level of interest has been created by data mining applications for decision tree and rule modeling, due to the nature of their modeling representation. Both decision tree and rule modeling algorithms produce outputs in symbolic notation, which is very amenable to inspection and interpretation. This characteristic allows business end users and analysts to understand the underlying decision boundaries in data, and take actions based upon them. Although alternate techniques such as neural networks may produce predictively accurate models, their outputs are highly quantitative in nature, and not easily understandable. However, when evaluating and determining a modeling technique to use from a given set of alternatives, end users have to weigh the compromise between predictive accuracy, level of understandability, and computational demand. Very often, these alternatives need to be traded off against one another, because algorithms often compromise one to gain performance in the other.

Clustering is another major class of data mining algorithms. In this family of techniques, using the same tabular data model described earlier, the algorithms attempt to automatically partition the data space into a set of regions or clusters, to which each of the examples in the table are assigned, either deterministically, or probabilistically. The goal of the search process used by these algorithms is to identify all sets of similar examples in the data, in some optimal fashion.

Obviously, the notion of similarity is highly subjective, and one of the more popular criteria that has been used for identifying similarity has been euclidean distances (k-means, hierarchical), based upon early research in pattern recognition [3, 5] and as of more recent, in Bayesian statistics (AutoClass), and neural networks (Self Organizing Map). In many cases, clustering results can only be judged by the value perceived by end users, and no strict evaluation guidelines

exist, as in predictive modeling.

While the early pattern recognition work produced clustering methods that were adequate for continuous valued space, data mining requirements have generated interest in newer techniques that can operate on variables that need not be continuous valued. Of these, techniques based upon relational voting are gaining in popularity. The idea here is to determine similarity between examples based upon how many features the examples agree upon, and not how close they are in euclidean space. It is obvious that this approach will work well on data sets that have discrete valued variables in them. However, real world data has both continuous valued as well as discrete data, and combinations of euclidean techniques and relational techniques need to be used for business and industrial applications.

The final class of data mining algorithms is for frequent pattern extraction. The goal here is to extract from the tabular data model all combinations of variable instantiations that exist in the data with some pre-defined level of regularity. This approach to data mining was formulated and solidified by a set of recent advances that were initiated in [1]. The basic kind of a pattern to be extracted is typically identified as an association. An association is a tuple of two sets, with a unidirectional causal implication between the two sets, $A \rightarrow B$. Attached with this tuple are two measures, confidence and support.

Confidence is a statistic that measures the fraction of times B exists in the data when A is present. Support is a statistic that measures the number of times A exists as a fraction of the total data. Thus, an association with a very high support and confidence is a pattern that occurs so often in the data that it should be obvious to the end user. Patterns with extremely low support and confidence should be regarded as outliers with no significance. It is patterns with combinations of intermediate values for confidence and support that provide the user with interesting and previously unknown information. Many variations of this basic association pattern have been formulated, with algorithms to extract them. These include patterns that are attached with time stamps, and temporal relations are extracted that may hold significance, either in terms of frequent occurrence, or in terms of frequent matching between groups of temporal patterns.

While this characterization of data mining algorithms is certainly not all inclusive, it does represent a major portion of the analysis. Many other types of data mining algorithms are also making their impact upon applications, deviation detection and attribute focusing being some of them. Many of these

algorithms are statistical algorithms that are geared towards computing normative behavior of the tabular data model, and then using that to focus into either examples or variables that tend to deviate from the derived norm. This identification is provided to the end-user in the form of detected exceptions, which could be valuable information to the user for taking corrective actions.

Finally, for all these data mining algorithms to work, methodologies for data selection, cleaning, and transformation play a necessary and critical role. For data selection, data needs to be extracted from different databases and joined, and perhaps sampled. Once selected, the data may need to be cleaned. If the data is not derived from a warehouse but from disparate databases, values may be represented using different notations in the different databases. Also, certain values may require special handling since they may imply missing or unknown information. After the selection and cleaning process, certain transformations may be necessary. These range from conversions from one type of data to another, to deriving new variables using mathematical or logical formulae. Many times, if a data warehouse does not already exist, this step of selection, cleaning, and transformation, may take up to 80% of a data mining analysis job for a large business data set.

Once the mining is done, visualization plays an important role in providing adequate bandwidth between the results of the data mining and the end user. There are a few generic tools derived from statistics and computer based 3-d modeling that play a useful role in assisting the end user to examine results of the data mining in useful and informative ways. Figure 1 illustrates a *big picture* view of data mining.

2 Key Applications

Utilizing the techniques described in the previous section, a growing body of applications is emerging that is changing the landscape of business decision support. Predictive modeling techniques are best used when a large body of historical data is available, and one uses this data to model a variable of interest, so that this variable may be forecast in future scenarios, and effective actions taken based upon that forecast. Some examples of these are:

- *Risk Analysis*: Given a set of current customers and an assessment of their risk worthiness, develop descriptions for these classes. Use these descriptions to classify a new customer into one of the risk categories.

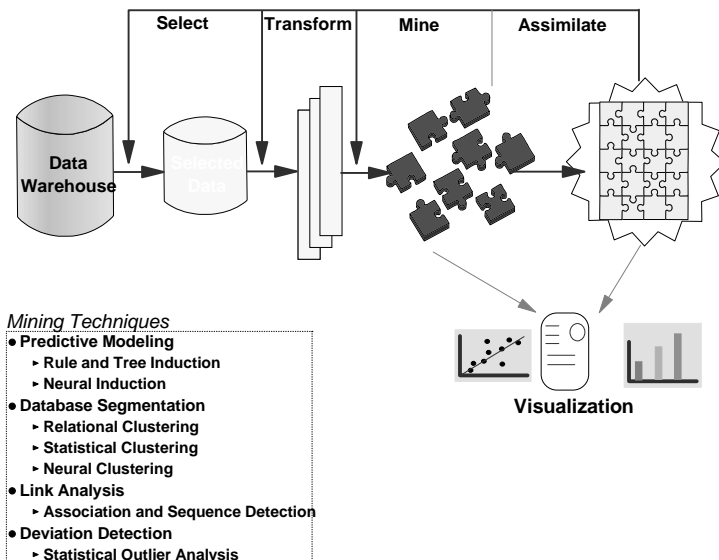


Figure 1: Data Mining: The Big Picture

- *Targeted Marketing*: Given a database of potential customers and how they have responded to a solicitation, develop a model of customers most likely to respond positively. Use the model for more focused new customer solicitation.
- *Customer Retention*: Given a database of past customers and their behavior prior to attrition, develop a model of customers most likely to leave. Use the model for determining the best course of action for these customers.
- *Portfolio Management*: A financial analyst may be interested in predicting the return of investment of a particular asset so that he can determine whether to include it in a portfolio he is managing.
- *Brand Loyalty*: A marketing executive may be interested in predicting whether a particular customer will switch brands on a product he uses.

Using frequent patterns extracted from data, link analysis and item set analysis applications can be built. These could be used for determining business values of promotional effectiveness, services subscriber analysis, demand forecasting. etc. One of the more powerful application of these techniques is in market basket analysis. Databases of sales transactions are examined, to extract patterns that identify what items sell together, what items sell better when relocated to new areas, and what product groupings improve department sales. The benefits of these are

exploited from the resulting store layout and organization, improved inventory management, more effective tie-in promotions and eventually increased department sales.

Clustering based approaches are one of the more pervasive applications of data mining. As databases grow and are populated with more and more data, it is often necessary to partition them into collections of related records for obtaining better summaries of the apparent distinct sub-populations that are present in the data. These applications are most appropriate in marketing planning and promotions, where one constantly wants to monitor and identify the segments in the marketplace served. For example, a bank may want to segment all its retail customers to get a better feel for the demographic and psychographic breakdown (rather than just go by a one dimensional breakdown, such as net assets). Clustering permits the bank to perform the segmentation across a diverse and large set of features (or variables) that the bank has access to for all its customers. Once the clustering is performed, the analyst can examine each cluster more closely, extract significant statistics, and use them in strengthening the bank's offerings on a more individual basis to each of the segments.

This is just a sampling of the many key applications that are being developed, or been deployed, or in inception phase. Needless to say, the underlying techniques are generic enough, and it the creativity of the analysts and application builders that permits the creation of these new and innovative applications.

3 Open Problems and New Directions

As the computer industry goes through its breathtaking phase of introducing new technologies and new solutions to the information processing marketplace, new directions are being enabled in the data mining and decision support arena. One significant enabler is the internet. As companies are beginning to increasingly rely upon the internet for exchange of information, data mining solutions are being crafted that will fit seamlessly in this new medium. One area of application is text mining, which attempts to apply the traditional data mining algorithms, but in the context of non-tabular unstructured data, such as document collections. Clustering and predictive modeling algorithms have been successfully applied to problems such as document indexing and topic identification in the research setting. Applications based upon these ideas are now emerging into the marketplace.

Another area of new applications is in studying and

analyzing internet traffic. Just as sales and bank data could be mined to help the retail store or bank to improve its product offering and marketing effectiveness, internet traffic on a web site can be analyzed to better understand where the real demand is, what pages are being looked at collectively, etc. and this information can be used by the service provider to better organize the site's web pages.

Finally, there is the possibility of mining over the internet, where data mining vendors offer their systems and algorithms as internet servers, and via electronic fee based access, clients can utilize these servers over the internet to mine their data. Early prototypes of these are being benchmarked in the laboratory, and will be soon entering the marketplace.

Research in the underlying algorithms is far from done. Many open problems remain. If one is to use scalability, accuracy, robustness, and interpretability as the criteria to judge data mining algorithms by, then no existing algorithms simultaneously excel in all criteria. This continues to be the holy grail for data mining algorithm research. Can existing techniques be modified, or new algorithms designed, that are scalable (so that the size of data doesn't pose a problem), robust (work well in a wide variety of domains), accurate (information extracted from the data continues to hold up outside and beyond the immediate data), and interpretable (providing insight and value to the end user). Furthermore, research continues in extending and adapting data mining algorithms so that they can operate on a richer collection of data types. Data is no longer just numerical or discrete. It may be unstructured text, or video, or audio, and the collection of these newer data types is dramatically growing. Developing mining techniques for extracting useful knowledge from this new diverse sources of data will keep research humming into the future.

References

- [1] R. Agrawal, T. Imielinski, and A. Swami. Mining Association Rules between Sets of Items in Large Databases. In *Proceedings, ACM SIGMOD Conference on Management of Data*, pages 207-216, 1993.
- [2] U. Fayyad, G. Piatesky-Shapiro, P. Smyth, and R. Uthurusamy, editors. *Advances in Knowledge Discovery and Data Mining*. AAAI Press / The MIT Press, 1995.
- [3] A.K. Jain and R.C. Dubes. *Algorithms for Clustering Data*. Prentics-Hall, 1988.

- [4] M. James. *Classification Algorithms*. John Wiley & Sons, 1985.
- [5] L. Kaufman and P.J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, 1990.
- [6] D. Michie, D. Spiegelhalter, and C. Taylor. *Machine Learning, Neural and Statistical Classification*. Ellis Horwood, 1994.
- [7] B. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996.
- [8] H. Scheffe. *The Analysis of Variance*. John Wiley & Sons, 1959.
- [9] S. Weiss and C.A. Kulikowski. *Computer Systems That Learn*. Morgan Kaufmann, 1991.