

Data Mining Analytics for Business Intelligence and Decision Support

Chid Apte, T.J. Watson Research Center, IBM Research Division

Knowledge Discovery and Data Mining (KDD) techniques are used for analyzing and discovering actionable insights from data. Today, data mining is being increasingly used and embedded in vertical solutions for business intelligence and decision support, such as platforms for Data Management, eCommerce, and critical large-scale solutions (e.g. CRM). This report highlights core algorithms that comprise data mining analytics, describes some business application scenarios for KDD, discusses issues dealing with business intelligence systems, and summarizes trends to watch out for.

1 Background

We are seeing today widespread and explosive use of database technology to manage large volumes of business data. The use of database systems in supporting applications that employ query based report generation continues to be the main traditional use of this technology. However, the size and volume of data being managed raises new and interesting issues. Can we utilize methods wherein the data can help businesses achieve competitive advantage, can the data be used to model underlying business processes, and can we gain insights from the data to help improve business processes? These are the goals of Business Intelligence (BI) systems, and Data Mining is the set of embeddable (in BI systems) analytic methods that provide the capabilities to explore, summarize, and model the data. Before applying these methods to data, the data has to be typically organized into history repositories, known as data warehouses. Data warehousing may require integration of multiple sources of data, which may involve dealing with multiple formats, multiple database systems, distributed databases, cleaning the data, and creating unified logical view of the underlying non-homogeneous data.

Online Analytics Processing (OLAP) is an extension of the Structured Query Language (SQL) framework to accommodate queries that would otherwise have been computationally impossible on a relational database management system. This is achieved by utilizing and storing pre-computed aggregates (e.g. credit-card sales in a certain geographic region over a certain time period) that are automatically updated as the underlying data changes. Deciding upon which aggregates to pre-compute is determined by the business end-user. Providing technical capabilities for automatic computation and updating of aggregates is the strength of OLAP analytics. Data mining analytics try to go beyond OLAP by providing abilities for discovering insights that are computer driven and not end-user driven. Data size is increasing at a rate far exceeding any rates that end-users can cope with. Providing solutions when end-users cannot reasonably supply all possible aggregates to pre-compute, or when it is not possible to express an insight as a pre-computed aggregate, is the goal of data mining analytics.

A collection of early foundational papers on data mining appear in (Fayyad, Piatetsky-Shapiro et al. 1995), and a recent special issue of the Communications of the ACM (Fayyad and Uthurusamy August 2002) contains a collection of papers that discuss current accomplishments and trends.

2 Data Mining

Data Mining may be viewed as automated search procedures for discovering credible and actionable insights from large volumes of high dimensional data. Often, there is emphasis upon

symbolic learning and modeling methods (i.e. techniques that produce interpretable results), and data management methods (for providing scalable techniques for large data volumes). Data Mining employs techniques from statistics, pattern recognition, and machine learning. Many of these methods are also frequently used in vision, speech recognition, image processing, handwriting recognition, and natural language understanding. However, the issues of scalability and automated business intelligence solutions drive much of and differentiate data mining from the other applications of machine learning and statistical modeling.

Typical business intelligence applications of data mining include Risk Analysis (given a set of current customers and their finance/insurance history data, build a predictive model that can be used to classify a new customer into a risk category), Targeted Marketing (given a set of current customers and history on their purchases and their responses to promotions, target new promotions to those most likely to respond), Customer Retention (given a set of past customers and their behavior prior to leaving, predict who is most likely to leave and take proactive action), and Fraud Detection (detect fraudulent activities either proactively or on-line real-time). Many other new application domains are surfacing as we continue to explore and expand upon new data mining opportunities.

It is worth emphasizing that there is more to building a successful business intelligence solution than just data mining. The process of identifying valid, novel, potentially useful, and understandable patterns in data requires one or more of selecting or sampling data from a data warehouse, cleaning or pre-processing it, transforming or reducing it, applying a data mining component to extract models or patterns, and evaluating the derived structure. Data mining is a key component in this methodology that is concerned with the algorithmic means by which structures are extracted from data while meeting computational efficiency constraints.

Key algorithm families used in data mining include predictive modeling (predict a specific attribute (database field) based upon the other attributes (fields) in the data), clustering (also known as segmentation, which groups data records into subsets where items in subsets are more similar to each other than to items in other subsets), frequent patterns (find interesting similarities between a few attributes in subsets of the data), change & deviation (detect and account for interesting sequence of information in data records), and dependencies (generate the joint probability density function that might have generated the data).

Predictive modeling may be defined as the estimation of a function f that maps points from an input space X to an output space Y , given only a finite sampling of the mapping. Typically this translates into predicting the value of a field (Y) in a database based on the other fields (X) in the database. Predictive modeling algorithms are designed to accurately construct an estimator f' of f from a typically finite sample of the data known as the training set. Training data is assumed to be potentially corrupted (i.e. noisy), and appropriate handling schemes need to be built into the modeling algorithm. If the predicted quantity is numeric then the prediction problem is that of regression modeling. If the predicted quantity is discrete then the prediction problem is that of classification modeling. In business intelligence applications, the more generalized form of predictive modeling, probabilistic modeling, is also frequently used since many decision support systems can work better with ranked (by probability) predictions, since they are more easily amenable to optimization and constraint satisfaction analytics.

A number of interesting technical issues are addressed in the design of predictive modeling algorithms. Transformations on the input space X to improve estimation capability are performed by feature extraction, construction, and selection methods. Evaluating the estimate f' in terms of how well it performs on data not present in the training set permits the maximization of prediction accuracy by avoiding under-fitting or over-fitting. Trading off model complexity versus model accuracy is addressed by methods such as bias-variance tradeoff, penalized likelihood, minimum message length (MML), and minimum description length (MDL) encoding.

Classification modeling enables the prediction of the most likely state of a categorical variable (the class) given the values of other variables. This can be viewed as a density estimation problem; i.e., deriving the value of Y given x from the joint density on Y and x . There are several approaches to building classification modeling algorithms, There are kernel density estimators, metric-space based methods (e.g. k-nearest neighbor), and projection into decision regions, i.e., dividing attribute space into decision regions and associating a prediction with each region. Projection methods are by far the most practical for data mining, and these include linear classifiers, neural networks, decision trees, and disjunctive normal form (DNF) rule-based classifiers.

Regression modeling enables the prediction of the most likely value of a numerical variable (the target column) given the values of other variables. This can be viewed as a numerical function approximation problem; i.e., deriving the value of Y given x from the joint probability distribution on Y and x . There are several approaches to regression modeling algorithms. There are statistical probability models (e.g. linear regression), projection into decision regions (dividing attribute space into decision regions and associating a constant value with each region), and hybrid methods (coupling projection methods with statistical methods. Projection and hybrid methods are by far the most practical data mining methods, and these include neural networks, decision trees, and disjunctive normal form (DNF) rule-based classifiers.

Predictive modeling is the most frequently used data mining technique for building business intelligence and decision support solutions. Essentially, it provides a robust and automated mechanism for building forecasting systems in data rich environments. There are three major steps in this data mining process. Historical data is first mined to **train** patterns/models for predicting future behavior. These behaviors can include typical business goals such as predicting response to direct mail, defects in manufactured parts, declining activity, credit risk, delinquency, likelihood to buy specific products, profitability, etc. These patterns/models are then used to **score** new transactions to determine their likelihood to exhibit the modeled behavior. These scores are then used to **act** upon for optimizing a business objective.

Clustering is a data mining technique (also known as segmentation) in which a finite sampling of points are grouped into sets of similar points. Points with common characteristics are essentially "clustered". While predictive modeling required that the target class (or value) membership is known in the training data, in clustering, this knowledge is not known a-priori, and is potentially being discovered by the clustering or segmentation process.

Typical clustering techniques involve a two-stage approach, an outer loop to determine the optimal cluster number k , and an inner loop to fit data points to clusters. Several different algorithmic techniques are available for fitting data points to clusters. Metric distance-based methods find best k -way partitions so that points in a partition are closer to each other than to points in other partitions. Model based methods hypothesize a best fit (very typically probabilistic) model for each cluster. Partition based methods use heuristic scoring function to iteratively enumerate and score various partition scenarios.

The k -mean clustering algorithm is widely used in data mining. Given k cluster centers $c_{1,j}, c_{2,j}, \dots, c_{k,j}$ at iteration j , compute $c_{1,j+1}, c_{2,j+1}, \dots, c_{k,j+1}$, we iterate through two steps. First, cluster assignment, in which *For each $i=1, \dots, m$, assign x_i to cluster $l(i)$ such that $cl(i),j$ is nearest to x_i .* Second, cluster center update, in which *For $l=1, \dots, k$ set $cl,j+1$ to be the mean of all x_i assigned to $c_{l,j}$.* *The iteration stops when $c_{l,j} = c_{l,j+1}$, $l=1, \dots, k$.* Extensions to this algorithm designed specifically for data mining applications include support for scalability, efficient placement of initial k means, and (harder problem) determining the number of clusters k .

Frequent patterns data mining extracts compact patterns that describe subsets of data. These patterns could be either row-wise or column-wise. A widely popular column-wise pattern detection technique is the Association rules data mining technique. Associations mining detects combinations of attribute values that occur with a minimum level of frequency (support) and certainty (confidence). Data mining specific associations algorithms provide scalable capability to find all such patterns in linear time under certain conditions of data sparseness. It should be pointed out that these patterns are not strictly statements about causal effects amongst attributes, but can still provide useful insights in existing large volumes of high dimensional data.

Change and deviation techniques can be used in detecting sequence information in data, where the sequential ordering could be either temporal in nature or some other ordering. The ordering information in the transactions is utilized for computing, under certain conditions of data sparseness, sequences with desired levels of frequency and certainty.

Dependency modeling techniques can be used for detecting causal structure within data. These causal models can be either in the form of probabilistic distributions governing the data, or functional dependencies between attributes in the data. Techniques for discovering causal structures include density estimation methods (expectation maximization) and explicit causal modeling methods (Bayesian networks).

Key business areas that data mining techniques can be potentially applied to include Business Profitability, Customer Relationships, and Business Process Efficiency. For example, discovering who are the best customers for selling products to, most effective market segments for a business, how to increase market share of products, reducing costs without impacting production, and optimizing inventories, are typically instances of recent successful applications of mining.

We give a few examples of how each of the data mining operation maps into potential applications. Predictive Modeling can be used in assigning risk levels to new insurance and financial contracts. Clustering / Segmentation can be used in identifying distinct market groups in customer populations. Frequent Patterns can be used for Market basket analysis (what gets

shopped together in a supermarket). Change and Deviation methods can be used for Fraud discovery in health claim data, and discovering shopping patterns over time. There appear to be numerous business application opportunities in diverse domains, including Retail/Distribution, Healthcare, Manufacturing, Utilities, Telecommunications, Transportation, Government, Cross Industry, and Financial Services.

Comprehensive textbooks on data mining algorithms and analytics include (Han and Kamber 2000), and (Hand, Mannila et al. 2001).

3 Summary: Current Status and Future Directions

The current situation of data mining in the marketplace is that it is primarily an enabler for business intelligence systems. Data mining algorithm suites are available as software packages, some loosely coupled with database technology. To successfully build a data mining application, there is usually heavy emphasis on data warehousing followed by exploratory data mining. The analysis and application building is typically conducted by consultants or in-house analytic teams. The key challenges to the successful completion of a data mining project are the data warehousing requirements, and the sophisticated analytics requirements.

To address these challenges, key research trends in data mining include systems research, for enabling transparent and pervasive usage, algorithms research, for providing scalable, optimized, and robust mining, and solutions research, for embedding data mining into vertically integrated applications.

The key goal in systems research is to enable transparent usage of data mining in environments where data typically resides and is being managed. This includes building database extenders (e.g. User Defined Functions for model training/scoring and sufficient statistics like histograms, counts, samples, etc.), parallel and distributed data mining (for supporting scalability via parallelization and inbuilt sampling), XML based APIs for database coupling and application embedding (to enable interoperability and training/scoring in different environments), and intelligent or semi-automated data warehousing for mining (by providing industry specific templates and meta-data mining).

The key goal in algorithms research is to enable robust and automated data mining, thereby making it easier for non-experts to conduct and run data mining applications. This includes building better techniques for automated evaluation metrics, automated feature extraction / transformation / selection, discovering relational and hierarchical structures amongst attributes, incorporating prior knowledge to account for costs / benefits / uncertainty / missing values, incremental and on-line mining, privacy preserving data mining, and heterogeneous data mining.

The key goal in solutions research is to develop solution specific data mining components that are optimized to the application at hand and can be embedded into a vertically integrated application. Some key application areas that are driving this research include business processes such as risk management, targeted marketing, and portfolio management; systems processes such as computer and network performance management, and Internet processes such as site profiling and performance tuning, and user personalization.

We see two areas in which data mining and operations research (and optimization techniques) will begin to intersect and interact more frequently as the data mining technology matures. While data mining can assist in the automated discovery of actionable insights from data, the efficient execution of the actions can only be effected by coupling the output of data mining with optimization methods. Very often the actionable insights need to be acted upon taking into account business constraints such as budgets and schedules. This can be effectively done only by applying optimization techniques to the outputs of data mining. A classic example is the Mail Stream Optimization (or Horizontal Marketing) solution (Bibelnieks and Campbell 2000; Campbell, Erdahl et al. 2001). A second area where the two disciplines will start interacting more frequently and productively is in the actual design of the data mining algorithms. Many data mining algorithms rely upon heuristic search techniques that are trying to optimize some objective function (e.g. minimizing predictive error on hold-out data). It is natural that optimization methods can contribute here to designing more effective data mining algorithms. A detailed exposition on potential ways in which mathematical programming might be utilized in data mining algorithm design appears in (Bradley, Fayyad et al. 1999). A recent specific example of the usage of low-rank kernel representations in a SVM (Support Vector Machine) classifier appears in (Fine and Scheinberg 2002).

In conclusion, given the current research and solution development directions, we see two key trends emerging; data mining will join Mathematical Programming and Optimization as a key scientific technology for building decision support systems, and using data mining should eventually become as easy and pervasive as working with databases and spreadsheets today.

4 References

Bibelnieks, E. and D. Campbell (2000). "Mail Stream Streamlining." Catalog Age **17**(12): 118-120.

Bradley, P. S., U. M. Fayyad, et al. (1999). "Mathematical Programming for Data Mining: Formulations and Challenges." INFORMS Journal on Computing **11**(3).

Campbell, D., R. Erdahl, et al. (2001). "Optimizing Customer Mail Streams at Fingerhut." Interfaces.

Fayyad, U., G. Piatetsky-Shapiro, et al., Eds. (1995). Advances in Knowledge Discovery and Data Mining, AAAI Press / The MIT Press.

Fayyad, U. and R. Uthurusamy, Eds. (August 2002). Communications of the CACM - Evolving data mining into solutions for insights, ACM Press, New York, NY.

Fine, S. and K. Scheinberg (2002). "Efficient SVM Training Using Low-Rank Kernel Representation." Journal of Machine Learning Research **2**(2): 243-264.

Han, J. and M. Kamber (2000). Data Mining: Concepts and Techniques, Morgan Kaufmann.

Hand, D. J., H. Mannila, et al. (2001). Principles of Data Mining, Bradford Books.