

*Fundamentally, these algorithms are driven by the nature of the data being analyzed, in both scientific and commercial applications.*

# DATA-DRIVEN EVOLUTION OF DATA MINING ALGORITHMS



ata mining is an application-driven field where research questions tend to be motivated by real-world data sets. In this context, a broad spectrum of formalisms and techniques has been proposed by researchers in a large number of applications. Organizing them is inherently rather difficult; that's why we highlight the central role played by the different types of data motivating the current research.

We begin with what is perhaps the best-known data type in traditional data analysis, namely,  $d$ -dimensional vectors  $\mathbf{x}$  of measurements on  $N$  objects or individuals, or  $N$  objects where for each object we have  $d$  measurements or attributes. Such data is often referred to as multivariate data and can be thought of as an  $N \times d$  data matrix. Classical problems in data analysis involving multivariate data include: classification (learning a functional mapping from a vector  $\mathbf{x}$  to  $y$ , where  $y$  is a categorical, or scalar, target variable of interest); regression (same as classification, except  $y$ , which takes real values); clustering (learning a function that maps  $\mathbf{x}$  into a set of categories, where the categories are unknown a priori); and density estimation (estimating the probability density function, or PDF, for  $\mathbf{x}$ ,  $p(\mathbf{x})$ ).

The dimensionality  $d$  of the vectors  $\mathbf{x}$  plays a significant role in multivariate modeling. In problems like

text classification and clustering of gene expression data,  $d$  can be as large as  $10^3$  or  $10^4$  dimensions. Density estimation theory shows that the amount of data needed to reliably estimate a density function scales exponentially in  $d$  (the so-called "curse of dimensionality"). Fortunately, many predictive problems, including classification and regression, do not need a full  $d$ -dimensional estimate of the PDF  $p(\mathbf{x})$ , relying instead on the simpler problem of determining a conditional probability density function  $p(y|\mathbf{x})$ , where  $y$  is the variable whose value the data miner wants to predict.

Traditional modeling methods from statistics and machine learning, including linear regression, logistic regression, discriminant analysis, and Naive Bayes models, are often the first tools used to model multivariate data. Newer predictive models, including additive regression, decision trees, neural networks, support vector machines, and Bayesian networks, have attracted attention in data mining research and applications, as modern computing power has allowed data miners to explore more complex models. These predictive models often sacrifice interpretability for increased flexibility in the functional forms they accommodate. The trade-off between flexibility and interpretability often drives the choice of method applied to a particular multivariate data set.

Recent research has shown that combining different models can be effective in reducing the instability that

---

PADHRAIC SMYTH, DARYL PREGIBON, AND CHRISTOS FALOUTSOS

---

**The navigation  
patterns of Web surfers, obtained from Web  
logs, also represent opportunities for prediction,  
clustering, personalization, and related techniques,  
often referred to as “WEB MINING.”**

results from predictions using a single model fit to a single set of data. A variety of model-combining techniques (with exotic names like bagging, boosting, and stacking) combine massive computational search methods with variance-reduction ideas from statistics; the result is relatively powerful automated schemas for building multivariate predictive models.

As the data miner's multivariate toolbox expands, a significant part of the art of data mining is the practical intuition of the tools themselves [8].

### Transaction Data

A common form of data in data mining in many business contexts is records of individuals conducting “transactions”; examples include consumers purchasing groceries in a store (each record describes a “market basket”) and individuals surfing a Web site (each record describes the pages requested during a particular session). Employing the multivariate viewpoint, we can conceptually view this data as a very sparse  $N \times d$  matrix of counts, where each of the  $N$  rows corresponds to an individual basket or session, each of the  $d$  columns corresponds to a particular item, and entry  $(i, j)$  is 1 if item  $j$  was purchased or requested as part of session  $i$  and is 0 otherwise.

Both  $N$  and  $d$  can be very large in practice. For example, a large retail chain or e-commerce Web site might record on the order of  $N = 10^6$  baskets per week and have  $d = 10^5$  different items in its stores available for purchase or downloading. These numbers pose significant challenges from both the point of view of being computationally tractable and being amenable to traditional statistical modeling. For example, in a store with  $10^5$  different items and  $10^6$  baskets per week, simply computing a pairwise correlation matrix requires  $O(Nd^2)$  time and  $O(d^2)$  memory, resulting in numbers of  $10^{16}$  for time and  $10^{10}$  for memory.

However, data miners routinely take advantage of the fact that transaction data is typically sparse; for example, since the average grocery basket might contain only 10 items, having only a few items in a basket means that only 10/50,000, or 0.02%, of the entries in the  $N \times d$  transaction matrix are nonzero. A substantial body of work in data mining research focuses

on the idea of using subsets of items represented in each market basket, the so-called itemsets  $I$ , as “information nuggets” in large high-dimensional transaction data sets; an example of an itemset is the combination of products bread, wine, and cheese in baskets in a grocery store. Several variants of efficient algorithms are available to find all frequent itemsets from a sparse set of transaction data, work originating with [1]; more recent developments are summarized in [7]. Frequent itemsets are itemsets  $I$  such that  $f_I > T$ , where the frequency  $f_I$  is the number of rows in which all the items in  $I$  were purchased and  $T$  is some preselected count threshold, such as  $T = 0.001 \times N$ .

Another strand of research takes a more statistical view of market basket data as a density estimation problem rather than a search problem. A methodology for finding statistically significant itemsets, that is, itemsets  $I$  whose empirical frequency varies significantly from the frequency expected by a baseline model (see illustrative visualizations at [www.ics.uci.edu/~smyth/cacm02/](http://www.ics.uci.edu/~smyth/cacm02/)). Determining statistical significance in this context is a subtle problem. A Bayesian approach can uncover complex multi-item associations ignored by more traditional hypothesis-testing techniques. It has been used by the U.S. Food and Drug Administration to search large post-market surveillance databases for significant but relatively rare adverse reactions—a good example of the marriage of computationally oriented data mining ideas with more traditional inferential theories from statistics. Increasingly, much of the research work in data mining occurs at this interface of computational and inferential approaches.

Frequent itemsets can also be viewed as constraints on the set of all possible high-order probability models for the data [11]. The technique of maximum entropy estimation provides theoretical framework for estimating joint and conditional probability distributions from the frequent itemsets that can then be used for forecasting and answering queries. Unfortunately, the maximum entropy approach scales exponentially in the number of variables as to model (in both time and memory), limiting the technique in practice to relatively short queries or low-dimensional models.

Viewing transaction data as a sparse  $N \times d$  matrix is a gross oversimplification of the true situation of the true structure of the data in most applications. Typically, real transaction data has significant additional structure at various levels of detail; for example, retail items are usually arranged in product hierarchies, and Web pages can be related to each other (through hyperlinks) or can be instances of a more general database schema. Thus, the columns of a data matrix, such as products and Web pages, can themselves have attributes (such as price and content), as well as implicit inter-item relationships. Similarly, the rows in a transaction data set can also have significant structure manifested by hourly, weekly, and seasonal temporal patterns.<sup>1</sup> While some of these techniques explicitly exploit this structure, many open research challenges remain. Clear, however, is that techniques exploiting special structure in the data are likely to produce much more valuable insights and predictions than techniques that choose to ignore this structure.

## Data Streams

The term “data stream” pertains to data arriving over time, in a nearly continuous fashion. In such applications, the data is often available for mining only once, as it flows by. Some transaction data can be viewed this way, such as Web logs that continue to grow as browsing activities occur over time. In many of these applications, the data miner’s interest often centers on the evolution of user activity; instead of focusing on the relationships of items (columns), the data miner focuses on modeling individuals or objects (rows).

Data streams have prompted several challenging research problems, including how to compute aggregate counts and summary statistics from such data [6]. A related problem is that of incremental learning, whereby a global model is assumed for the data stream, and the model is estimated incrementally as data arrives. A good example of this approach for online adaptation of classification tree models uses analytical probabilistic bounds to guide the degree to which the model needs to be updated over time.

Another aspect of data stream research involves scaling traditional ideas in statistical data analysis to massive, heterogeneous, nonstationary environments. Using large streams of call-record data in, for example, the telecommunications industry, statistical models (called signatures) can be built for individual telephone customers [9]. Note that the collection of customer

signatures resulting from this methodology can be viewed in a database context as a statistical view of the underlying transaction data. Thus, the derived data can help provide approximate (in the statistical sense) answers to queries. Numerous applications of these techniques tackle problems in forecasting, fraud detection, personalization, and change detection.

## Graph- and Text-based Data

The possibility of discovering patterns in large graphs also motivates data mining interest. We can think of representing  $N$  objects as nodes in a graph, with edges representing relationships among objects. Such “data graphs” appear in multiple settings; for example, the Web can be viewed as a graph where nodes are pages and hypertext links are edges. Similarly, user browsing behavior can be viewed as a bipartite graph where nodes are either users or Web pages, and the edges are pages users have visited. An inevitable question arising from a graphical view of the Web is: What kind of structure can be automatically discovered from its topology? Research suggests, for example, the graph structure underlying the Web is distinctly nonrandom and possesses many interesting properties.

Graphs can be represented by an adjacency matrix conveying the nodes as row/column labels and edges as cell entries. Such matrices are indeed large, and fortunately, sparse. That is, all nodes in a real graph are not created equal; some have an extremely high degree, or outgoing or incoming edges, while the vast majority barely have degree 1. If the nodes are sorted according to their degree, the result is often “laws” of the form

$$\text{degree} \propto 1/\text{rank}^a$$

where  $a$  is often termed the “degree” exponent [4].

The matrix representation of a graph suggests that many classical methods in linear algebra are likely to be extremely useful for analyzing the properties of graphs. Indeed, the singular value decomposition is the engine behind many powerful tools, including latent semantic indexing, the “hubs and authorities” algorithm, and Google’s PageRank algorithm. Reflecting what can be discovered from connectivity information alone, PageRank uses a recursive system of equations, defining the importance of each page in terms of the importance of the pages pointing to it. The importance (or page rank) of each page can then be determined by solving this set of linear equations. Once again, sparseness is important from a computational point of view. Since the number of outlinks per page is on average extremely low relative to the total number of pages on the Web, this system of linear equations is sparse, and

<sup>1</sup>Transactions may also be grouped by individual (such as via frequent shopper cards or via cookies for Web data) and by location (such as different stores and different versions of a Web server), providing further hierarchical structure.

## How can the algorithm designer and the scientist represent **PRIOR KNOWLEDGE** so the data mining algorithm does not just rediscover what is already known?

an iterative algorithm typically converges on a solution rather quickly.

Hyperlink connectivity represents only one type of Web data. The navigation patterns of Web surfers, obtained from Web logs, also represent opportunities for prediction, clustering, personalization, and related techniques, often referred to as “Web mining” [10].

Web content, including text documents, is another vast and readily available data source for data mining [2]. Considerable progress in text classification and clustering has been made by representing text as “term vectors” (a vector where component  $j$  is 1 if the document contains term  $j$  and 0 otherwise). Nevertheless, modeling documents at a richer semantic level is clearly worthwhile for, say, trying to identify the relations among sets of objects, such as documents [5].

### Scientific Data

While many data mining applications focus on commercial applications, such as credit scoring, fraud detection, and Web personalization, data mining as a tool for scientific discovery also motivates research interest. For example, data in the form of DNA and protein sequences, microarray-based gene expression measurements, and biological images has revolutionized the fields of biology and medicine. Biologists often spend more time looking at data than through a microscope. Since much biological research is data-rich and relatively theory-poor, data mining research promises significant opportunities for assisting biologists pursuing new scientific discoveries. Rather than viewing the field of computational biology as just applications, data miners find themselves confronted with interesting and fundamental research challenges from a number of perspectives, including modeling, inference, and algorithmic. For example, the discovery of “motifs” in DNA sequences is an example of a biologically motivated data mining problem. Motif discovery can involve prior knowledge as to the number of motifs (such as one per sequence) and their exact or expected lengths. However, little knowledge is typically available as to where the motifs occur in each sequence or what symbols they contain. Related research is driven by development of both score


functions for patterns (to be interesting, a pattern must differ from the background in a systematic way) and efficient search techniques to locate the likely candidates from the combinatorially large space of possible patterns in a set of sequences. Ideas from systematic search, heuristic search, and stochastic search have all proved useful in this context. Several publicly available algorithms are used in computational biology for motif-discovery, each combining basic statistical models with massive search capabilities [12].

Scientists in other disciplines also have an increased awareness of the importance of data mining; for example, in astronomy, the Sloan Digital Sky Survey generates 5TB of data annually, leading to significant data engineering challenges (see [www.sdss.org](http://www.sdss.org)). An important research topic concerning such data is how to develop efficient algorithms to perform common data analysis tasks, including clustering and density estimation, on massive data sets. Multiresolution *kd*-trees, or a flexible data structure for indexing data in multiple dimensions, can provide orders of magnitude speed-ups in the density estimation of astronomical data using mixture models [3].

One research area conspicuous by its absence in data mining research, yet tremendously important in practically any scientific context, is human-computer interaction for discovery; for example, how can the algorithm designer and the scientist represent prior knowledge so the data mining algorithm does not just rediscover what is already known? and How can scientists “get inside” and “steer” the direction of a data mining algorithm? While some research on these topics has been pursued in a number of areas, including artificial intelligence and statistics, it has had relatively little effect on data mining in general.

### Conclusion

The data mining innovations being implemented worldwide often involve collaborations among domain experts, computer scientists, and statisticians. We expect these application-driven developments will continue to proliferate as data owners seek new and better ways to gain insight into their data. We can hope that a more synergistic view of data mining,

combining ideas from computer science and statistics, will gradually emerge to provide a unifying theoretical framework for many of these efforts. 

The contribution by the author Christos Faloutsos is based on work supported by the National Science Foundation under grants IIS-9988876 and IIS-0083148. The contribution by the author Padhraic Smyth is based on work supported in part by the National Science Foundation under grants IIS-9703120 and IIS-0083489, by grants from NASA and the Jet Propulsion Laboratory, the National Institute of Standards and Technology, Lawrence Livermore Laboratories, the University of California, Irvine, Cancer Center, Microsoft Research, and an IBM Faculty Partnership award.

## REFERENCES

1. Agrawal, R., Imielinski, T., and Swami, A. Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD Conference* (Washington, D.C., May 26–28). ACM Press, New York, 1993, 207–216.
2. Chakrabarti, S. Data mining for hypertext: A tutorial survey. *ACM SIGKDD Explor.* 1, 2 (Jan. 2000), 1–11.
3. Connolly, A., Genovese, C., Moore, A., Nichol, R., Schneider, J., and Wasserman, L. Fast algorithms and efficient statistics: Density estimation in large astronomical data sets. *Astronom. J.* (2002).
4. Faloutsos, M., Faloutsos, P., and Faloutsos, C. On power-law relationships of the Internet topology. In *Proceedings of the ACM SIGCOMM Conference* (Cambridge, MA, Aug. 31–Sept. 3). ACM Press, New York, 1999, 251–262.
5. Getoor, L., Friedman, N., Koller, D., and Pfeffer, A. Learning probabilistic relational models. In *Relational Data Mining*, S. Dzeroski and N. Lavrac, Eds. Springer-Verlag, Berlin, 2001, 307–333.
6. Gilbert, A., Kotidis, Y., Muthukrishnan, S., and Strauss, M. Surfing wavelets on streams: One-pass summaries for approximate aggregate queries. In *Proceedings of Very Large Data Bases (VLDB 2001)* (Rome, Italy, Sept. 11–14). Morgan Kaufmann, San Francisco, CA, 2001, 79–88.
7. Han, J. and Kamber, M. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, San Francisco, CA, 2001.
8. Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, New York, 2001.
9. Lambert, D., Pineiro, J., and Sun, D. Updating timing profiles for millions of customers in real-time. *J. Amer. Statist. Assoc.* 96, 453 (Mar. 2001), 316–330.
10. Masand, B. and Spiliopoulou, M. *Web Usage Analysis and User Profiling*. Springer-Verlag, Berlin, 1998.
11. Pavlov, D., Mannila, H., and Smyth, P. Probabilistic models for query approximation with large sparse binary data sets. In *Proceedings of the Uncertainty in AI conference (UAI-2000)* (Stanford University, Stanford, CA, June 30–July 3). Morgan Kaufmann, San Francisco, CA, 2000, 465–472.
12. Wang, J., Shapiro, B., and Shasha, D. *Pattern Discovery in Biomolecular Data: Tools Techniques and Applications*. Oxford University Press, New York, 1999.

**PADHRAIC SMYTH** (smyth@ics.uci.edu) is an associate professor in the Department of Information and Computer Science at the University of California, Irvine.

**DARYL PREGIBON** (daryl@research.att.com) is a member of AT&T Labs, Research, Florham Park, NJ.

**CHRISTOS FALOUTSOS** (christos@cs.cmu.edu) is a professor in the Department of Computer Science at Carnegie Mellon University, Pittsburgh, PA.