Data Mining: Data Analysis on a Grand Scale?*

Padhraic Smyth
Information and Computer Science
University of California, Irvine
CA 92697–3425

July 6, 2000

July 6, 2000

Technical Report UCI-ICS 00–20

Information and Computer Science

University of California, Irvine

*Revised version to appear as Review Paper for *Statistical Methods in Medical Research*, September 2000

Abstract

Modern data mining has evolved largely as a result of efforts by computer scientists to address the needs of "data owners" in extracting useful information from massive observational data sets. Because of this historical context, data mining to date has largely focused on computational and algorithmic issues rather than the more traditional statistical aspects of data analysis. This paper provides a brief review of the origins of data mining as well as discussing some of the primary themes in current research in data mining, including scalable algorithms for massive data sets, discovering novel patterns in data, and analysis of text, Web, and related multi-media data sets.

1 Introduction

The phrase "data mining" has had a varied history within the past 30 to 40 years. In the 1960's, as digital computers were beginning to be applied to data analysis problems, it was noticed that if one searched long enough (using the computer) that one could always find some relatively complex model to fit a data set arbitrarily well. This could of course happen even if the resultant model were entirely spurious and did not represent any true underlying structure, e.g., if the data were entirely random in nature (Armstrong, 1967). Thus, terms such as "data mining" and "data dredging" were coined to describe such activities, along with related terms such as "data snooping" and "data fishing" (Selvin and Stuart, 1966; Lovell, 1983). In fields such as econometrics the term "data mining" still has quite a negative connotation (e.g., Leamer, 1978; Hendry, 1995, Chapter 15.1).

Nonetheless, despite this history, by the early 1990's the term "data mining" was somewhat independently adopted by computer scientists to describe algorithmic and database-oriented methods that search for previously unsuspected structure and patterns in data. The data sets involved are often (but not always) massive in nature. A precise definition of this notion of data mining is quite difficult to pin down, since as currently practiced it encompasses quite a wide variety of data analytic techniques and methods without any necessarily single coherent theme. Nonetheless, it is fair to say that much of this modern work in data mining can be characterized as placing a significant emphasis on the role of algorithmic and computational issues in data analysis, rather than on more traditional statistical issues such as inference and estimation. Another distinction is that data mining is almost always practiced in a retrospective manner on observational data, and does not involve considerations of experimental design and related concepts.

In this paper we will adopt this modern (computer science) usage of the term data mining. We will largely focus on data mining as evidenced by research published in the mainstream conferences and journals in the field, e.g., the annual ACM International Conference on Knowledge Discovery and Data Mining, the annual ACM Conference on Management of Data (SIGMOD), and the Journal of Data Mining and Knowledge Discovery. It is currently fashionable to attach the term data mining to research papers or product descriptions that involve data analysis in any form. For example, consider a medical research project which uses the term data mining in the title of the study or paper to describe building classification trees (for example) on a relatively small data set for medical diagnosis. Such papers are not of primary interest in this review since they can equally well be viewed as the application of relatively well-known ideas in applied statistics.

In this review paper we will instead focus on themes and strands of work which can be viewed as relatively unique to data mining and which complement traditional statistical methods. The paper

begins in Section 2 with a brief review of general resources in the area of data mining and continues in Section 3 with a brief history of the field. The following three sections discuss some of the main themes in data mining as currently practiced, focusing in particular on scalable algorithms (Section 4), finding patterns in data (Section 5), and text and Web data applications (Section 6). Section 7 contains concluding comments. As with any general review, many of the comments contains in this paper are subjective in nature and cannot be backed up with theorems!

2 Other Reviews and Resources for Data Mining

In terms of other general reviews of data mining, there are several which complement the viewpoint of this paper. From a statistical perspective Hand (1998) and Glymour et al. (1996, 1997) discuss aspects of the general relationship between data mining and statistics. Fayyad et al. (1996) provide an overview of the state of the field in 1996. In the area of automated machine discovery, Valdez-Perez (1999) discusses the role of automated discovery systems in science. A number of recent books have also appeared on data mining, largely emphasizing business and marketing applications of data mining algorithms and intended primarily for a non-technical business audience (e.g., Adrians and Zantige (1996), Berry and Linoff (1997)). Other texts such as those of Weiss and Indurkhya (1997) and Witten and Frank (1999) provide more of a research-oriented viewpoint on data mining, but from a largely machine learning (computer science) perspective. To date there is no text available which treats data mining in a statistical context. The online Web site (and associated newsletter) `www.kdnuggets.com` provides many online resources covering both commercial and research activity in data mining.

The evolution of research in data mining can be traced through a series of workshops and conferences entitled "Knowledge Discovery in Databases" (KDD). The first few workshops (e.g., Piatetsky-Shapiro, 1991) were relatively small (approximately 50 attendees) and were motivated by a realization among machine learning and AI researchers that technology was beginning to change the nature of data collection and analysis. "Data owners" such as scientists, businesses, and medical researchers, were able to gather, store, and manage previously unimaginable quantities of data due to technological advances and economic efficiencies in sensors, digital memory, and data management techniques. As data volumes and archives began to grow very rapidly in the 1990's, so too did interest from data owners in the research conducted under the KDD umbrella. In 1994 the first International Conference on Knowledge Discovery and Data Mining was held (Fayyad and Uthurasamy (1995)). It has evolved into the primary annual forum for data mining research (Simoudis and Han, 1996; Heckerman, Mannila, and Pregibon, 1997; Agrawal and Stolorz, 1998; Chaudhuri and Madigan, 1999; Ramakrishnan and Stolfo, 2000). Edited research papers from the early KDD workshops were published in two edited volumes, Piatetsky-Shapiro and Frawley (1991) and Fayyad et al. (1996), providing snapshots of early research in the field.

3 A Brief History of Data Mining

It is worthwhile to begin by reviewing briefly the evolution of modern data mining. From a statistical perspective perhaps the most noticeable feature of data mining research is the emphasis on computational aspects of data analysis, in concert with a relative lack of emphasis on traditional

statistical concepts such as sufficient statistics, likelihood, or model diagnostics. This "computational culture" is a direct consequence of the fact that data mining has been (to date) largely driven by computer scientists.

3.1 Data Mining and Machine Learning

Within computer science, two particular subfields have contributed most heavily to the development of data mining in the past 10 years, namely, machine learning and databases. Machine learning involves the study of how machines and humans can learn from data and has been an important component of research in artificial intelligence (AI) since the inception of AI in the 1950's. Early work in this field was strongly linked to theories in cogntive science, trying to build algorithms and machines which could adapt to data in a manner thought to be similar to human learning (see Russell and Norvig (1995), Chapter 1, for a review). In more recent years (since the early 1980's) much research in machine learning has shifted from modeling how humans learn to the more pragmatic aims of constructing algorithms which learn and perform well on specific tasks (such as prediction). Naturally this has led to a much greater overlap with applied statistics, with particular emphasis on classification (discrimination) techniques, but again with somewhat of a computational flavor. For example, machine learning research has traditionally placed an emphasis on the human-interpretability of any model which is learned from data, leading to much work on predictive models such as trees and rules which can (for example) be readily understood by clinicians in a medical context, at least for relatively simple trees and rule sets.

The work of Quinlan (1993) on decision tree classifiers, largely paralleling the more statistically motivated work of Breiman et al. (1984) on CART, is a good example of how similar methodologies and algorithms were pursued largely independently by researchers in both machine learning and statistics. Within machine learning, artificial neural networks (Bishop, 1995; Ballard, 1997), nearest-neighbour classifiers (Aha, Kibler, and Albert, 1989; Atkeson, Schaal, and Moore (1997)), simple conditional independence models such as naive Bayes (Duda and Hart, 1973; Domingos and Pazzani (1997)), and (more recently) support-vector machines (Scholkopf, Burges, and Smola, 1999), have all been widely researched. In recent years, statisticians and machine learning researchers have sought out common ground, such that the boundaries between applied statistics and machine learning are more blurred than in the past (e.g., Michie, Spiegelhalter, and Taylor, 1994; Bishop, 1995; Ripley, 1996; Mitchell 1997). It is noteworthy, however, that for largely historical reasons certain standard statistical techniques, such as logistic regression for example, have received little attention in the machine learning literature. In addition, there are many aspects of statistics which are largely absent from the machine learning literature. Because of this lineage, researchers in machine learning (and subsequently in data mining) typically have formal backgrounds in computer science but may have little background in modern statistical methods other than standard undergraduate coursework. Thus, to a statistician, many papers on data mining may appear to be written in a foreign language, with much discussion of algorithms and computational complexity but relatively little in terms of mathematical characterization of the statistical aspects of the problem. Nonetheless, despite the nature of the presentation, these papers can contain useful ideas and methodologies for statistically-oriented researchers.

The significance of machine learning to present-day data mining lies in the fact that many of the researchers involved in data mining, and many of the algorithms being used in data mining,

have their intellectual roots in machine learning. This partly explains (for example) the prevalence of tree-based and rule-based algorithms in data mining papers and software tools, and the relative paucity of many of the more traditional statistical concepts such as parameter estimation, maximum likelihood, hypothesis testing, and so forth. Just as the interests of certain applied statisticians in the 1990's led to significant "crossover" work between computer science and machine learning (e.g., Geman, Bienenstick, and Doursat, 1992; Breiman, 1996, Friedman, 1997), there are similar elements of crossover work between statisticians and data miners in data mining (e.g., Du Mouchel et al. (1999)), although to date on a smaller scale. In the commercial sector, vendors of statistical software packages have been quick to note the advantages of including the phrase "data mining" in their product names and promotional literature, although it is not obvious that these packages contain much that is conceptually different from the older "non data mining" versions.

3.2 Data Mining and Database Research

Another strand of data mining research emerged in the 1990's within the database research community, somewhat independently and largely in parallel with developments in machine learning. Database research got underway as a research field in the 1960's as computer scientists realized that applications which relied on transaction processing (such as banking, airline reservations, and so forth) could not be readily handled using simple collections of relatively independent and loosely coupled files. The introduction of relational database concepts (Codd, 1970) and high-level data models (Chen, 1976) proved to be major conceptual breakthroughs in the field, providing general and principled frameworks for data modeling and access. Topics such as updating the database in a systematic manner, answering structured queries about the data, controlling access and security in the context of multiple users, and so forth, became the foundations of modern database management. By the late 1980's and early 1990's, relational database technology had successfully established itself in the commercial sector, i.e., many businesses and organizations were now using these relational models and tools to manage their data. Worth noting is the fact that these relational database systems were never explicitly designed to support data analysis tasks. Instead they are primarily designed for the purposes of storage, query, and transaction management, i.e., supporting day-to-day operations of organizations that handle large volumes of data (e.g., airlines, banks, hospitals, retail organizations, etc.).

In transactional business environments (such as banks, etc.) interest in *data warehousing* began to grow in the early 1990's (Inmon, 1996), namely maintaining a historical repository of all transactions which had ever been recorded. Database researchers quickly realized that now not only did their customers want to store, manage, and access their data in a systematic fashion, but now they also wished to be able to analyze it. This analysis could not take place in the traditional statistical fashion since these data sets were typically far too large to be handled by conventional statistical software packages. Thus was born the concept of data analysis algorithms which are designed to operate directly on relational databases, forming the main component of modern database-oriented research in data mining.

The paper by Agrawal, Imielinski, and Swami (1993) on association rule mining is probably the earliest example of such work, demonstrating how simple association rules can be "mined" from a relational database in an efficient manner. An example of an association rule is "if a individual purchases bread and milk then they are likely to also purchase butter with probability 0.8." This

early work on association rules spurred significant interest in the database research community, and data mining attained an increasingly significant presence at database research conferences such as SIGMOD by the late 1990's. This strand of work is largely characterized by an emphasis on very efficient data structures and algorithms for operating on data which is not resident in main memory (typically on a disk, perhaps stored in a relational database), and searching for sets of simple local patterns such as association rules. More recently there has been more infusion of statistical ideas in the database research community, involving for example development of computationally efficient algorithms for algorithms such as classification trees and mixture modeling. For example, Gehrke et al. (1999) report substantial computational and memory efficiencies in their implementation of CART using special-purpose data structures, and apply their algorithm to data sets involving millions of points. In a similar fashion, Bradley, Fayyad, and Reina (1998) describe a heuristic algorithm for an implementation of the Expectation-Maximization (EM) algorithm applied to Gaussian mixture modeling on massive data sets, which seeks to minimize the number of passes through the data set.

Just as the influence of machine learning research on data mining has led to somewhat of a bias towards classification problems, the database influence has led to an emphasis on the data access aspects of analyzing massive data sets. Overall this has been a positive contribution in the sense it has led to an increased awareness among data analysts that the traditional approach of viewing the data as existing in a single "flat file" often does not scale very well to massive data sets. For example, data on a group of medical patients may well be distributed across different tables, located on physically different storage devices. The beauty of database technology is that the user is isolated from the details of where such data are stored. The user simply issues queries (in a formal representation language such as the Structured Query Language (SQL)) and the database management system then takes care of the details of finding and returning the relevant data in as efficient a manner as possible.

In the context of data mining research one of the larger issues to be faced is whether this general "standard interface" approach can support sophisticated statistical modeling. At present the answer is no, in the sense that a conventional query language such as SQL provides a relatively awkward and potentially inefficient interface for performing the underlying mathematical operations inherent to statistical modeling (e.g., matrix operations for linear regression). Thus, it is somewhat of an open question as to whether it is better to develop special purposes "SQL-like" languages for data mining or instead to focus on algorithmic ideas (such as sampling) which minimize interaction with the database and perform traditional statistical analyses using traditional tools and environments on reduced data sets which can fit in main memory. While there are various trade-offs involved it is again worth noting that the existing database interfaces (such as SQL) were not originally designed for supporting statistical model building.

In next three sections we discuss specific strands of research in data mining which may be of interest to researchers involved in medical data analysis, and which involve concepts and techniques which are largely outside the mainstream statistical literature (at least at present).

4 Scalable Algorithms for Massive Data Sets

4.1 General Challenges Imposed by Massive Data Sets

As mentioned earlier, one of the main challenges in dealing with massive data sets is the scaling effects which often occur as data sets grow in size. For simplicity, assume we have an $N \times p$ data matrix with p measurements (variables, columns) characterizing each of N objects (individuals, rows). When we talk about massive data sets we often implicitly assume we are talking about very large values of N (e.g., data on several hundred thousand patients) but it is important to note that many of these massive data sets also may contain large numbers of measurements (p) as well, e.g., up to several hundred test results on patients in a medical study. The *time complexity* of a data analysis algorithm is typically expressed in a worst-case sense as a function of N and p and any other parameters which may enter into the algorithm or the modeling, e.g., $O(Np)$ for an algorithm which is linear in both N and p. Algorithms whose time complexity scales poorly as a function of N (e.g., as N^2 or N^3) are often completely impractical for large data sets, e.g., hierarchical clustering algorithms typically scale as $O(N^2)$ in both time and memory. Sensitivity to p is slightly better since p is typically not as large as N: $O(p^2)$ is often fine for many problems, but $O(p^3)$ or higher will begin to be problematic for p of the order 10^3 or greater. Thus, data mining researchers interested in massive data set applications often focus on algorithms which scale in the "near-linear" range in N and usually no worse than p^2 in p (see Huber (1997) for further discussion).

The other relevant aspect of data analysis for large data sets concerns the physical storage location of the data relative to the central processing unit (CPU). In simple terms, we can think of two primary types of storage media (memory) in a computer system—in reality there can be other distinctions such as cache memory, tape-memory etc, but here we just focus on this simplified viewpoint. Primary memory consists of random-access memory (RAM) and has the benefit of allowing relatively fast random access of any bytes stored in RAM, on the order of 10^{-7} to 10^{-8} seconds with current technology. Specifically, this is how long it takes the system to bring the data from memory to the CPU, after which a computation can be performed. Secondary memory consists (for our purposes) of disk storage. The access time here (how long it takes to access a random location on the disk) is on the order of 10^{-2} seconds. There are many other issues involved here, and storage technology is constantly changing, but nonetheless this relative difference in access time between primary and secondary memory is fairly fundamental and is predicted to remain on the order of 10^4 to 10^5 (Gray and Shenoy, 2000). An analogy would be that if the data in primary memory are thought of as being on the bookshelf in your office within 1 meter of your hand, the data in secondary memory are effectively 100 kilometers away!

Thus, in determining the time complexity above, we can think of the physical location of the data (whether primary or secondary memory) as affecting the overall complexity by a multiplicative constant proportional to the average access time, e.g., if the algorithm requires one computation per data point, and each data point is accessed randomly, then the time taken by the algorithm will be proportional to cN, where N is the number of data points and c is the time taken to access the data point (to bring it to the processor). This is somewhat of a simplification, but illustrates the main point that algorithms which frequently access the disk will be much slower than algorithms which operate on data entirely in main memory. If we can organize the data so that it can be sequentially scanned from the disk then the cost of disk access decreases, since sequential scanning of a disk

can be carried out much more efficiently than random access of the same amount of data. But many widely used data analysis algorithms either repeatedly access different subsets of the data in an unpredictable manner (such as classification trees) or require multiple passes through the entire data set (e.g., applications of the EM algorithm). Even if such algorithms scale reasonably in N and p, while they may run in reasonable time on data in main memory they will typically be impractical for large data sets which exceed main memory capacity.

Of course what constitutes small or large depends on the context. It is quite easy to now have 1 Gbyte (10^9 bytes) of RAM (primary memory) on a modern workstation (compared to machines with only 64 kilobytes of RAM 20 years ago). The secondary memory problem arises if one's data set is too large to be read from disk into available primary memory. For example, many retail transaction data sets and many image data sets are in the terabyte range (10^{12} bytes). Clearly, such data sets are well beyond the realm of what most statisticians are used to thinking about and the rules of data analysis for such data sets may be different both from an organizational and mathematical viewpoint. An important statistical issue, which we will only mention in passing here, is the fact that as data sets become this large, homogeneity assumptions (such as independent and identically distributed measurements) become less reliable. There is relatively little work in data mining focusing on such statistical aspects of massive data sets, although it seems clearly that statistical methods such as hierarchical models may be ideally suited to deal with such heterogeneity.

4.2 Scalable Versions of Existing Algorithms

The primary consequence of the above discussion on memory is that a naive implementation of many data analysis algorithms will spend a large fraction of time waiting for data to be transferred from disk when faced with massive data sets. One approach to this problem in data mining has been to develop new versions of existing data analysis algorithms which provably return the same results as the original algorithm, but which involve data management strategies which minimize the overall amount of time spent accessing data. An example of this general approach is that of Gehrke et al. (1999) who propose a family of algorithms called BOAT (Bootstrapped Optimistic Algorithm for Tree Construction). The BOAT approach uses two scans through the entire data set. In the first scan an "optimistic tree" is constructed using a small random sample from the full data (and which can fit in primary memory). The second scan then takes care of any differences between the initial tree and the tree which would have been built using all of the data: the resulting tree is then the same tree that the naive algorithm would have constructed (in a potentially inefficient manner). (The details of how this is achieved in two scans is beyond the scope of this paper, involving various clever data structures to keep track of tree-node statistics). By explicitly focusing on how to deal with data in secondary memory this approach allows well-known algorithms to be scaled-up to massive data sets in a relatively efficient manner. For example, Gehrke et al. report fitting classification trees to 9-dimensional synthetically-generated data sets with 10 million data vectors in about 200 seconds. In a similar vein, the work of Moore and Lee (1998) on cached sufficient statistics for multi-variate categorical data takes advantage of clever data structures to efficiently store information on a full data set in a greatly reduced form. Computational speed-ups of 50 to 5000-fold on various classification algorithms (compared to naive implementation of the algorithms) have been reported (Moore, 1999), where again the final model returned by the algorithm is exactly the same as that which would have been returned by the naive implementation.

However, there are many algorithms which are not so amenable to scaling in this manner. Iterative algorithms such as EM may require many scans through the full data set to converge, and full scans may be quite expensive. Thus, researchers have turned to heuristic algorithms which are scalable to massive data sets but cannot be guaranteed to produce the same result as the original naive algorithm. For example, Bradley, Fayyad, and Reina (1998) describe a heuristic algorithm for scaling both the k-means clustering algorithm and EM-based Gaussian mixture modeling to massive data sets which are not resident in primary memory. In their approach, the algorithm samples the data to find regions of high density and then gradually constructs the cluster model or mixture model while minimizing scans over the full database.

4.3 Novel Scalable Algorithms

A different approach to developing scalable data mining algorithms has been to invent new data analysis algorithms which can be easily supported by conventional database interfaces. The best known example in this category is the afore-mentioned framework of association rules for transaction data. A transaction data set typically consists of N transactions recorded over time. Each transaction consists of a set of individual activities or purchases which occurred during a single "session," e.g., a list of items purchased at a retail store, a list of financial transactions conducted during a single session at an automated teller machine, or a list of prescription medicines authorized by a doctor after examining a patient. Thus, for each of the N transactions the data set typically contains the list of items "transacted," the names or identification numbers of the persons involved, the time and date, and other details such as the price of individual items and so forth.

A simple view of such data is as a large binary $N \times p$ matrix, where there are p individual items, and there is a 1 in entry (i, j) if item j was involved in transaction i, and 0 otherwise. Thus, this matrix is typically very sparse, e.g., in retail environments we may have $p = 50,000$ individual products (items) that one could purchase, but a typical transaction may only involve on the order of 10 of these products. An association rule consists of a simple statement of the following form:

$$\text{IF items } \pi \text{ are purchased, THEN item } j \text{ is also purchased with confidence } p \qquad (1)$$

where π is a set of items (columns in the matrix), not including item j, and p is usually interpreted as the conditional probability $p(j|\pi)$, i.e., the conditional probability of item j being purchased given that items π were purchased. The joint probability $p(\pi, j)$ is often referred to as the *support*. Both the support and confidence of a rule are estimated empirically from the data.

The basic idea behind association rule algorithms is to find *all* association rules in the data which have support above some threshold t_S and confidence above some threshold t_C, e.g., $t_S = 0.1$ and $t_C = 0.8$. Sets of items which have joint probability greater than the support threshold are known as *frequent itemsets*. To find all itemsets of size k one can take advantage of the fact that for an itemset of size k to be frequent, all its subsets of size $k-1$ must also be frequent (a simple consequence of joint probability). Thus, given a list of frequent itemsets of size $k-1$, one can generate a candidate list of itemsets of size k by checking that all subsets of itemsets of size k are themselves frequent. This process of generating candidate itemsets is carried out based on the itemsets alone, and does not involve scanning the data. Once the candidate itemsets of size k have been found, the database is scanned to find the actual empirical support for each of the candidate itemsets. Since counting is a relatively simple and standard operation for databases, this can be

carried out in a computationally efficient manner (e.g., linear in N and p). The algorithms typically search the rule-space in a systematic manner, starting at itemsets of size $k = 1$ and incrementally increasing k, where moving from k to $k + 1$ involves both the generation of candidate itemsets and the scanning of the database to find those that are frequent. Typically for sparse data (and a support threshold value t_S that is not "too large") the number of frequent itemsets will be zero above a relatively small k value, e.g., $k \approx 10$. After all frequent itemsets are found, the algorithm makes one final pass through the database to determine which of the frequent itemsets correspond to association rules with confidence above t_C. Agrawal et al. (1996) report results on synthetic data involving 1000 items and up to 10 million transactions. They empirically demonstrate on these data sets that the computation time scales up linearly as a function of the number of transactions. Similar results have since been reported on a wide range of sparse transaction data sets and many variations of the basic algorithm have been developed (e.g., Brin et al. (1997, 1999)).

The work on association rules differs from more traditional statistical analysis of binary data in two significant aspects. Firstly, the emphasis is on *patterns* (in the form of rules) rather than on *global models* such as a log-linear model. The algorithms produce a set of rules or patterns, which are local in the sense that they apply to specific regions of the p-dimensional multi-variate space. Because the rules are local and evaluated individually, there is no notion of how the set of rules can be combined in a coherent manner for interpretation or prediction. In other words, since the rules are found individually, there is no attempt made by the algorithm to integrate them into a model, e.g., for the purposes of prediction. One approach here is to view the rules as constraints on a large p-dimensional contingency table and use iterative proportional fitting to construct joint probability models which are consistent with these constraints (Pavlov, Mannila, and Smyth (1999)). The resulting model can then be used for prediction. The generalization of this idea is to extract simple summaries of the data in a computationally efficient manner (e.g., based on counting operations) and then to build a model from the resulting summary data.

The second non-traditional aspect of association rules is the emphasis on computational efficiency rather than on the interpretation of the results. It is fair to say that in published work on association rules that most of the emphasis in evaluating different algorithms is placed on computational efficiency (e.g., run-time as a function of N or p) with little or no emphasis on the interpretation of the actual rules returned. This points to a potential problem with the application of such methods in general, namely, that users with little knowledge of statistics may interpret the association rules in an inappropriate manner, e.g., perhaps mistaking correlation for causation. Indeed, while association rules have been one of the primary success stories in data mining and are now available in various data mining software toolkits, it is difficult to find any specific published reference which describes a successful application of the method to a real problem, i.e., the question can be asked as to what association rules are good for exactly? The answer may be that the method is primarily a computationally efficient exploratory data analysis technique for massive transactional data sets.

Novel scalable algorithms also exist for other types of data analysis. For example, the BIRCH algorithm of Zhang, Ramakrishnan, and Livny (1998) provides a scalable approach to clustering, which is similar in spirit to k-means, but which is essentially a new clustering algorithm in its own right.

4.4 Other Approaches to Scaling

There are a number of other general approaches to developing scalable algorithms which have been proposed in data mining, e.g.,

- The obvious idea of running the algorithm on a smaller random sample of the full data set is often used in practice, especially for data analysis tasks involving iterative and interactive phases of model-building. Note that merely generating a random sample from a large database stored on disk may itself be a non-trivial task from a computational viewpoint.

- Du Mouchel et al. (1999) propose a statistically-motivated methodology for "data-squashing" which amounts to creating a set of M weighted "pseudo" data points, where M is much smaller than the original number N, and where the pseudo data points are automatically chosen by the algorithm to mimic as closely as possible the statistical structure of the original larger data set. The method is empirically demonstrated to provide one to two orders of magnitude reduction in prediction error on a logistic regression problem compared to simple random sampling of a data set. This is similar in spirit to ideas in Moore and Lee (1998), Bradley, Fayyad, and Reina (1998) and Pavlov, Mannila, and Smyth (1999), namely generating a smaller approximate representation of the original large data set which in some sense matches the statistical characteristics of the original data set as closely as possible. One advantage of this general approach is that once the reduced data set is created, the original data set can in effect be "thrown away" and computationally intensive visualization or model-building (e.g., using cross-validation methods for model and parameter selection) can take place entirely on the reduced data set in main memory.

- For non-stationary data sets which are collected over time, an online recursive approach is often quite effective, i.e., "pipelining" the data through the analysis system as it arrives and recursively updating model parameters in an online adaptive fashion. Cortes and Pregibon (1998) describe an impressive system at AT&T which adaptively updates estimates on whether a telephone line is a business or a residence, for about 350 million customers per night, based on about 300 million records of daily phone calls. Logistic regression models are trained offline (on numbers whose business/residence classification is known) and the probability of a number being a business is modeled by a logistic regression model with input variables based on characteristics of calls, such as time of day, length of calls, etc.

- Provost and Kolluri (1999) describe a variety of other techniques for scaling up to massive data sets, including technology-driven approaches such as using parallel computing for data analysis problems which can be parallelized.

5 Pattern Discovery Algorithms

Another general area of work in data mining has focused on searching for unexpected and interpretable patterns (in a somewhat more general sense than association rules) on data sets which are large enough that they are not amenable to visualization but are not necessarily in the massive category as described earlier (e.g., high-dimensional categorical data sets which fit in main memory). Typically these methods use various measures (such as entropy-based measures) to quantify how

informative a particular rule or pattern may be relative to background knowledge (Silberschatz and Tuzhilin, 1996), where background knowledge is usually expressed in a very simple form such as prior probability distributions on individual variables.

An early example of such work was the RX discovery project of Blum (Blum, 1982; Blum and Walker, 1986). The RX system searched through a subset of a data set of patient records to find candidate hypotheses such as "A precedes B in time, and A is correlated with B." The most interesting such hypotheses were then tested on the entire data set to evaluate their statistical significance. As with association rules, this type of unconstrained search raises some important concerns from a statistical viewpoint. The issue of multiple-testing is a real concern when searching through a large set of potential hypotheses in an automated fashion since there is a non-zero probability that some non-existent association will appear significant just by chance. The probability of incorrectly accepting such a spurious hypothesis rises as more and more hypotheses are tested. In addition, in many of these systems there appears to be an implied notion of causality, for example in the way the correlation information is presented as rules such as "if A then B." For naive users of such systems the difference between correlation and causation may not be apparent and the potential for misuse and misinterpretation is significant. Other more subtle dangers, such as the presence of hidden variables and Simpson's paradox, are discussed in Glymour et al. (1996, 1997).

Despite the potential pitfalls of unfettered automated "discovery" algorithms, the general idea of having a computer search a large database for unexpected patterns is certainly worthwhile as long as there is some human input into the process, and can legitimately be viewed as a form of large-scale semi-automated exploratory data analysis. A variety of machine learning and data mining algorithms have expanded on this general idea (Quinlan, 1987; Smyth and Goodman, 1992; Segal and Etzioni, 1994; Cohen, 1995; Domingos, 1996). The "patient rule induction method" (PRIM) of Friedman and Fisher (1999) provides a general statistical framework for rule induction, with broad applicability to multivariate data with mixed categorical and continuous-valued variables (many rule induction algorithms can only deal with categorical variables). The algorithm can be thought of as finding local "boxes" (hyper-rectangles) in a multi-dimensional space where some objective function of interest is maximized. As an example, one might have 20 demographic variables measured on medical patients with an additional binary class label indicating presence or absence of some medical condition. The objective function in this case could be defined as the log-odds that an individual has the condition, given that they lie within a particular box. The algorithm searches for "boxes" by shrinking the box boundaries in specific dimensions so as to greedily maximize or minimize the objective function within the box. Since the box boundaries are defined to be parallel to the variable axes, they can be interpreted as simple thresholds, and a box can be expressed in the form of a simple rule consisting of a conjunction of threshold conditions on variables on the left-hand side (e.g., "IF $X \geq 3.2$ and $0.2 \leq Y \leq 0.7$ and $Z \leq 1.8$") and a condition expressed in terms of the mean value of the objective function within the box on the right hand side (e.g., "THEN E[f] = 5.3"). The algorithm finds the first box, removes it from the multivariate space, then searches for the best box in the remaining space, and so on in an iterative manner, in an attempt to "cover" the input space. Fisher and Friedman (1999) illustrate interesting applications of the method to a multivariate geological data set and a consumer marketing data set. Note that this particular algorithm is not intended to be scalable in the sense described earlier, i.e., as described it is primarily intended for application to data sets which reside in main memory.

This general theme of discovering local patterns (rather than global models) has emerged in

several different strands of work in data mining. For example, Mannila, Toivonen, and Inkeri Verkamo (1995) describe formal methods for representing sequential patterns in sequences of events and then develop search algorithms that are similar in spirit to association rule algorithms for efficiently finding such patterns which occur frequently in large event sequence data sets. A typical pattern might be that the event $A\ OR\ B$ always precedes (within some time window) event C with probability p, where A, B and C are individual event types. The authors report that the method was applied to finding patterns in log-files of telecommunication alarms and the resulting patterns were considered useful by domain experts.

Bay and Pazzani (1999) describe *contrast sets*, a framework for determining statistically significant differences between two or more groups in a d-dimensional multivariate catagorical data sets. A contrast set consists of a conjunction of k variables and values, $1 \leq k \leq d$, which are statistically different across the groups. For example, applying this technique to a UC census data set, and comparing individuals with PhD degrees versus individuals with Bachelor's degrees, the algorithm discovered several interesting differences. Individuals in the PhD group were about 3 times more likely to work over 60 hours per week than the Bachelor's group, but were twice as likely to earn a salary greater than \$50,000 per annum. An interesting aspect of this problem is the huge search space involved, i.e., there are an exponential number of possible contrast sets. Bay and Pazzani describe a variety of heuristic search techniques for sysematically searching the space of hypotheses (candidate contrast sets). The multiple hypothesis testing problem is addressed by using modified Bonferroni corrections for significance testing.

These are but a few of a large class of data mining techniques for discovering local patterns from data sets. These patterns typically only "cover" a portion of the input space (e.g., PRIM finds local boxes within the full multi-dimensional space). It is not entirely clear how to evaluate these methods in the standard statistical framework, since there is usually no direct notion of how one can generate predictions from these patterns. Instead, the methods appear to have more in common with exploratory data analysis techniques, and could potentially be very useful in uncovering previously unknown relationships and structure in data. Having said this, it is frequently the case in practice that very large numbers of patterns are produced by these algorithms of which only a small subset of these are actually of interest or useful. The classic example is an algorithm which "discovers" from a database of medical patient records that all patients who are pregnant are also female. This is a valid relationship, but is trivial in the sense that it is already well-known. There has been some work in data mining on how patterns can be evaluated relative to a set of prior beliefs (e.g., see Padmanabhan and Tuzhilin (1998)) but it is fair to say that this is an area where there is much room for improvement.

6 Text and Multi-Media Mining

A significant recent focus of data-mining activity has involved the application of data mining concepts to online collections of text documents and multi-media objects such as images, video, audio, and so forth. Of interest here from a statistical viewpoint is the application of statistical concepts (such as population variability) to objects which are not necessarily best characterized by fixed-dimensional vector representations. The World-Wide-Web (WWW) is of course a major focus of attention in its own right. Phrases such as "text-mining" and "Web-mining" are often used for

these activities, although in many cases the techniques used appear to be fairly direct extensions and combinations of earlier ideas in the field of information retrieval (e.g., Van Rijsbergen, 1979) and applied statistics.

In the field of information retrieval the research focus has traditionally been on the following problem: find the documents (from a large corpus) which are most relevant to a specific query posed by a human to the system. The phrase "document" is interpreted broadly and can range from single paragraphs, to Web pages, to entire books. A widely used technique for solving this retrieval problem is to represent all queries and documents as individual *term-vectors*. A term can be a single word or phrase, and a term-vector is a d-dimensional vector of such terms, where component i is 1 if term i is present in the query or document and 0 otherwise (there are more general ways to do this, but this is the essential idea). Thus, in effect, documents and queries are reduced to points in a d-dimensional space and any of a variety of distance functions can be used to determine the similarity of queries and documents. Of course this simple term-vector representation loses a considerable amount of information compared to the original document, e.g., the relative position and context of individual terms are lost in the conversion to vector form. Nonetheless, this relatively simple vector-space approach works reasonably (and surprisingly) well and has been widely adopted in current information retrieval research (Witten, Moffat, and Bell (1999)).

Text data mining (insofar as it can be defined at this point, see Hearst (1999)) differs from information retrieval in that it can be viewed as the process of automated or semi-automated discovery of knowledge from text. As an example, unsupervised clustering algorithms can be used on collections of documents (perhaps represented as term-vectors) to discover which sets of documents are most closely related (at least in a term-vector sense). More specifically, hierarchical clustering algorithms can be used to automatically produce a taxonomy of documents. This can provide a practical alternative to manual cataloging of large document collections (e.g., in Web applications) since a human can simply assign labels to the clusters determined by the algorithm without having to pre-define what the taxonomy should be (e.g., Cutting et al. (1992)). This type of semi-supervised (or semi-automated) discovery appears to be quite a useful framework in general for the way in which humans actually perform data mining, i.e., rather than a fully-automated system which autonomously discovers patterns of interest, having a semi-automated process involving the human in interpretation and evaluation of patterns discovered by the algorithm. An early application of the idea of text mining is the work of Swanson (1987) who developed a system for automatically discovering links between previously disconnected strands of research in the medical literature. The system uses chains of implication within the medical literature to automatically search for hypotheses for causes of rare diseases (Swanson and Smalheiser (1994, 1997)).

In the context of Web-based data analysis, Chakrabarti et al. (1999) describe a general class of algorithms which treat the Web as a large graph, with links from one page to another represented as directed edges in this graph. Their approach estimates "authority" and "hub" weights for each of a set S of candidate Web pages which have already been retrieved on the basis of being potentially relevant to a given search term (a query). A page has a high authority weight if many good hubs point to it, and a page has a high hub weight if it points to many good authority pages. An authority page is intended to capture the notion of a page which is authoritative on a given subject, while a hub page is intended to be a page which contains collections of links to authorities. The definition of authority and hub weights specify a system of recursive equations since authority and hub weights depend on each other. The two sets of solution weights are in fact the principal

eigenvectors of AA^T and A^TA, where A is the adjacency matrix of the graph defined by S. The set of documents S are then ranked using the resulting weights and returned to the user. In related work, Kumar et al. (1998) demonstrate how the concepts of hubs and authorities can be used to automatically discover "cybercommunities," i.e., groups of Web page authors who share common interests.

The Web appears particularly well-suited to exploratory data mining ventures since it is relatively poorly understood and quite complex. For example, understanding navigation patterns of Web users is of considerable interest in e-commerce and network traffic contexts. Cadez et al. (2000) describe the application of mixtures of Markov models to modeling the sequences of page requests from users visiting a large commercial Web site over a 24-hour period. Each Web page at the site is categorized into one of approximately 18 categories. Thus, each user is represented by a discrete-valued sequence of page requests, where the sequences are of different lengths for different users. Approximately 1 million users were then clustered into 50 groups using the EM algorithm, each group represented by a Markov model, describing 50 qualitatively different sets of navigation patterns.

The ubiquity and availability of Web-related data is likely to lead to increasing interest from a data mining viewpoint in Web-related data sets. This type of data analysis poses a number of interesting challenges to traditional statistical methods since the data sets tend to be heterogenous and multi-modal (e.g., text, images, etc.), highly structured (the connectivity of Web pages), non-stationary (Web pages and their usage changes continually) and massive. Much of this work has relevance to medical research since medical data can also be viewed as highly structured, multi-modal, and non-stationary in nature, i.e., test results, time-series, diagnostic images, text annotations, and so forth. However, to date there has been relatively little data mining work directed at these types of data sets in medical contexts, other than standard applications of regression and classification algorithms for predictive modeling.

7 Conclusions

In this review we have discussed aspects of data mining which are somewhat distinct from traditional statistical research, with a particular emphasis on

- scalable data analysis algorithms that can operate efficiently on data which reside outside of main memory,
- algorithms which search for local patterns in data (rather than global models), and
- algorithms and techniques for non-traditional data sources such as document and image collections and the Web.

The utility of any of these techniques in a medical research context is not yet clear. Nonetheless, as medical data sets become larger, more heterogeneous, and contain more complex structure, at least some of these concepts from data mining may play a useful role in medical data analysis tasks. For example, pattern-finding algorithms such as PRIM could be quite useful for retrospective exploratory analysis of clinical trials data, keeping in mind the potential dangers of data-dredging mentioned earlier.

We noted throughout that data mining as currently practiced has its roots in computer science, rather than in statistics. More specifically, data mining has inherited many of the concepts and techniques underlying classification-oriented algorithms which were prevalent in machine learning during the 1980's and 1990's. It is also strongly influenced by research in the database area, where traditionally the emphasis has been on how to manage data rather than on how to interpret or analyze it. These "cultural biases" can be expected to become less pronounced as more statisticians and application-specific experts become involved in the data mining fray. However, it is nonetheless important that researchers who have traditionally relied on statistical methods in their work, be aware of the "computational" viewpoint which tends to prevail in data mining. Data mining offers several new and interesting techniques for data analysis on a grand scale, but it also requires a "marriage" with more fundamental statistical techniques in order to be successfully used in real-world applications. As more statisticians become involved in data mining we can expect to see more cross-fertilization of both statistical and computer science concepts occurring in data mining research.

Acknowledgements

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

References

Adriaans, P. and Zantige, D. (1996) *Data Mining*, Harlow, UK: Addison-Wesley.

Agrawal, R., Imielinski, T., and Swami, A. (1993) Mining associations between sets of items in massive databases, in *Proceedings of the 1993 ACM SIGMOD International Conference on the Management of Data*, New York, NY: ACM Press, 207–216.

Agrawal, R. and Stolorz, P. (eds.) (1998) *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, Menlo Park, CA: AAAI Press.

Aha, D., Kibler, D., and Albert, M. (1989) Instance-based learning, *Machine Learning*, 6, 37–66.

Armstrong, J. S., (1967) Derivation of theory by means of factor analysis or Tom Swift and his electric factor analysis machine, *American Statistician*, 21, 415–22.

Atkeson, C. G., Schaal, S., and Moore, A. W.(1997) Locally weighted learning, *Artificial Intelligence Review*, 11, 11–73.

Ballard, D. H. (1997) *An Introduction to Natural Computation*, Cambridge, MA: MIT Press.

Bay, S. and Pazzani, M. (1999) Detecting change in categorical data: mining contrast sets, in *Proceedings of the Fifth ACM International Conference on Knowledge Discovery and Data Mining*, New York, NY: ACM Press, 302–305.

Berry, M. J. A. and Linoff, G. (1997) *Data Mining Techniques For Marketing, Sales, and Customer Support*, New York, NY: John Wiley and Sons.

Bishop, C. (1995) *Neural Networks for Pattern Recognition*, Oxford, UK: Clarendon Press.

Blum, R. L. (1982) Discovery, confirmation, and incorporation of causal relationships from a large time-oriented clinical database: the RX project, *Computers and Biomedical Research*, 15, 165–187.

Bradley, P., Fayyad, U. M., and Reina, C. (1998) Scaling EM (expectation-maximization) to large databases, Technical Report MSR-TR-98-35, Microsoft Research, Redmond, WA.

Breiman, L. (1996) Stacked regressions, *Machine Learning*, 24(1), 49–64.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J., (1984) *Classification and Regression Trees*, Belmont, CA: Wadsworth Statistical Press.

Brin, S., Motwani, R., Ullman, J., and Tsur, S. (1997) Dynamic itemset counting and implication rules for market basket data, in *Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data*, New York, NY: 255–264.

Brin, S., Rastogi, R., and Shim, K. (1999) Mining optimized gain rules for numeric attributes, *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY: ACM Press, 135–144.

Cadez, I., Heckerman, D., Meek, C., Smyth, P., and White, S. (2000) Visualization of navigation patterns on a Web site using model-based clustering, in *Proceedings of the Sixth ACM International Conference on Knowledge Discovery and Data Mining*, New York, NY: ACM Press, to appear.

Chakrabarti, S., Dom, B. E., Ravi Kumar, S., Raghavan, P., Rajagopalan, S., Tomkins, A., Gibson, D., Kleinberg, J. (1999) Mining the Web's link structure, *IEEE Computer*, 32, 8, 60–67.

Chaudhuri, S. and Madigan, D. (eds.) (1999) *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY: ACM Press.

, Chen, P. (1976) The entity-relationship model—toward a unified view of data, *Transactions on Database Systems*, 1(1), 9–36.

Codd, E. (1970) A relational model for large shared data banks, *Communications of the ACM*, 13(6), 377–387.

Cohen, W. (1995) Fast effective rule induction, *Proceedings of the Twelfth International Conference on Machine Learning*, San Mateo: CA, Morgan Kaufmann, 115–123.

Cortes, C. and Pregibon, D. (1998) Giga-mining, in *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, R. Agrawal and P. Stolorz (eds.), Menlo Park, CA: AAAI Press, 174–178.

Cutting, D. R., Pedersen, J. O., Karger, D., and Tukey, J. W. (1992) Scatter/Gather: a cluster-based approach to browsing large document collections, in *Proceedings of the 15th Annual International ACM/SIGIR Conference*, New York, NY: ACM Press, 318–329.

Domingos, P. (1996), Unifying instance-based and rule-based induction, *Machine Learning*, 24, 141–168.

Domingos, P. and Pazzani, M. (1997) On the optimality of the zero-one classifier under zero-one loss, *Machine Learning*, 29, 103–130.

Du Mouchel, W., Volinsky, C., Johnson, T., Cortes, C., and Pregibon, D. (1999) Squashing flat files flatter, in *Proceedings of the Fifth ACM International Conference on Knowledge Discovery and Data Mining*, New York, NY: ACM Press, 6–15.

Duda, R. O. and Hart, P. E. (1973) *Pattern Classification and Scene Analysis*, New York, NY: Wiley.

Fayyad U. M., and Uthurasamy, R. (eds.) (1995) *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, Menlo Park, CA: AAAI Press.

Fayyad U. M., Piatetsky-Shapiro G., Smyth P., and Uthurasamy, R. (eds.) (1996) *Advances in Knowledge Discovery and Data Mining*, Cambridge, MA: MIT Press.

Fayyad U.M., Piatetsky-Shapiro G., and Smyth P. (1996) From data mining to knowledge discovery: an overview, in *Advances in Knowledge Discovery and Data Mining*, U. M.Fayyad, G.Piatetsky-Shapiro, P.Smyth, and R.Uthurusamy (eds.), Cambridge, MA: MIT Press, 1–34.

Friedman, J. (1997) On bias, variance, 0/1-Loss, and the curse-of-dimensionality, *Journal of Data Mining and Knowledge Discovery*, 1(1), 55–77.

Gehrke, J., Ganti, V., Ramakrishnan, R., Loh, W-Y. (1999) BOAT—Optimistic decision tree construction. *Proceedings of the SIGMOD Conference 1999*, New York, NY: ACM Press, 169–180.

Geman, S., Bienenstock, E., and Doursat, R. (1992) Neural networks and the bias/variance dilemma, *Neural Computation*, 4, 1–58.

Glymour C., Madigan D., Pregibon D., and Smyth P. (1996) Statistical inference and data mining, *Communications of the ACM*, 39(11), 35–41.

Glymour C., Madigan D., Pregibon D., and Smyth P. (1997) Statistical themes and lessons for data mining, *Journal of Data Mining and Knowledge Discovery*, 1, 11–28.

Gray, J. and Shenoy, P. (2000) Rules of thumb in data engineering, *Proceedings of the Sixteenth International Conference on Data Engineering*, Los Alamitos, CA: IEEE Computer Society.

Hand, D. J. (1998) Data mining—statistics and more, *The American Statistician*

Hearst, M. A. (1999) Untangling text data mining, in *Proceedings of ACL '99: the 37th Annual Meeting of the Association for Computational Linguistics*, New Brunswick, NJ: The Association for Computational Linguistics.

Heckerman, D., Mannila, H., and Pregibon, D. (eds.) (1997) *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, Menlo Park, CA: AAAI Press.

Hendry, D. F. (1995) *Dynamic Econometrics*, New York, NY: Oxford University Press.

Huber, P. (1997) From large to huge: a statisticians reactions to KDD and DM, in *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, Menlo Park, CA: AAAI Press, 304–308.

Inmon, W. (1996) *Building the Data Warehouse*, New York, NY: Wiley, 2nd edition.

Kumar, S. R., et al. (1999) Trawling emerging cyber-communities automatically, *Proceedings of the 8th World Wide Web Conference*, Amsterdam: Elsevier Science, 403–415.

Leamer, E. E. (1978) *Specification Searches: Ad Hoc Inference with Non-Experimental Data*, New York, NY: John Wiley.

Lovell, M. (1983) Data mining, *Review of Economics and Statistics*, 65, 1–12.

Mannila, H., Toivonen, H. and Inkeri Verkamo, A. (1995) Discovering frequent episodes in sequences, in *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, Menlo Park, CA: AAAI Press, 210–215.

Michie, D., Spiegelhalter, D. J., and Taylor, C. C. (eds.) (1994) *Machine Learning, Neural and Statistical Classification*, New York, NY: Ellis Horwood.

Mitchell, T. M. (1997) *Machine Learning*, New York, NY: McGraw Hill.

Moore, A. W., (1999) Cached sufficient statistics for automated discovery and data mining from massive data sources, online white paper, Department of Computer Science, Carnegie Mellon University, July 1999.

Moore, A. W. and Lee, M. (1998) Cached sufficient statistics for efficient machine learning with large data sets, *Journal of Artificial Intelligence Research*, 8, 67–91.

Padmanabhan, B. and Tuzhilin, A. (1998) A belief-driven discovery method for discovering unexpected patterns, In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, Menlo Park, CA: AAAI Press, 94–100.

Pavlov, D., Mannila, H., and Smyth, P. (1999) Prediction with local patterns using cross-entropy, in *Proceedings of the Fifth ACM International Conference on Knowledge Discovery and Data Mining*, New York, NY: ACM Press, 357–361.

Piatetsky-Shapiro, G. (1991) Report on the AAAI-91 Workshop on knowledge discovery in databases and knowledgebases, *IEEE Expert*, 6(5), 74–76.

Piatetsky-Shapiro, G. and Frawley, W. (eds.) (1991) *Knowledge Discovery in Databases*, Menlo Park, CA: AAAI Press.

Provost, F. and Kolluri, V. (1999) A survey of methods for scaling up inductive algorithms, *Journal of Data Mining and Knowledge Discovery*, 3(2), 131–169.

Quinlan, J. R. (1987) Generating production rules from decision trees, *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*, San Mateo, CA: Morgan Kaufmann, 304–307.

Quinlan, J. R. (1993) *C4.5: Programs for Machine Learning*, San Mateo: CA, Morgan Kaufmann.

Ramakrishnan, R., and Stolfo, S. (eds.) (2000) *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY: ACM Press.

Ripley, B. D. (1996) *Pattern Recognition and Neural Networks*, Cambridge, UK: Cambridge University Press.

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

Scholkopf, C., Burges, J. C., and Smola, A. J. (1999) *Advances in Kernel Methods*, Cambridge, MA: MIT Press.

Segal, R. and Etzioni, O. (1994) Learning decision lists using homogenous rules, *Proceedings of the Twelfth National Conference on Artificial Intelligence*, Menlo Park, CA: AAAI Press, 619–625.

Selvin, H. and Stuart, A. (1966) Data dredging procedures in survey analysis, *American Statistician*, 20(3), 20–23.

Silberschatz, A. and Tuzhilin, A. (1996) What makes patterns interesting in database systems, *IEEE Transactions on Knowledge and Data Engineering*, 8(6), 970–974.

Simoudis, E., and Han, J. (eds.) (1996) *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, Menlo Park, CA: AAAI Press.

Smyth, P. and Goodman, R. (1992) An information-theoretic approach to rule induction from databases, *IEEE Transactions on Knowledge and Data Engineering*, 4(4), 301–306.

Swanson, D. R. (1987) Two medical literatures that are logically but not bibliographically connected, *Journal of the American Society for Information Retrieval*, 38(4), 228–233.

Swanson, D. R. and Smalheiser, N. R. (1994) Assessing a gap in the biomedical literature: magnesium deficiency and neurologic disease, *Neuroscience Research Communications*, 15, 1–9.

Swanson, D. R. and Smalheiser, N. R. (1997) An interactive system for finding complementary literatures: a stimulus to scientific discovery, *Artificial Intelligence*, 91, 183–203.

Valdes-Perez, R. E., (1999) Principles of human computer collaboration for knowledge discovery in science, *Artificial Intelligence*, 107(2), 335–346.

Van Rijsbergen, C. J. (1979) *Information Retrieval*, London, Butterworh Press, 2nd edition.

Walker, M. G. and Blum, R. L. (1986) Towards automated discovery from clinical databases: the RADIX project, in *Proceedings of the Fifth Conference on Medical Informatics*, volume 5, 32–36.

Witten, I. H., Moffat, A., and Bell, T. C. (1999) *Managing Gigabytes: Compressing and Indexing Documents and Images*, San Francisco, CA: Morgan Kaufmann, Second Edition.

Witten, I. H. and Frank, E. (1999) *Data Mining: Practical Machine Learning Tools and Techniques with JAVA Implementations*, San Francisco, CA: Morgan Kaufmann.

Weiss, S. M. and Indurkhya, N. (1998) *Predictive Data Mining: A Practical Guide*, San Francisco, CA: Morgan Kaufmann Publishers.

Zhang, T., Ramakrishnan, R., Livny, M., (1997) BIRCH: A new data clustering algorithm and its applications, *Journal of Data Mining and Knowledge Discovery*, 1(2), 141–182.