Machine Learning and Data Mining

Machine learning algorithms enable discovery of important "regularities" in large data sets.

Tom M. Mitchell

Over the past decade, many organizations have begun to routinely capture huge volumes of historical data describing their operations, products, and customers. At the same time, scientists and engineers in many fields have been capturing increasingly complex experimental data sets, such as gigabytes of functional magnetic resonance imaging (MRI) data describing brain activity in humans. The field of data mining addresses the question of how best to use this historical data to discover general regularities and improve the process of making decisions.

PHILIP NORTHOVER/PNORTHOV.FUTURE.EASYSPACE.COM/INDEX.HTML

The increasing interest in data mining, or the use of historical data to discover regularities and improve future decisions, follows from the confluence of several recent trends: the falling cost of large data storage devices and the increasing ease of collecting data over networks; the development of robust and efficient machine learning algorithms to process this data; and the falling cost of computational power, enabling use of computationally intensive methods for data analysis.

The field of data mining, sometimes called "knowledge discovery from databases," "advanced data analysis," and machine learning, has already produced practical applications in such areas as analyzing medical outcomes, detecting credit card fraud, predicting customer purchase behavior, predicting the personal interests of Web users, and optimizing manufacturing processes. It has also led to a set of fascinating scientific questions about how computers might automatically learn from past experience.

Figure 1. Data mining application. A historical set of 9,714 medical records describes pregnant women over time. The top portion is a typical patient record ("?" indicates the feature value is unknown). The task for the algorithm is to discover rules that predict which future patients will be at high risk of requiring an emergency C-section delivery. The bottom portion shows one of many rules discovered from this data. Whereas 7% of all pregnant women in the data set received emergency C-sections, the rule identifies a subclass at 60% at risk for needing C-sections.

Data

Patient103 time=1 →	*Patient103 time=2* → ...	*Patient103 time=n*
Age: 23	Age: 23	Age: 23
FirstPregnancy: No	FirstPregnancy: No	FirstPregnancy: No
Anemia: No	Anemia: No	Anemia: No
Diabetes: No	Diabetes: Yes	Diabetes: No
PreviousPrematureBirth: No	PreviousPrematureBirth: No	PreviousPrematureBirth: No
Ultrasound: ?	Ultrasound: Abnormal	Ultrasound: ?
Elective C–Section: ?	Elective C–Section: ?	Elective C–Section: No
Emergency C–Section: ?	Emergency C–Section: ?	**Emergency C–Section: Yes**
...	...	...

Rules learned

If No previous vaginal delivery, and
Abnormal 2nd Trimester Ultrasound, and
Malpresentation at admission
Then Probability of Emergency C-Section is 0.6

Training set accuracy: 26/41 = .63
Test set accuracy: 12/20 = .60

Prototypical Applications

Figure 1 shows a prototypical example of a data mining problem. Given a set of historical data, we are asked to use it to improve our medical decision making. The data consists of a set of medical records describing 9,714 pregnant women. We want to improve our ability to identify future high-risk pregnancies—specifically, those at high risk of requiring an emergency Cesarean-section delivery. In this database, each pregnant woman is described in terms of 215 distinct features, such as her age, whether she is diabetic, and whether this is her first pregnancy. These features (in the top portion of the figure) together describe the evolution of each pregnancy over time.

The bottom portion of the figure illustrates a typical result of data mining, including one of the rules learned automatically from this data set. This particular rule predicts a 60% risk of emergency C-section for mothers exhibiting a particular combination of three features, out of the 215 possible features. Among the women known to exhibit these three features, the data indicates that 60% have historically given birth by emergency C-section. As summarized at the bottom of the figure, this regularity holds over both the training data used to formulate the rule and a separate set of test data used to verify the reliability of the rule over new data. Physicians may want to consider this rule as a useful factual statement about past patients when weighing treatment for similar new patients.

What algorithms can be used to learn rules like the one in the figure? This rule was learned through a symbolic rule-learning algorithm similar to Clark's and Nisbett's CN2 [3]. Decision-tree learning algorithms, such as Quinlan's C4.5 [9], are also frequently used to formulate rules of this type. When rules have to be learned from extremely large data sets, specialized algorithms stressing computational efficiency may also be used [1, 4]. Other machine learning algorithms commonly applied to this kind of data mining problem include neural networks [2], inductive logic programming [8], and Bayesian learning algorithms [5]. Mitchell's 1997 textbook [7] describes a broad range of machine learning algorithms used for data mining, as well as the statistical principles on which they are based.

Figure 2. Typical data and rules for analyzing credit risk.

Data

Customer 103: (time=t0)	*Customer 103: (time=t1)* ...	*Customer 103: (time=tn)*
Years of credit: 9	Years of credit: 9	Years of credit: 9
Loan balance: $2,400	Loan balance: $3,200	Loan balance: $4,500
Income: $52k	Income: ?	Income: ?
Own House: Yes	Own House: Yes	Own House: Yes
Other delinquent accts: 2	Other delinquent accts: 2	Other delinquent accts: 3
Max billing cycles late: 3	Max billing cycles late: 4	Max billing cycles late: 6
Repay loan?: ?	Repay loan?: ?	**Repay Loan?: No**
...	...	...

Rules learned from synthesized data:

If Other-Delinquent-Accounts > 2, and
Number-Delinquent-Billing-Cycles > 1
Then Repay-Loan? = No

If Other-Delinquent-Accounts = 0, and
(Income > $30k) OR (Years-of-Credit > 3)
Then Repay-Loan? = Yes

Although machine learning algorithms are central to the data mining process, it is important to note that the process also involves other important steps, including building and maintaining the database, data formatting and cleansing, data visualization and summarization, the use of human expert knowledge to formulate the inputs to the learning algorithm and evaluate the empirical regularities it discovers, and determining how to deploy the results. Thus, data mining bridges many technical areas, including databases, human-computer interaction, statistical analysis, and machine learning algorithms. My focus here is on the role of machine learning algorithms in the data mining process.

The patient-medical-records application example in Figure 1 represents a prototypical data mining problem in which the data consists of a collection of time-series descriptions; we use the data to learn to predict later events in the series—emergency C-sections—based on earlier events—symptoms before delivery. Although I use a medical example to illustrate these ideas, I could have given an analogous example of learning to predict, say, which bank-loan applicants are at high risk of failing to repay their loans (see Figure 2). As shown in this figure, data in such applications typically consists of time-series descriptions of customer bank balances and other demographic information, rather than medical symptoms.

Other data mining applications include predicting customer purchase behavior, customer retention, and the quality of goods produced by a particular manufacturing line (see Figure 3). All are applications for which data mining has been applied successfully and in which further research promises even more effective techniques.

The State of the Art and Beyond

The field of data mining is at an interesting crossroads; we now have a first generation of machine learning algorithms (such as those for learning decision trees, rules, neural networks, Bayesian networks, and logistic regressions) that have been demonstrated to be of significant value in a variety of real-world data mining applications. Dozens of companies around the world now provide commercial implementations of these algorithms (see www.kdnuggets.com), along with efficient interfaces to commercial databases and well-designed user interfaces. But these first-generation algorithms also have significant limitations. They typically assume the data contains only numeric and symbolic features and no text, image features, or raw sensor data. They assume the data has been carefully collected into a single database with a specific data mining task in mind. Furthermore, today's algorithms tend to be fully automatic and therefore fail to allow guidance from knowledgeable users at key stages in the search for data regularities.

Given these limitations, and the strong commercial interest despite them, and the accelerating university research in machine learning and data mining, we might well expect the next decade to produce an order of magnitude advance in the state of the art. Such an advance could be motivated by development of new algorithms that accommodate dramatically more diverse sources and types of data, a broader range of automated steps in the data mining process, and mixed-initiative data mining in which human experts collaborate more closely with the computer to form hypotheses and test them against the data.

To illustrate one important research issue, consider again the problem of predicting the risk of an emergency C-section for pregnant women. One key limitation of current data mining methods is that they cannot utilize the full patient record that is today routinely captured in hospital medical records. This is because hospital records for pregnant women often contain sequences of images (such as the ultrasound images taken during pregnancy), other raw

Figure 3. More data mining problems.

Customer purchase behavior

Customer 103: (time=t0)	*Customer 103: (time=t1)*	...	*Customer 103: (time=tn)*
Sex: M	Sex: M		Sex: M
Age: 53	Age: 53		Age: 53
Income: $50,000	Income: $50,000		Income: $50,000
Own House: Yes	Own House: Yes		Own House: Yes
MS Products: Word	MS Products: Word		MS Products: Word
Computer: 386 PC	Computer: Pentium		Computer: Pentium
Purchase Excel?: ?	Purchase Excel?: ?		**Purchase Excel?: Yes**
...	...		...

Customer retention

Customer 103: (time=t0)	*Customer 103: (time=t1)*	...	*Customer 103: (time=tn)*
Sex: M	Sex: M		Sex: M
Age: 53	Age: 53		Age: 53
Income: $50,000	Income: $50,000		Income: $50,000
Own House: Yes	Own House: Yes		Own House: Yes
Checking: $5,000	Checking: $20,000		Checking: $0
Savings: $15,000	Savings: $0		Savings: $0
Current customer?: Yes	Current customer?: Yes		**Current customer?: No**

Process optimization

Product72: (time=t0)	*Product72: (time=t1)*	...	*Product72: (time=tn)*
Stage: mix	Stage: cook		Stage: cool
Mixing–speed: 60rpm	Temperature: 325°		Fan–speed: medium
Viscosity: 1.3	Viscosity: 3.2		Viscosity: 1.3
Fat content: 15%	Fat content: 12%		Fat content: 12%
Density: 2.8	Density: 1.1		Density: 1.2
Spectral peak: 2800	Spectral peak: 3200		Spectral peak: 3100
Product underweight?: ?	Product underweight?: ?		**Product underweight?: Yes**
...	...		...

instrument data (such as fetal distress monitors), text (such as the notes made by physicians during periodic checkups during pregnancy), and even speech (such as the recording of phone calls), in addition to the numeric and symbolic features in Figure 1.

Although our first-generation data mining algorithms work well with the numeric and symbolic features, and although some learning algorithms are available for learning to classify images or to classify text, we still lack effective algorithms for learning from data that is represented by a *combination* of these various media. As a result, the state of the art in medical outcomes analysis is to ignore the image, text, and raw sensor portions of the medical record, or at best to summarizes them in an oversimplified form (such as by labeling complex ultrasound images as simply "normal" or "abnormal").

However, it is natural to expect that if predictions could be based on the full medical record, we should achieve much greater prediction accuracy. Therefore, a topic of considerable research interest is development of algorithms that can learn regularities in rich, mixed-media data. This issue is important in many data mining applications, ranging from mining historical equipment maintenance records, to mining records at customer call centers, to analyzing MRI data on brain activity during different tasks.

The research issue of learning from mixed-media data is just one of many current research issues in data mining. The left-hand side of Figure 4 lists a number of additional research topics in data mining; the right-hand side indicates a variety of applications for which these research issues are important including:

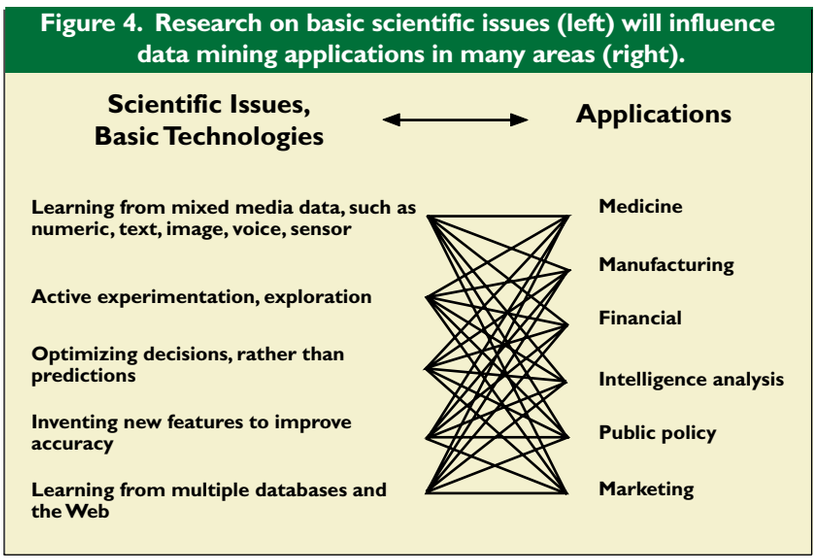
Optimizing decisions, rather than predictions. The research goal here is to develop machine learning algorithms that go beyond learning to predict likely outcomes, and learn to suggest preemptive actions that achieve the desired outcome. For example, consider again the birth data set mentioned earlier. Although it is clearly helpful to learn to predict which women suffer a high risk of birth complications, it would be even more useful to learn which preemptive actions might help reduce this risk. Similarly, in modeling bank customers, it is one thing to predict which of them might close their accounts and move to new banks; even more useful would be to learn which actions might help retain them before they left.

This problem of learning which actions achieve a desired outcome, given only previously acquired data, is much more subtle than it may first appear. The difficult issue is that the available data often represents a biased sample that does not correctly represent the underlying causes and effects; for example, whereas the data may show that mothers giving birth at home suffer fewer complications than those giving birth in a hospital, one cannot necessarily conclude that sending a woman home reduces her risk of complications.

The observed regularity might instead be due to the fact that a disproportionate number of high-risk women choose to give birth in a hospital. Thus, the problem of learning to choose actions raises impor-

A topic of considerable interest is development of algorithms that can learn regularities in rich, mixed media data.

Figure 4. Research on basic scientific issues (left) will influence data mining applications in many areas (right).



tant and basic questions, such as: How can the system learn from biased samples of data? How can the system incorporate conjectures by human experts about the effectiveness of various intervention actions? If successful, this research will allow the application of historical data much more directly to the questions of greatest concern to decision makers.

Scaling to extremely large data sets. Whereas most learning algorithms perform acceptably on data sets with tens of thousands of training examples, many important data sets are significantly larger. For example, large retail customer databases and Hubble telescope data can easily involve a terabyte or more. To provide reasonably efficient data mining methods for such large data sets requires additional research. Research during the past few years has already produced more efficient algorithms for such problems as learning association rules [1] and efficient visualization of large data sets [6]. Further research in this direction might well lead to even closer integration of machine learning algorithms into database management systems.

Active experimentation. Most current data mining systems passively accept a predetermined data set. We need new computer methods that actively generate optimal experiments to obtain additional useful information. For example, when modeling a manufacturing process, it is relatively easy to capture data while the process runs under normal conditions. But this data may lack information about how the process performs under important non-standard unpredictable conditions. We need algorithms that propose optimal experiments to collect the most informative data, taking into account—precisely—the expected benefits, as well as the risks, of the experiment.

Learning from multiple databases and the Web. The volume and diversity of data available over the Internet and corporate intranets is large and growing rapidly. Therefore, future data mining methods should be able to use this huge variety of data sources to expand their access to data and to learn useful regularities. For example, one large U.S. equipment manufacturer uses data mining to construct models of the interests and maintenance needs of its corporate customers. In this application, the company mines a database consisting primarily of records of past pur-

chases and the servicing needs of its various customers, with only a few features describing the type of business each customer performs.

But as it turns out, nearly all of these customers have public Web sites that provide considerable information about their current and planned activities. Significant improvement could be expected in the manufacturer's data mining of strategic information if the data mining algorithms combined this Web-accessible information with the information available in the manufacturer's own internal database. To achieve this result, however, we need to develop new algorithms that can successfully extract information from Web hypertext. If successful, this line of research could yield several orders of magnitude increase in the variety and currency of data accessible to many data mining applications.

Inventing new features to improve prediction accuracy. In many cases, the accuracy of predictions can be improved by inventing a more appropriate set of features to describe the available data. For example, consider the problem of detecting the imminent failure of a piece of equipment based on the time series of sensor data collected from the equipment. Millions of features describing this time series can be generated easily by taking differences, sums, ratios, and averages of primitive sensor readings, along with previously defined features. Given a sufficiently large and long-duration data set, it should be feasible to automatically explore this large space of possible defined features to identify the small fraction of them most useful for future learning. This work could lead to increased accuracy in many prediction problems, such as equipment failure, customer attrition, credit repayment, and medical outcomes.

Active research takes many other directions, including how to provide more useful data visualization tools, how to support mixed-initiative human-machine exploration of large data sets, and how to reduce the effort needed for data warehousing and for combining information from different legacy databases. Still, the interesting fact is that even current first-generation approaches to data mining are being put to routine use by many organizations, producing important gains in many applications.

We might speculate that progress in data mining over the next decade will be driven by three mutually reinforcing trends:

- Development of new machine learning algorithms that learn more accurately, utilize data from dramatically more diverse data sources available over the Internet and intranets, and incorporate more human input as they work;
- Integration of these algorithms into standard database management systems; and
- An increasing awareness of data mining technology within many organizations and an attendant increase in efforts to capture, warehouse, and utilize historical data to support evidence-based decision making.

We can also expect more universities to react to the severe shortage of trained experts in this area by creating new academic programs for students wanting to specialize in data mining. Among the universities recently announcing graduate degree programs in data mining, machine learning, and computational statistics are Carnegie-Mellon University (see www.cs.cmu.edu/~cald), the University of California, Irvine (www.ics.uci.edu/~gcounsel/masterreqs.html), George Mason University (vanish.science.gmu.edu), and the University of Bristol (www.cs.bris.ac.uk/Teaching/Machine-Learning). C

References

1. Agrawal, R., Imielinski, T., and A. Swami. Database mining: A performance perspective. *IEEE Trans. on Knowl. Data Eng. 5,* 6 (1993), 914–925.
2. Chauvin, Y. and Rumelhart, D. *Backpropagation: Theory, Architectures, and Applications.* Lawrence Erlbaum Associates, Hillsdale, N.J., 1995.
3. Clark, P. and Niblett, R. The CN2 induction algorithm. *Mach. Learn. 3,* 4 (Mar. 1989), 261–284.
4. Gray, J., Bosworth A., Layman A., and Pirahesh, H. *Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Totals.* Microsoft Tech. Rep. MSR-TR-95-22, Redmond, Wash., 1995.
5. Heckerman, D., Geiger, D., and Chickering, D. Learning Bayesian networks: The combination of knowledge and statistical data. *Mach. Learn. 20*, 3 (Sept. 1995), 197–243.
6. Faloutsos C. and Lin, K. FastMap: A fast algorithm for indexing, data mining, and visualization. *ACM SIGMOD* (1995), 163–174.
7. Mitchell, T. *Machine Learning.* McGraw-Hill, New York, 1997.
8. Muggleton, S. *Foundations of Inductive Logic Programming.* Prentice Hall, Englewood Cliffs, N.J., 1995.
9. Quinlan J. *C4.5 Programs for Machine Learning.* Morgan Kaufmann, San Mateo, Calif., 1993.

Tom M. Mitchell (Tom.Mitchell@cmu.edu) is the Fredkin Professor of AI and Learning and director of the Center for Automated Learning and Discovery in the School of Computer Science at Carnegie-Mellon University in Pittsburgh.

This research is supported in part by DARPA under contract F30602-97-1-0215 and by contributions from the Corporate Members of the Center for Automated Learning and Discovery in the School of Computer Science at Carnegie-Mellon University.