To appear in *Communications of the ACM*,
Vol. 42, No. 11, November 1999.

Machine Learning and Data Mining

Tom M. Mitchell
Center for Automated Learning and Discovery
School of Computer Science
Carnegie Mellon University

1 Introduction

Over the past decade many organizations have begun to routinely capture huge volumes of historical data describing their operations, their products, and their customers. At the same time, scientists and engineers in many fields find themselves capturing increasingly complex experimental datasets, such as the gigabytes of functional MRI data that describe 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 to improve future decisions.

> *Data Mining:* using historical data to discover regularities and improve future decisions.

The rapid growth of interest in data mining follows from the confluence of several recent trends: (1) the falling cost of large data storage devices and the increasing ease of collecting data over networks, (2) the development of robust and efficient machine learning algorithms to process this data, and (3) the falling cost of computational power, enabling the use of computationally intensive methods for data analysis.

The field of data mining, sometimes referred to as knowledge discovery from databases, machine learning, or advanced data analysis, has already produced highly practical applications in areas such as credit card fraud detection, medical outcomes analysis, predicting customer purchase behavior, predicting the 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 experience.

2 Data Mining Examples

A prototypical example of a data mining problem is illustrated in Figure 1. Here we are provided a set of historical data and asked to use this data to make improved decisions in the future. In this example the data consists of a set of medical records describing 9,714 pregnant women. The decision we wish to improve is our ability to identify future high risk pregnancies (more specifically, pregnancies that are 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 this is a first pregnancy, whether she is diabetic, and so on. As shown in the top portion of Figure 1, these features together describe the evolution of the pregnancy over time.

The bottom portion of Figure 1 illustrates a typical result of data mining. It shows one of the rules that has been learned automatically from this set of data. This particular rule predicts a 60 percent risk of emergency C-section for mothers that exhibit a particular combination of three features (e.g., "no previous vaginal delivery") out of the 215 possible features. Among women known to exhibit these three features, the data indicates that 60 percent have historically given birth by emergency C-section. As summarized at the bottom of the figure, this regularity holds both over the training data used to formulate the rule, and over a separate set of test data used to verify the reliability of the rule over new data. Physicians may wish to consider this rule as a useful factual statement about past patients when they consider treatment of similar new patients.

What algorithms are used to learn rules such as the one in Figure 1? This rule was learned by a symbolic rule learning algorithm similar to Clark 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 must be learned from extremely large data sets, specialized algorithms that stress computational efficiency may 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 textbook [7] provides a description of a broad range of machine learning algorithms, as well as the statistical principles on which they are based.

Although machine learning algorithms such as these are central to the

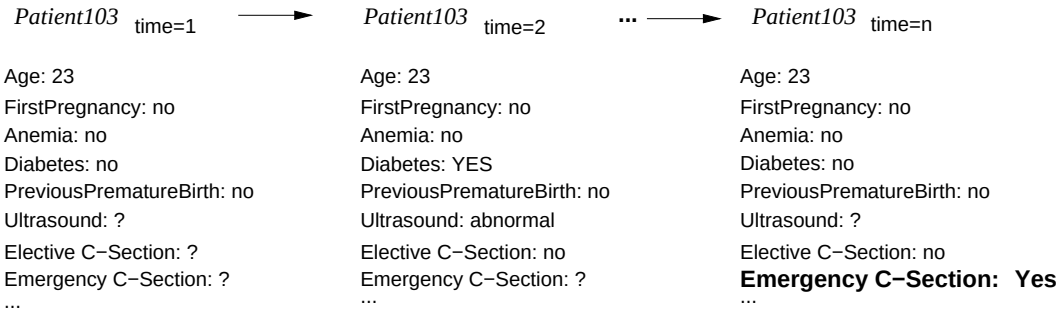
data mining process, it is important to note that the data mining process also includes other important steps such as building and maintaining the database, data formatting and data cleansing, data visualization and summarization, the use of human expert knowledge to formulate the inputs to the learning algorithm and to evaluate the empirical regularities it discovers, and the eventual deployment of the results. Thus data mining bridges many technical areas including data bases, human-computer interaction, statistical analysis and machine learning algorithms. In this article, we focus primarily on the role of machine learning algorithms in the data mining process.

The above medical example thus represents a prototypical data mining problem in which the data consists of a collection of time series descriptions, and we use the data to learn to predict later events in the series (emergency C-section) based on earlier events (symptoms before delivery). Although we used a medical example for illustration, we could have given an analogous example of learning to predict which bank loan applicants are at high risk of failing to repay their loan (see Figure 2). As shown in the figure, in this application data typically consists of time series descriptions of customers' bank balances and other demographic information, rather than medical symptoms. Yet other applications, shown in Figure 3, include predicting customer purchase behavior, customer retention, and the quality of goods produced on a particular manufacturing line. All of these are applications for which data mining has been successfully applied in practice, and where further research promises yet more effective techniques.

3 The State of the Art, and What Next?

What is the current state of the art in data mining? The field is at an interesting crossroads: we now have a first generation of data mining algorithms (e.g., logistic regression, decision tree and rule learning algorithms, neural network learning methods, and Bayesian learning methods) that have been demonstrated to be of significant value across a variety of real-world applications. Dozens of companies now provide commercial implementations of these algorithms (for a list, see www.kdnuggets.com), along with efficient interfaces to commercial databases and well-designed user interfaces. But these first generation data mining algorithms work best for problems where one has a large set of data collected into a single database, where the data

Data:



Learned rule:

```
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
```

Figure 1: A typical data mining application. A historical set of 9714 medical records describes pregnant women over time. The top portion of the figure illustrates a typical patient record, where "?" indicates that the feature value is unknown. The task here is to identify classes of patients at high risk of receiving an emergency Cesarean section. The bottom portion of the figure shows one of many rules discovered from this data. Whereas 7% of all pregnant women in this dataset received emergency C-sections, this rule identifies a subclass at 60% risk.

Data:

Customer103: (time=t0)	*Customer103:* (time=t1)	...	*Customer103:* (time=tn)
Years of credit: 9	Years of credit: 9		Years of credit: 9
Loan balance: $2,400	Loan balance: $3,250		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
```

Figure 2: Typical data and rules for credit risk analysis.

Customer purchase behavior:

Customer103: (time=t0)	*Customer103:* (time=t1)	...	*Customer103:* (time=tn)
Sex: M	Sex: M		Sex: M
Age: 53	Age: 53		Age: 53
Income: $50k	Income: $50k		Income: $50k
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:

Customer103: (time=t0)	*Customer103:* (time=t1)	...	*Customer103:* (time=tn)
Sex: M	Sex: M		Sex: M
Age: 53	Age: 53		Age: 53
Income: $50k	Income: $50k		Income: $50k
Own House: Yes	Own House: Yes		Own House: Yes
Checking: $5k	Checking: $20k		Checking: $0
Savings: $15k	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**
...	...		...

Figure 3: Additional examples of data mining problems.

is described by numeric or symbolic features, where the data does not contain text and image features interleaved with these numeric and symbolic features, and where the data has been carefully and cleanly collected with a particular decision making task in mind.

While this first generation of data mining algorithms is already of significant practical value, data mining methods are still in their infancy. We might well expect the next decade to produce an order of magnitude advance in the state of the art, through development of new algorithms that will accomodate dramatically more diverse sources and types of data, that will automate a broader range of the steps involved in the data mining process, and that will support mixed-initiative data mining in which human experts collaborate with the computer to form hypotheses and test them against the data.

To illustrate one important research issue, consider again the problem of predicting risk of emergency C-section for pregnant women. One key limitation of current data mining methods is that in fact they cannot utilize the full patient record that is already routinely captured in hospital medical records! This is because current hospital records for pregnant women often contain sequences of images (e.g., the ultrasound images taken during pregnancy), other raw instrument data (e.g., fetal distress monitors), text (e.g., the notes made by physicians during periodic checkups during pregnancy), and even speech (e.g., recordings of phone calls), in addition to the numeric and symbolic features described 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, the fact is that we currently lack effective algorithms for learning from data that is represented by a *combination* of these various media. As a result, the current state of the art in medical outcomes analysis is to ignore the image, text, and raw sensor portion of the medical record, or at best to summarize these in some oversimplified form (e.g., labeling the complex ultrasound image as simply "normal" or "abnormal"). However, it is clear that if predictions could be based on the full medical record, we would expect much greater prediction accuracy. Therefore, a topic of considerable current research interest is the development of algorithms that can learn regularities over 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 fMRI data on brain activity during different tasks.

This 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, while the right hand side of this figure indicates a variety of applications for which these research issues are important. Below we discuss these additional research issues in turn:

- *Optimizing decisions rather than predictions.* The goal here is to use historical data to improve the choice of actions in addition to the more usual goal of predicting outcomes. 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 pre-emptive actions could be taken to reduce this risk. Similarly, in modeling bank customers it is one thing to predict which customers may close their accounts and move to a new bank, but even more useful to learn which actions may be taken to retain the customer before they depart. This problem of learning which actions acheive 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; for instance, whereas the data may show that mothers giving birth at home suffer fewer complications than women who give birth in the hospital, one cannot necessarily conclude that sending a woman home will reduce her risk of complications. This empirical regularity might instead be due to the fact that a disproportionate number of high risk women choose to give birth in the hospital. Thus, the problem of learning to choose actions raises important and basic questions such as how to learn from biased samples of data, and how to incorporate conjectures by human experts about the effectiveness of various intervention actions. If successful, this research will allow applying historical data much more directly to the questions faced by decision-makers.

- *Scaling to extremely large data sets.* Whereas most learning algorithms perform acceptably on datasets with tens of thousands of training examples, data sets such as large retail customer data bases, and the Hubble telescope data can easily reach a terabyte or more. To provide reasonably efficient data mining methods for such large data sets requires additional research. Research during the past few has already

produced more efficient algorithms for problems such as learning association rules [1], and efficient visualization of large data sets [6]. Further research in this direction might lead to an even closer integration of machine learning algorithms into database 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, in modeling a manufacturing process it is relatively easy to capture data while the process runs under normal conditions. However, this data may lack information about how the process will perform under important non-standard conditions. We need algorithms that will 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 over multiple databases and the World Wide Web.* The volume and diversity of data available over the Internet and over corporate intranets is very large and growing rapidly. Therefore it is natural that future data mining methods will use this huge variety of data sources to expand their access to data and their ability to learn useful regularities. For example, one large equipment manufacturer currently uses data mining to construct models of the interests and maintenance needs of their corporate customers. In this application, they mine a database that consists primarily of records of past purchases and servicing needs of various customers, with only a few features that describe the type of business that each customer performs. As it turns out, nearly all of these corporate customers have public web sites that provide considerable information about their current and planned activities. If the data mining algorithms could combine this information with the information available in the internal database, one would expect significant improvements. Of course to achieve this, we will need new methods that can successfully extract information from web hypertext. If successful, this line of work may result in 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. It is easy to generate millions of features that describe this time series by taking differences, sums, ratios, averages, etc. of primitive sensor readings and previously defined features. Our conjecture is that given a sufficiently large and long-duration data set it should be feasible to automatically explore this large space of possible defined features in order to identify the small fraction of these features most useful for future learning. If successful, this work would lead to increased accuracy in many prediction problems, such as predicting equipment failure, customer attrition, credit repayment, medical outcomes, etc.

There are many other directions of active research as well, including work on 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 as the future of this field unfolds, we will see several directions in which it will advance including (1) new algorithms that learn more accurately, that are able to utilize data from dramatically more diverse data sources available over the internet and intranets, and that are able to incorporate more human input as they work (2) integration of these data mining algorithms into standard database systems, (3) an increasing effort within many organizations on capturing, warehousing and utilizing historical data to support evidence-based decision making.

We can also expect to see more universities react to the severe shortage of trained experts in this area, by creating new academic programs for students wishing to specialize in data mining. In fact, several universities have recently announced graduate degree programs in data mining, machine learning, and computational statisics, including Carnegie Mellon University (see www.cs.cmu.edu/~cald), University of California at Irvine (www.ics.uci.edu/~gcounsel/masterreqs.html), George Mason University (van-

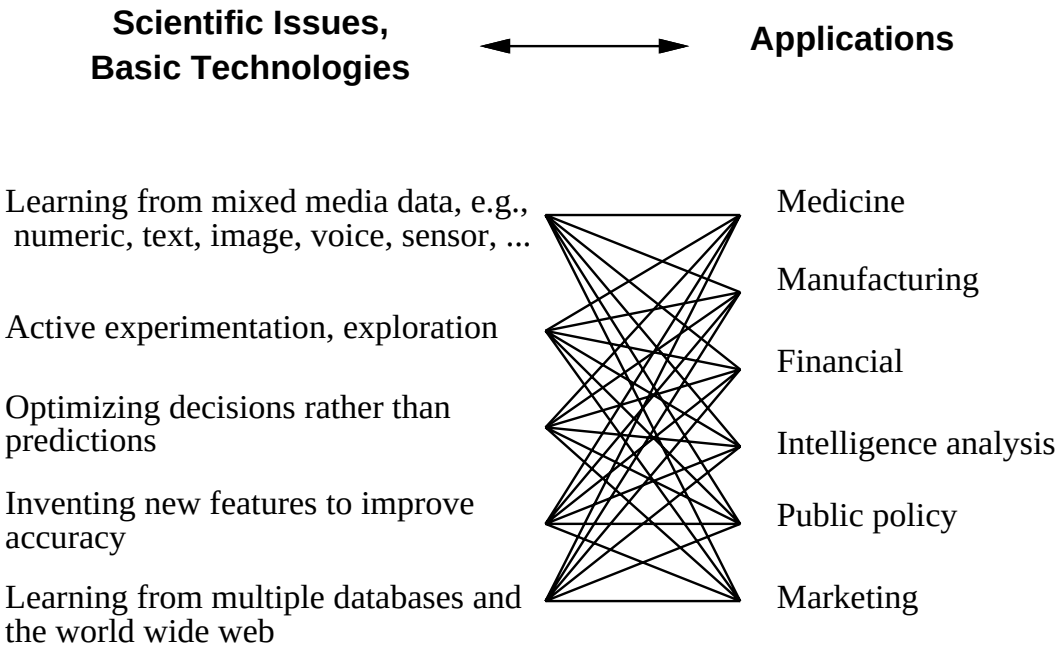


Figure 4: Research on basic scientific issues (left) will impact future data mining applications in many areas (right).

ish.science.gmu.edu) and the University of Bristol (www.cs.bris.ac.uk/Teaching/MachineLearning).

Acknowledgements

The ideas presented here have been shaped in many discussions with faculty and students in the Center for Automated Learning and Discovery (CALD) at Carnegie Mellon. This paper has also benefited from comments by Christos Faloutsos, Tosh Munakata, and an anonymous referee. This research was supported in part by the Darpa HPKB program under contract F30602-97-1-0215, and in part by contributions from the Corporate Members of CALD.

4 References

1. Agrawal, R., Imielinski, T. and A. Swami (1993). Database mining: a performance perspective, *IEEE Trans. on Knowledge and Data Engineering*, 5(6), pp. 914-925.

2. Chauvin, Y., and Rumelhart, D. (1995). BACKPROPAGATION*: theory, architectures, and applications*, edited collection, Lawrence Erlbaum Assoc., Hillsdale, NJ.

3. Clark, P., and Niblett, R. (1989). The CN2 induction algorithm, *Machine Learning*, 3, Kluwer Academic Publishers, pp. 261-284.

4. Gray, J., Bosworth A., Layman A., and Pirahesh, H.(1995). Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Totals, Microsoft Technical Report MSR-TR-95-22.

5. Heckerman, D., Geiger, D., and Chickering, D. (1995). Learning Bayesian networks: the combination of knowledge and statistical data, *Machine Learning*, 20, Kluwer Academic Publishers, p. 197.

6. Faloutsos C., and Lin, K. (1995). FastMap: A Fast Algorithm for Indexing, Data-Mining and Visualization, *ACM SIGMOD*, pp 163-174.

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

8. Muggleton, S. (1995) *Foundations of inductive logic programming*, Englewood Cliffs, NJ: Prentice Hall.

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