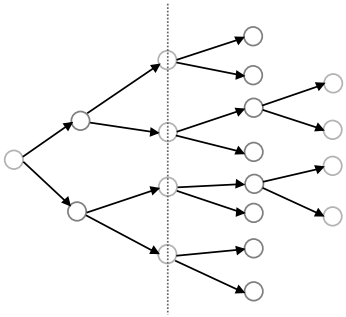
from *SIAM News*, Volume 34, Number 5

A Top-Ten List for Data Mining

By Robert Grossman

Data mining is the semi-automatic extraction of models, patterns, changes, anomalies, and other statistically significant structures from large data sets. During the past decade, the amount of data available for analysis has grown exponentially, while the number of scientists, engineers, and mathematicians available to analyze it has remained essentially level. In some sense, the role of data mining is to fill this gap. Without effective algorithms and good software implementations, most of the data written will never be read. This would be sad, and so in that sense, the goal of data mining is to make the world a happier place—at least that's what I tell my friends.

Five years ago, as the field of data mining was beginning to pick up momentum, an outsider could have accused it of being concerned with the building of traditional statistical models on labeled numerical vectors that came from a single application in a single location, if with a great deal more enthusiasm than a standard statistician would bring to the problem. I argue in this note that the field is much richer now. The First SIAM International Conference on Data Mining took place in Chicago, April 5–7, 2001, with about two hundred fifty people in attendance. This article is basically my "top-ten list" of things that could be learned at the conference. (Remember that the role of such a list can be to irritate as much as to educate.)



Here is a common set up in data mining: One is given a data set with points x and labels y. One is also given a class of models M. (Model types include linear classifiers, tree-based classifiers, neural networks, and so forth.) The data mining game is played in two steps. In the learning step, the goal is to use the data set to find a statistical model f in M such that $y = f(x)$ as often as possible. In the validation step, additional data is used to measure the misclassification rate—that is, how often y is not equal to $f(x)$. There is an essential tension between how well f fits the learning set and how badly it misclassifies the validation set. Probability models can be used to make this precise. The f that is produced is called a classifier, and the vector x is called a feature vector.

What has just been described is a standard problem in statistics: model fitting, i.e., estimating f in M from data. The intent of data mining is the partial automation of this process. It is still a challenge.

Given this background, here is my top-ten list:

1. Data mining has two aspects: data mining "in the small" and data mining "in the large." Data mining in the large is concerned with fitting statistical models to data sets. Data mining in the small is concerned with the smart counting of local patterns, such as association rules, clusters, and anomalies. Several speakers at the conference emphasized this point of view. One speaker, for example, described how smart counting of local structures can improve the fitting of global models.

2. Most data is distributed. The traditional approach to data mining has been to collect all the data at one central location and to build models. From one perspective, much of the interest in data mining is in finding patterns when a data set is combined with other data sets. Because most of the "other" data is in some other location, most data is distributed.

3. Most data is unlabeled. Several talks mentioned co-learning, a technique in which classifiers on labeled data are improved by the addition of records from a larger collection of unlabeled data. Given that most data is unlabeled, a technique like this can be quite useful.

4. Most data is non-numerical. Early work in data mining tended to focus on finding patterns and models in numerical data. Today, more and more data is non-numerical, and more and more emphasis is being placed on the mining of non-numerical data (with examples including text data, Web data, bioinformatics data, and multimedia data).

5. The Web can be viewed as 4 billion feature vectors in $\mathbf{R}^{100,000}$. There is a simple way to map a page of text, or hypertext, to a feature vector x. First, fix a collection of words, called a dictionary. Assume that the dictionary has 100,000 words. Given a Web page, count the number of times each word occurs and create a vector of the counts in $\mathbf{R}^{100,000}$. For example, if the word "system" is the 305th word in the dictionary and if "system" occurs 3 times in a document, then the 305th coordinate in the feature vector for the document would be 3. It is now easy to compute the distance between two documents: Simply compute the distance between two feature vectors. This application has provided a minor renaissance for those interested in sparse linear algebra. It has also invalidated excuses for a paper not to contain numerical studies: There is plenty of data, and it is easy to get.

6. Mining grid data is still hard. Rainfall varies spatially and temporally, but a problem as simple as asking for correlations between rainfall and the incidence of infectious diseases at various locations is still a challenge. A number of papers at the conference dealt with the mining of spatial–temporal data or grid data, but work in this area is really just beginning.

7. Nearest neighbor is not dead. Perhaps the simplest classification method is to use the learning set as the model itself. Given an unlabeled data point, the classifier labels it by using the label of its nearest neighbor, or its nearest k neighbors, and voting. For small data sets this is a practical algorithm, and for larger data sets it is a tempting algorithm. Simple variants of it are the basis for several e-commerce recommendation systems, where the neighbors are called mentors and the products bought by the mentors are the ones suggested by the recommendation system.

8. Every ten years, a community actually develops a new and useful algorithm. I estimated the period as ten years (it turns out to be exactly ten years) after a careful and extensive analysis of more than 50,000 technical articles (it turns out to have been exactly 50,000). Unfortunately, the analysis is too complex to be presented here. Once annointed, of course, the algorithm turns out to be very old. In the data mining area of classification, one could argue that the last thirty years have produced neural networks, tree-based classifiers, and support vector machines. Support vector machines are the new kids on the block and were well represented at the conference. Support vector machines are useful in part because they produce classifiers by reducing classification to well-understood optimization problems.

9. One never talks about anything important. Data is collected so that people can use it in some way, either to gain insights or take actions. Broadly speaking, there are three phases in the move from data to action: Phase 1 is to clean and transform the data. Phase 2 is to apply data mining algorithms. Phase 3 is to deploy the results in some way or to gain some new understanding of the data. In general, 50% of a project is devoted to phase 1, 50% to phase 3, and 0% to phase 2. This conference, like all other data mining conferences, was basically concerned with algorithms and systems for phase 2.

10. Forget knowledge. Data mining is commonly defined as the automatic discovery of valid, novel, and potentially useful knowledge from data. Although defining data mining in this way is common, using data mining in this way is not. These days, it is probably more helpful to view data mining as a map from learning sets to models or patterns, where models capture global structure and patterns capture local structure. There is now an XML language called the Predictive Model Markup Language (PMML) for models and patterns, and so, very concretely, data mining can be thought of as a map from learning sets to PMML.

Robert Grossman divides his time between Magnify, Inc., and the Laboratory for Advanced Computing, University of Illinois at Chicago. With Jiawei Han of Simon Fraser University and Vipin Kumar of the University of Minnesota, he co-chaired the organizing committee for the First SIAM International Conference on Data Mining.