Appears in the
National Academy of Engineering (NAE)
US Frontiers of Engineering 2000

Data Mining and Visualization

Ron Kohavi

Blue Martini Software
2600 Campus Drive
San Mateo, CA, 94403, USA

ronnyk@bluemartini.com

Abstract

Data Mining is the process of identifying new patterns and insights in data. As the volume of data collected and stored in databases grows, there is a growing need to provide data summarization (e.g., through visualization), identify important patterns and trends, and act upon the findings. Insight derived from data mining can provide tremendous economic value, often crucial to businesses looking for competitive advantages. A short review of data mining and important theoretical results is provided, followed by recent advances and challenges.

1 Introduction

Yahoo!'s traffic increased to 680 million page views per day on average…
Yahoo!'s communication platform delivered 4.4 billion messages… in June [2000]
-- Yahoo! Press Release, July 11, 2000

The amount of data stored on electronic media is growing exponentially fast. Today's data warehouses dwarf the biggest databases built a decade ago [1], and making sense of such data is becoming harder and more challenging. Online retailing in the Internet age, for example, is very different than retailing a decade ago because the three most important factors of the past (location, location, and location) are irrelevant for online stores.

One of the greatest challenges we face today is making sense of all this data. Data mining, or knowledge discovery, is the process of identifying new patterns and insights in data, whether it is for understanding the Human Genome to develop new drugs, for discovering new patterns in recent Census data to warn about hidden trends, or for understanding your customers better at an electronic webstore in order to provide a personalized one-to-one experience. The examples in this paper are from the e-commerce world, but data mining has been used extensively in multiple domains including many scientific applications. The paper is also restricted to structured mining; significant literature exists for text mining and information retrieval.

The paper is organized as follows. Section 2 introduces data mining tasks and models, followed by a quick tour of some theoretical results in Section 3. Section 4 reviews the recent advances, followed by some challenges in Section 5 and a summary in Section 6.

2 Data Mining Tasks and Models

The most exciting phrase to hear in science, the one that heralds new discoveries, is not 'Eureka!' (I found it!) but 'That's funny ...'
Isaac Asimov

Data mining, sometimes referred to as knowledge discovery [2], is at the intersection of multiple research areas, including Machine Learning [3, 4, 5, 6], Statistics [7, 8, 9], Pattern Recognition [10, 11, 12], Databases [13, 14], and Visualization [15, 16]. Good marketing and business-oriented data mining books are also available [17, 18, 19]. With the maturity of databases and constant improvements in computational speed, data mining algorithms that were too expensive to execute are now within reach.

Data mining serves two goals:

1. **Insight**: identify patterns and trends that are comprehensible, so that action can be taken based on the insight. For example, characterize the heavy spenders on a web site, or people that buy product X. By understanding the underlying patterns, the web site can be personalized and improved. The insight may also lead to decisions that affect other channels, such as brick-and-mortar stores' placement of products, marketing efforts, and cross-sells.

2. **Prediction**: a model is built that predicts (or scores) based on input data. For example, a model can be built to predict the propensity of customers to buy product X based on their demographic data and browsing patterns on a web site. Customers with high scores can be used in a direct marketing campaign. If the prediction is for a discrete variable with a few values (e.g., buy product X or not), the task is called *classification*; if the prediction is for a continuous variable (e.g., customer spending in the next year), the task is called *regression*.

The majority of research in data mining has concentrated on building the best models for prediction. Part of the reason, no doubt, is that a prediction task is well defined and can be objectively measured on an independent test-set. Given a dataset that is labeled with the correct predictions, it is split into a training set and a test-set. A *learning algorithm* is given the *training set* and produces a model that can map new unseen data into the prediction. The model can then be evaluated for its accuracy in making predictions on the unseen *test-set*. Descriptive data mining, which yields human insight, is harder to evaluate, yet necessary in many domains because the users may not trust predictions coming out of a black box or because legally one must explain the predictions. For example, even if a Perceptron algorithm [20] outperforms a loan officer in predicting who will default on a loan, the person requesting a loan cannot be rejected simply because he is on the wrong side of a 37-dimensional hyperplane; legally, the loan officer must explain the reason for the rejection.

The choice of a predictive model can have a profound influence on the resulting accuracy and on the ability of humans to gain insight from it. Some models are naturally easier to understand than others. For example, a model consisting of if-then rules is easy to understand, unless the number of rules is too large. Decision trees, are also relatively easy to understand. Linear models get a little harder, especially if discrete inputs are used. Nearest-neighbor algorithms in high dimensions are almost impossible for users to understand, and non-linear models in high dimensions, such as Neural Networks are the most opaque.

One way to aid users in understanding the models is to visualize them. MineSet [21], for example, is a data mining tool that integrates data mining and visualization very tightly. Models built can be viewed and interacted with. Several movies are available at: http://www.sgi.com/software/mineset/demos.html. Figure 1 shows a visualization of the Naïve-Bayes classifier. Given a target value, which in this case was who earns over $50,000 in the US working population, the visualization shows a small set of "important" attributes (measured using mutual information or cross-entropy). For each attribute, a bar chart shows how much "evidence" each value (or range of values) of that attribute provides for the target label. For example, higher education levels (right bars in the education row) imply higher salaries because the bars are higher. Similarly, salary increases with age up to a point and then decreases, and salary increases with the number of hour worked per week. The combination of a back-end algorithm that bins the data, computes the importance of hundreds of attributes, and then a visualization that shows the important attributes visually, makes this a very useful tool that helps identify patterns. Users can interact with the model by clicking on attribute values and seeing the predictions that the model makes.

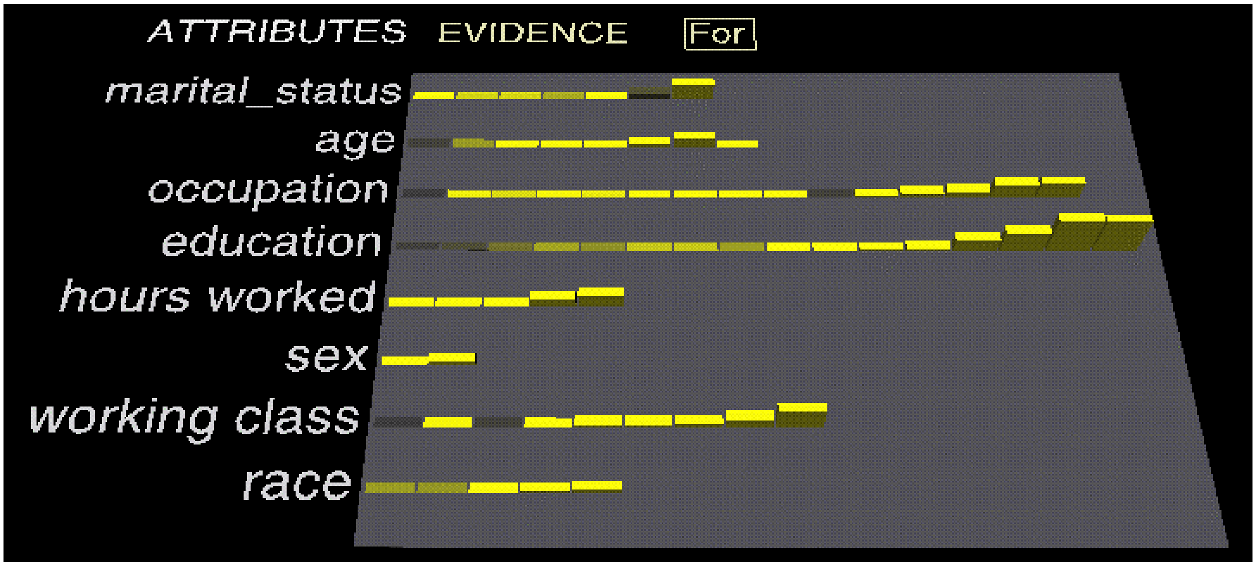

Figure 1: A visualization of the Naive-Bayes classifier

3 Data Mining Theory

Reality is the murder of a beautiful theory by a gang of ugly facts
Robert L. Glass [22]

This section provides a short review of some theoretical results in data mining.

1. **No free lunch**. A fundamental observation is that learning is impossible without assumptions. If all concepts are equally likely, then not only is learning impossible, but no algorithm can dominate another in generalization accuracy [23, 24]. The result is similar to the proof that data compression is not always possible (yet everyone enjoys the saving provided by data compression algorithms). In practice, learning is very useful because the world does not present us with uniform worst-case scenarios.
2. **Consistency**. While parametric models (e.g., linear regression) are known to be of limited power, non-parametric models can be shown to learn "any reasonable" target concept given enough data. For example, nearest-neighbor algorithms with a growing neighborhood have been shown to have asymptotically optimal properties under [25]. Similar results exist for the consistency of decision tree algorithms [26]. While asymptotic consistency results are comforting because they guarantee that with enough data the learning algorithms will converge to the target concept one is trying to

learn, our world is not so ideal. We are always given finite amounts of data from which to learn and rarely do we reach *asymptopia*.

An excellent example of the problem of nearest-neighbors not being so "near" is as follows [8]. Assume a 20 dimensional unit ball (radius = 1) centered at the origin with 100,000 points uniformly distributed. The median distance from the origin to the closest point is 0.55, more than half way to the boundary. Most points are therefore closer to the boundary of the sample space than to another point! In a few dimensions, standard visualization methods work well; in higher dimensions our intuition is commonly wrong and data mining can help.

3. **PAC learning.** Probably Approximately Correct learning [27,28] is a concept introduced to provide guarantees about learning. Briefly, assuming that the target can be described in a given hypothesis space (e.g., disjunctions of conjunctions of length k), a PAC learning algorithm can learn the approximate target with high probability. The two parameters typically given as input to a PAC learning algorithm are *epsilon* and *delta*. The algorithm must satisfy that at least (1-delta) fraction of the time, the error between the actual target concept and the predictions made is bounded by epsilon. PAC learning theory defines bounds on the number of examples needed to provide such guarantees. One of the more interesting results in PAC learning theory is that a weak learning algorithm, which can classify more accurately than random guessing (e.g., epsilon < 0.5), can always be boosted into a strong learning algorithm, which can produce classifiers of arbitrary accuracy [29] (more training data will be needed, of course). This theoretical result led to interesting practical developments mentioned below.
4. **Bias-Variance decomposition**. The expected error of any learning algorithm for a given target concept and training set size can be decomposed into two terms: the bias and the variance [30]. The importance of the decomposition is that it is valid for finite training set sizes, not asymptotically, and that the terms can be measured experimentally. The bias measures how closely the learning algorithm’s average guess (over all possible training sets of the given training set size) matches the target. The variance measures how much the learning algorithm’s guess varies for different training sets of the given size. Many unsuccessful and painful routes have been taken by researchers trying to improve a learning algorithm by enlarging the space of models, which can reduce the bias, but may also increase the variance. For example, Figure 2 shows how 10 data points, assumed to be slightly noisy, can be reasonably fit with a quadratic polynomial, and perfectly fit with a 9-th degree polynomial that overfits the data. A learning algorithm trying to fit high-degree polynomials will generate very different polynomials for different training sets, and hence have high variance. A learning algorithm that always fits a linear model will be more stable, but will be biased for quadratic and higher-order models. Making the analogy to decision tree models, finding the smallest decision tree that perfectly fits the data (an NP-hard problem) takes a long time and often results in worse generalizations than using a simple greedy algorithm that approximately fits the data. The reason is that the smallest perfect trees generated for similar data sets of the same size vary significantly in structure and predictions and hence the expected error has a large variance term. For several algorithms, it is known how to move along this bias-variance tradeoff through regularization techniques.

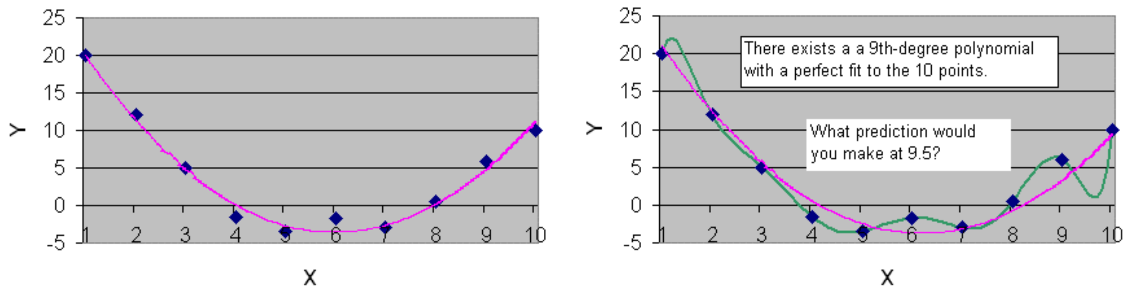

Figure 2: the left figure shows a quadratic fit to data, but the fit is not perfect. The right figure shows a 9th degree polynomial fit that perfectly passes through all the data points. If the data is expected to contain some noise, the model on the left will probably make a better prediction at x=9.5 than the squiggly model on the right, which (over)fits the data perfectly.

4 Recent Advances

The advancement of the arts, from year to year, taxes our credulity and seems to presage the arrival of that period when human improvement must end
Henry Elsworth, US Patent Office, 1844

This section provides a brief summary of recent advances in the field of data mining. Several advances specific to machine learning are described in the AI Magazine [31].

1. **Multiple model learning.** Two learning techniques developed in the last few years have had significant impact: Bagging and Boosting. Both methods learn multiple models and vote them in order to make a prediction and both were shown to be very successful in improving prediction accuracy on real data [32, 33]. Bagging [34] generates bootstrap samples by repeatedly sampling the training set with replacement. A model is built for each sample and they are then uniformly voted. Boosting algorithms, and specifically the AdaBoost algorithm [35], generate a set of classifiers in sequence. Each classifier is given a training set where examples are reweighted to highlight those previously misclassified.
2. **Associations.** A common problem in retailing is to find combinations of products that when bought together imply the purchase of another product. For example, an association might be that the purchase of hot dogs and Coke implies the purchase of chips with high probability. Several algorithms were developed to find such associations for market basket analysis [13]. Given a minimum support (percentage of the data that has to satisfy the rule) and a minimum confidence (the probability that the right hand side is satisfied given the left-hand side), the algorithms find all associations. Note that unlike prediction tasks, this is a descriptive task where the result is well defined and the algorithms must be sound and complete. The main observation in these algorithms is that in order for a combination of size L to have minimum support, each of its subsets of size (L-1) must have minimum support.
3. **Scalability (both speed and dataset size).** Several advances have been made in scaling algorithms to larger datasets and parallelizing them [36,37, 14]

5 Challenges

Laggards follow the path of greatest familiarity. Challengers, on the other hand, follow the path of greatest opportunity, wherever it leads.
Competing for the Future / Hamel and Prahalad

This section provides some challenging problems.

1. **Make Data Mining Models Comprehensible to Business Users.** Business users need to understand the results of data mining. Few data mining models are easy to understand and techniques need to be developed to explain or visualize existing ones (e.g., [38]) or new models that are simple to understand with matching algorithms need to be derived. This is particularly hard for regression models. A related problem is that association algorithms usually derive too many rules (e.g., 100,000) and we need to find ways to highlight the "interesting" rules or families of associations.
2. **Make Data Transformations and Model Building Accessible to Business Users.** An important issue that has not been mentioned above is the need to translate user's questions into a data mining problem in relational format. This often requires writing SQL, Perl scripts, or small programs. Even defining *what* the desired transformations and features should be is a knowledge-intensive task requiring significant understanding of the tasks, the algorithms, and their capabilities. Can we design a transformation language more accessible to business users? Can we automatically transform the data?
3. **Scale algorithms to large volumes of data.** It is estimated that the amount of text in the Library of Congress can be stored in about 17 terabytes of disk space [17]. The package-level detail database used to track shipments at UPS is also 17 terabytes. Most data mining algorithms can handle a few gigabytes of data at best, so there are three to four orders of magnitude to grow before we can attack the largest databases that exist today. In addition, most algorithms learn in batch, but many applications require real-time learning.
4. **Close the loop: identify causality, suggest actions, and measure their effect.** Discoveries may reveal correlations that are not causal. For example, human reading ability correlates with shoe size, but wearing larger shoes will not improve one's reading ability. (The correlation is explained by the fact that children have smaller shoe sizes and cannot read as well.) Controlled experiments and measurements of their effects can help pinpoint the causal relationships. One advantage of the online world is that experiments are easy to conduct: changing layout, emphasizing certain items, and offering cross-sells can all be easily done and their effect can be measured. For electronic commerce, the World Wide Web is a great laboratory for experiments, but our learning techniques need to improve to offer interventions and take them into account.
5. **Cope with privacy issues.** Data mining holds the promise of reducing the amount of junk mail we receive by providing us with more targeted messages. However, data collection can also lead to abuses of the data, raising with many social and economic issues. This is doubly true in the online world where every page and every selection we make can be recorded.

6 Summary

You press the button, and we'll do the rest
-- Kodak advertisement

Taking pictures and developing them (or loading them into a computer) has become trivial: there is no need to focus, adjust the shutter speed and aperture, or know anything about chemistry to take great pictures. Data mining and related technologies have had significant advances, but we have yet to build the equivalent of the point-and-click cameras. This short review of the basic goals of data mining, some theory, and recent advances should provide those interested with enough information to see the value of data mining and use it to find nuggets; after all, almost everyone has access to the main ingredient that is needed: *data*.

7 Acknowledgments

I would like to thank Carla Brodley, Tom Dietterich, Pedro Domingos, Rajesh Parekh, Ross Quinlan, and Zijian Zheng for their comments and suggestions.

References

[1] Kimball, R. & Merz, R. (2000), The Data Webhouse Toolkit: Building the Web-Enabled Data Warehouse, John Wiley & Sons.

[2] Fayyad, U. M., Piatetsky-Shapiro, G. & Smyth, P. (1996a), From data mining to knowledge discovery: An overview, in ‘Advances in Knowledge Discovery and Data Mining’, AAAI Press and the MIT Press, chapter 1, pp. 1-34.

[3] Dietterich, T. G. & Shavlik, J. W., editors (1990), Readings in Machine Learning, Morgan Kaufmann.

[4] Quinlan, J. R. (1993b), C4.5: Programs for Machine Learning, Morgan Kaufmann.

[5] Mitchell, T. M. (1997), Machine Learning, McGraw-Hill.

[6] Kearns, M. J. & Vazirani, U. V. (1994), An Introduction to Computational Learning Theory, MIT Press.

[7] Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. (1984), Classification and Regression Trees, Wadsworth International Group.

[8] Friedman, J., Hastie, T. & Tibshirani, R. (to appear), The Elements of Statistical Learning: Prediction, Inference and Data Mining.

[9] Fukunaga, K. (1990), Introduction to Statistical Pattern Recognition, Academic Press.

[10] Duda, R. & Hart, P. (1973), Pattern Classification and Scene Analysis, Wiley. Second edition to appear with Stork in October 2000.

[11] Ripley, B. D. & Hjort, N. L. (1995), Pattern Recognition and Neural Networks, Cambridge University Press.

[12] Bishop, C. M. (1995), Neural Networks for Pattern Recognition, Oxford University Press.

[13] Srikant, R. & Agrawal, R. (1995), Mining generalized association rules, in ‘Proceedings of the 21st International Conference on Very Large Databases’.

[14] Shafer, J., Agrawal, R. & Mehta, M. (1996), SPRINT: a scalable parallel classifier for data mining, in ‘Proceedings of the 22nd International Conference on Very Large Databases (VLDB)’.

[15] Tufte, E. R. (1983), The Visual Display of Quantitative Information, Graphics Press, Cheschire, CT.

[16] Cleveland, W. S. (1993), Visualizing Data, Hobart Press.

[17] Berry, M. J. & Linoff, G. S. (2000), Mastering Data Mining, John Wiley & Sons, Inc.

[18] Berson, A., Thearling, K. & Smith, S. J. (1999), Building Data Mining Applications for CRM, McGraw-Hill.

[19] Dhar, V. (1996), Seven Methods for Transforming Corporate Data Into Business Intelligence, Prentice Hall.

[20] Minsky, M. L. & Papert, S. (1988), Perceptrons : an Introduction to Computational Geometry, MIT Press. Expanded edition.

[21] Brunk, C., Kelly, J. & Kohavi, R. (1997), MineSet: an integrated system for data mining, in D. Heckerman, H. Mannila, D. Pregibon & R. Uthurusamy, eds, ‘Proceedings of the third international conference on Knowledge Discovery and Data Mining’, AAAI Press, pp. 135-138. http://mineset.sgi.com.

[22] R. Glass (1996), The relationship between theory and practice in software engineering, CACM Nov 1996, vol. 39, No 11.

[23] Wolpert, D. H. (1994), The relationship between PAC, the statistical physics framework, the Bayesian framework, and the VC framework, in D. H. Wolpert, ed., ‘The Mathematics of Generalization’, Addison Wesley.

[24] Schaffer, C. (1994), A conservation law for generalization performance, in ‘Machine Learning: Proceedings of the Eleventh International Conference’, Morgan Kaufmann, pp. 259-265.

[25] Fix, E. & Hodges, J. (1951), Discriminatory analysis--nonparametric discrimination: Consistency properties, Technical Report 21-49-004, report no. 04, USAF School of Aviation Medicine, Randolph Field, Tex.

[26] Gordon, L. & Olshen, R. A. (1984), ‘Almost sure consistent nonparametric regression from recursive partitioning schemes’, Journal of Multivariate Analysis 15, 147-163.

[27] Valiant, L. G. (1984), ‘A theory of the learnable’, Communications of the ACM 27, 1134-1142.

[28] Kearns, M. J. & Vazirani, U. V. (1994), An Introduction to Computational Learning Theory, MIT Press.

[29] Schapire, R. E. (1990), ‘The strength of weak learnability’, Machine Learning 5(2), 197-227.

[30] Geman, S., Bienenstock, E. & Doursat, R. (1992), ‘Neural networks and the bias/variance dilemma’, Neural Computation 4, 1-48.

[31] Dietterich, T. G. (1997), ‘Machine learning research: Four current directions’, AI Magazine 18(4).

[32] Quinlan, J. R. (1996a), Bagging, boosting, and c4.5, in ‘Proceedings of the Thirteenth National Conference on Artificial Intelligence’, AAAI Press and the MIT Press, pp. 725-730.

[33] Bauer, E. & Kohavi, R. (1999), ‘An empirical comparison of voting classification algorithms: Bagging, boosting, and variants’, Machine Learning 36, 105-139.

[34] Breiman, L. (1996b), ‘Bagging predictors’, Machine Learning 24, 123-140.

[35] Freund, Y. & Schapire, R. E. (1997), ‘A decision-theoretic generalization of on-line learning and an application to boosting’, Journal of Computer and System Sciences 55(1), 119-139.

[36] Provost, F. & Kolluri, V. (1999), ‘A survey of methods for scaling up inductive algorithms’, Data Mining and Knowledge Discovery 3(2), 131-169.

[37] Freitas, A. A. & Lavington, S. H. (1998), Mining Very Large Databases With Parallel Processing, Kluwer Academic Publishers.

[38] Barry Becker, Ron Kohavi, and Dan Sommerfield, Visualizing the Simple Bayesian Classifier, *KDD Workshop on Issues in the Integration of Data Mining and Data Visualization*, 1997.