Automated Learning and Discovery: State-Of-The-Art and Research Topics in a Rapidly Growing Field

Sebastian Thrun, Christos Faloutsos, Tom Mitchell, Larry Wasserman

Center for Automated Learning and Discovery
Carnegie Mellon University
Pittsburgh, PA 15213

September 1998

1 Introduction

The field of automated learning and discovery—often called data mining, machine learning, or advanced data analysis—is currently undergoing a major change. The progressing computerization of professional and private life, paired with a sharp increase in memory, processing and networking capabilities of today's computers, make it increasingly possible to gather and analyze vast amounts of data. For the first time, people all around the world are connected to each other electronically through the Internet, making available huge amounts of online data at an exponential rate.

Sparked by these innovations, we are currently witnessing a rapid growth of a new industry, called the data mining industry. Companies and governments have begun to realize the power of computer-automated tools for systematically gathering and analyzing data. For example, medical institutions have begun to utilize data-driven decision tools for diagnostic and prognostic purposes; various financial companies have begun to analyze their customers' behavior in order to maximize the effectiveness of marketing efforts; the Government now routinely applies data mining techniques to discover national threats and patterns of illegal activities in intelligence databases; and an increasing number of factories apply automatic learning methods to optimize process control. These examples illustrate the immense societal importance of the field.

At the same time, we are witnessing a healthy increase in research activities on issues related to automated learning and discovery. Recent research has led to revolutionary progress, both in the type methods that are available, and in the understanding of their characteristics. While the broad topic of automated learning and discovery is inherently cross-disciplinary in nature—it falls right into the intersection of disciplines like statistics, computer science, cognitive psychology, robotics, and it users such as medicine, social sciences, and public policy—these fields have mostly studied this topic in isolation. So where is the field, and where is it going? What are the most promising research directions? What are opportunities of cross-cutting research, and what is worth pursuing?

2 The CONALD Meeting

To brainstorm about these and similar questions, Carnegie Mellon University (CMU) recently hosted the CONALD meeting (short for "Conference on Automated Learning and Discovery"). CONALD's goal was to bring together leading scientists from the various disciplines involved, to brainstorm about the following central questions:

1. **State of the art.** What is the state of the art? What are examples of successful systems?
2. **Goals and impact.** What are the long-term goals of the field? What will be the most likely future impact of the area?
3. **Promising research topics.** What are examples of the most promising research topics that should be pursued in the next three to five years, and beyond?
4. **Opportunities for cross-disciplinary research.** Which are the most significant opportunities for cross-cutting research?

CONALD, which took place in June 1998, drew approximately 250 participants. The majority of CONALD's attendants were computer scientists or statisticians. CONALD featured seven plenary talks, given by **Tom Dietterich** (Oregon State University) with a talk on "Learning for Sequential Decision Making," **Stuart Geman** (Brown University) with a talk on "Probabilistic Grammars and their Applications," **David Heckerman** (Microsoft Research) with a talk on "A Bayesian Approach to Causal Discovery," **Michael Jordan** (MIT, now UC Berkeley) with a talk on "Graphical Models and Variational Approximation," **Daryl Pregibon** (AT&T Research) with a talk on "Realtime Learning and Discovery in Large Scale Networks," **Herb Simon** (CMU) with a talk on "Using Machine Learning to Understand Human Learning," and **Robert Tibshirani** (Univ. of Toronto, now Stanford Univ.) who talked about "Learning from Data: Statistical Advances and Challenges."

Apart from the plenary talks, which were aimed at familiarizing researchers from different scientific communities with each other's research, the meeting featured a collection of seven workshops (see Table 1), where workshop participants discussed a specific topic in depth. Each workshop was organized by an inter-disciplinary team of researchers, and the topic of the workshops related to research done in several areas. Workshop organizers invited up to two leading scientists per workshop, using funds provided by the National Science Foundation (NSF). At the last day, all workshop participants met in a single room for two sessions called "thesis topics," where workshop chairs summarized the results of their workshop and laid out concrete, promising research topics, as examples of feasible and promising topics for future research.

3 The Need For Cross-Disciplinary Research

A key objective of the CONALD meeting was to investigate into the role of a cross-disciplinary approach. There was a broad consensus that the issues at stake are highly interdisciplinary. Workshop participants and organizers alike expressed that each discipline has studied unique aspects of the problem and therefore can contribute a unique collection of approaches. Statistics—undoubtedly the field with the longest-reaching history—has developed powerful methods for gathering, learning from and reasoning with data, often studied in highly restrictive settings. Researchers in AI have

1. **Learning Causal Bayesian Networks**
   organized by Richard Scheines and Larry Wasserman.
2. **Mixed-Media Databases**
   organized by Shumeet Baluja, Christos Faloutsos, Alex Hauptmann and Michael Witbrock.
3. **Machine Learning and Reinforcement Learning for Manufacturing**
   organized by Sridhar Mahadevan and Andrew Moore.
4. **Large-Scale Consumer Databases**
   organized by Mike Meyer, Teddy Seidenfeld, and Kannan Srinivasan.
5. **Visual Methods for the Study of Massive Data Sets**
   organized by Bill Eddy and Steve Eick.
6. **Learning from Text and the Web**
   organized by Yiming Yang, Jaime Carbonell, Steve Fienberg, and Tom Mitchell.
7. **Robot Exploration and Learning**
   organized by Howie Choset, Maja Matarić, and Sebastian Thrun.

Table 1: CONALD workshop topics. A detailed report of each workshop can be found at CONALD's Web site and as Technical Report [1].

---

explored learning from huge datasets with high-dimensional feature spaces (e.g., learning from text). Database researchers have devised efficient method for storing and processing huge datasets, and they have devised highly efficient methods for answering certain types of questions (such as membership queries). Various applied disciplines have contributed specific problem settings and datasets of societal importance. Many participants expressed that by bringing together these various disciplines, there is an opportunity to integrate each other's insides and methodologies, to gain the best of all worlds. In addition, an interdisciplinary discourse is likely to reduce the danger of wasting resources by re-discovering each other's results.

Historically, issues of automated learning and discovery have been studied by various scientific disciplines, such as statistics, computer science, cognitive psychology, and robotics. In many cases, each discipline pursued its research in isolation, studying specific facets of the general problem, and developing a unique set of methods, theory, and terminology. To illustrate this point, Table 2 shows a modified version of a "statistics-to-AI dictionary," shown by Rob Tibshirani in his plenary talk (and later augmented by Andrew W. Moore) to illustrate difference in terminology.

Statistics Term	AI Term
statistics	data learning
regression	progression, straight neural network
discrimination	pattern recognition
prediction sum squares	generalization ability
fitting	learning
empirical error	training set error
sample	training set
experimental design	active learning

Table 2: An extended version of Tibshirani's Statistics-to-Artificial Intelligence dictionary illustrates the differences in terms used in two scientific field concerned with the same questions.

4 State Of The Art, Promising Research Directions

Characterizing the state-of-the-art is not an easy endeavor, as the space of commercially available approaches is large, and research prototypes exist for virtually any problem in the area of automated learning and discovery. Thus, we will attempt to give some broad characterizations that, in our opinion, most people agreed to.

There was an agreement that function fitting (which often goes by the name of supervised learning, pattern recognition, regression, approximation, interpolation—not all of which mean the exact same thing) appears now to be a well-understood problem. This is specifically the case when feature spaces are low-dimensional and sufficient data are available. There exists now a large collection of popular and well-understood function fitting algorithms, such as splines, logistic regression, Back-Propagation, decision trees, and so on. Many of these tools form the backbone of commercial data-mining tools, where they analyze data and predict future trends.

Clustering of data in low-dimensional spaces has also been studied extensively, and today we possess a large collection of methods for clustering data, which are specifically applicable if feature spaces are low-dimensional and plenty of data is available.

Of course, data mining is more than just applying a learning algorithm to a set of data. Existing tools provide powerful mechanisms for data preparation, visualization, and interpretation of the results. Many of the tools work well in low-dimensional, numerical feature spaces—yet they cease to work if the data is high-dimensional and non-numerical (such as text). There was a reasonable consensus that such work is important, and better methodologies are needed to do data preparation, processing, and visualization.

The workshop sessions generated a large number of research topics. Despite the fact that the workshops were organized around different problem/application domains, many of these topics co-occurred in multiple workshops. Among the most notable were the following:

- **Active learning/experimental design.** Active learning (AI jargon) and experimental design (statistics jargon) addresses the problem of choosing which experiment to run during learning. It assumes that during learning, there is an opportunity to influence the data collection. For example, a financial institution worried about customer retention might be interesting *why* customers discontinue their business with this institution, so that potential candidates can be identified and the appropriate actions can be taken. It is impossible, however, to interview millions of customers, and in particular those who changed to another provider are difficult to ask—so whom should one call to learn the most useful model? In robot learning, to name a second example, the problem of "exploration" is a major problem. Robot hardware is slow; yet, most learning methods depend crucially on a wise choice of learning data. Active learning addresses the question of how to explore.

- **Cumulative learning.** Many practical learning problems are characterized by a continual feed of data. For example, databases of customer transaction or medical records grow incrementally. Often, the sheer complexity of the data and/or the statistical algorithm used for their analysis prohibits evaluating the data from scratch every day. Instead, data have to be analyzed cumulatively, as they arrive. This problem is especially difficult if the laws underlying the data generation may change in non-obvious ways: For example, customers' behavior can be influenced by a new regulation, a fashion, a weather pattern, a product launched by a competitor, a recession, or a scientific discovery. Can we devise cumulative learning algorithms that can incrementally incorporate new data, and that can adapt to changes in the process that generated the data?

- **Multitask learning.** Many domains are characterized by families of highly related (though not identical) learning problems. Medical domains are of this type. While each disease poses an individual learning task for which dedicated databases exist, many diseases share similar physical causes and symptoms, making it promising to transfer knowledge across multiple learning tasks. Similar issues arise in user modeling, where knowledge may be transferred across individual users, and in financial domains, where knowledge of one stock might help predicting the future value of another. Can we devise effective multi-task learning algorithms, which generalize more accurately through transferring knowledge across learning tasks?

- **Learning from labeled and unlabeled data.** In many application domains, it is not the data that is expensive; instead, obtaining labels for the data is a difficult and expensive process. For example, software agents that adaptively filter on-line news articles can easily access vast amounts of data almost for free; however, having a user label excessive amounts of data (e.g., expressing his level if interest) is usually prohibitive. Can we devise learning algorithms that exploit the unlabeled data when learning a new concept? If so, what is the relative value of labeled data compared to unlabeled data?

- **Relational learning.** In many learning problems, instances are not described by a static set of features. For example, when finding patterns in intelligence databases, the relation between entities (companies, people) is of crucial importance. Entities in intelligence databases are people, organizations, companies and countries; and the relation of them is of crucial importance when finding patterns of criminal activities such as money laundering. Most of today's learning algorithms require fixed feature vectors for learning. Can we devise relational learning algorithms that consider the relation of multiple instances when making decisions?

- **Learning from extremely large datasets.** Many datasets are too large to be read by a computer more than a few times. For example, many grocery stores collect data of each transaction, often producing gigabytes of data every day. This makes it impossible to apply algorithms that require many passes through the data. Other databases, such as the Web, are too large and too dynamic to permit exhaustive access. Many of the existing algorithms exhibit poor scaling abilities when the dataset is huge. Can we devise learning algorithms that scale up to extremely large databases?

- **Learning from extremely small datasets.** At the other extreme, there are often databases that are too small for current learning methods. For example, in face recognition problems, there is often just a single image of a person available, making it difficult to identify this person automatically in other images. In robotics, the number of examples is often extremely limited; yet many of the popular learning algorithms (e.g., genetic programming, reinforcement learning) require huge amounts of data. How can we reduce the amount of data required for learning? How can we assess the risk involved in using results obtained from small datasets?

- **Learning with prior knowledge.** In many cases, substantial prior knowledge is available about the phenomenon that is being learned. For example, one might possess knowledge about political events, laws and regulations, personal preferences, and economics essential to the prediction of exchange rates between currencies. How can we incorporate such knowledge into our statistical methods? Can we find flexible schemes that facilitate the insertion of diverse, abstract, or uncertain prior knowledge?

- **Learning from mixed media data.** Many data sets contain more than just a single type of data. For example, medical datasets often contain numerical data (e.g., test results), images (e.g., X-rays), nominal data (e.g., person smokes/does not smoke), acoustic data (e.g., the recording of a doctor's voice), and so on. Existing algorithms can usually only cope with a single type of data. How can we design methods that can integrate data from multiple modalities? Is it better to apply separate learning algorithms to each data modality, and to integrate their results, or do we need algorithms that can handle multiple data modalities on a feature-level?

- **Learning casual relationships.** Most existing learning algorithms detect only correlations, but are unable to model causality and hence fail to predict the effect of external controls. For example, a statistical algorithm might detect a strong correlation between chances to develop lung cancer and the observation that a person has yellow fingers. What is more difficult to detect is that both, lung cancer and yellow fingers are caused by a hidden effect (smoking), and hence providing alternative means to reduce the yellowness of the fingers (e.g., a better soap) is unlikely to change the odds of a person developing cancer—despite the correlation! Can we devise learning algorithms that discover causality? If so, what type of assumptions have to be made, to extract causality from a purely observational database, and what implications do they have?

- **Visualization and interactive data mining.** In many applications, data mining is an interactive process, which involves automated data analysis and control decisions by an expert of the domain. For example, patterns in many large-scale consumer or medical databases are often discovered interactively, by a human expert looking at the data, rearranging it, and using computer tools to search for specific patterns. Data visualization is specifically difficult when data is high-dimensional, specifically when it involves non-numerical data such as text. How can we visualize large and high-dimensional data sets? How can we design interactive tools that best integrate the computational power of computers with the knowledge of the person operating them?

This list has been compiled based on the outcomes of the individual workshops, the invited talks, and various discussions that occurred in the context of CONALD. Virtually all of these topics are cross-disciplinary in nature. While this list is necessarily incomplete, it covers the most prominent research issues discussed at CONALD. Detailed reports of each individual workshop, which describe addition topics and open problems, can be found at CONALD's Web site

`http://www.cs.cmu.edu/~conald`

and is also available as Technical Report [1].

We believe that CONALD has successfully contributed to an ongoing dialogue between different disciplines that, for a long time, have studied different facets of one and the same problem: decision making based on historical data. We believe that the cross-disciplinary dialogue, which more than everything else characterized CONALD, is essential for the health of the field, and that it opens up many new, exciting research directions.

Acknowledgements

CONALD was held to celebrate CMU's new Center for Automated Learning and Discovery, a cross-disciplinary center created in 1997, whose sponsorship is gratefully acknowledged. In addition, CONALD was financially supported by the National Science Foundation (NSF), which is gratefully acknowledged. We also thank an anonymous reviewer who helped to improve this report. Finally, we would like to thank all participants who contributed to a lively meeting.

References

[1] Technical Report CMU-CALD-98-100, Carnegie Mellon University, Center for Automated Learning and Discovery. Same authors and title.