International Journal of Information Technology & Decision Making
Vol. 5, No. 4 (2006) 597–604

World Scientific
www.worldscientific.com

10 CHALLENGING PROBLEMS IN DATA MINING RESEARCH

QIANG YANG

Department of Computer Science
Hong Kong University of Science and Technology
Clearwater Bay, Kowloon, Hong Kong, China

XINDONG WU

Department of Computer Science
University of Vermont
33 Colchester Avenue, Burlington, Vermont 05405, USA
xwu@cs.uvm.edu

CONTRIBUTORS: PEDRO DOMINGOS, CHARLES ELKAN, JOHANNES GEHRKE, JIAWEI HAN, DAVID HECKERMAN, DANIEL KEIM, JIMING LIU, DAVID MADIGAN, GREGORY PIATETSKY-SHAPIRO, VIJAY V. RAGHAVAN, RAJEEV RASTOGI, SALVATORE J. STOLFO, ALEXANDER TUZHILIN and BENJAMIN W. WAH

In October 2005, we took an initiative to identify 10 challenging problems in data mining research, by consulting some of the most active researchers in data mining and machine learning for their opinions on what are considered important and worthy topics for future research in data mining. We hope their insights will inspire new research efforts, and give young researchers (including PhD students) a high-level guideline as to where the hot problems are located in data mining.

Due to the limited amount of time, we were only able to send out our survey requests to the organizers of the IEEE ICDM and ACM KDD conferences, and we received an overwhelming response. We are very grateful for the contributions provided by these researchers despite their busy schedules. This short article serves to summarize the 10 most challenging problems of the 14 responses we have received from this survey. The order of the listing does *not* reflect their level of importance.

Keywords: Data mining; machine learning; knowledge discovery.

1. Developing a Unifying Theory of Data Mining

Several respondents feel that the current state of the art of data mining research is too "ad-hoc." Many techniques are designed for individual problems, such as classification or clustering, but there is no unifying theory. However, a theoretical framework that unifies different data mining tasks including clustering, classification, association rules, etc., as well as different data mining approaches (such as statistics, machine learning, database systems, etc.), would help the field and provide a basis for future research.

There is also an opportunity and need for data mining researchers to solve some longstanding problems in statistical research, such as the age-old problem of avoiding spurious correlations. This is sometimes related to the problem of mining for "deep knowledge," which is the hidden cause for many observations. For example, it was found that in Hong Kong, there is a strong correlation between the timing of TV series by one particular star and the occurrences of small market crashes in Hong Kong. However, to conclude that there is a hidden cause behind the correlation is too rash. Another example is: can we discover Newton's laws from observing the movements of objects?

2. Scaling Up for High Dimensional Data and High Speed Data Streams

One challenge is how to design classifiers to handle ultra-high dimensional classification problems. There is a strong need now to build useful classifiers with hundreds of millions or billions of features, for applications such as text mining and drug safety analysis. Such problems often begin with tens of thousands of features and also with interactions between the features, so the number of implied features gets huge quickly.

One important problem is mining data streams in extremely large databases (e.g. 100 TB). Satellite and computer network data can easily be of this scale. However, today's data mining technology is still too slow to handle data of this scale. In addition, data mining should be a continuous, online process, rather than an occasional one-shot process. Organizations that can do this will have a decisive advantage over ones that do not. Data streams present a new challenge for data mining researchers.

One particular instance is from high speed network traffic where one hopes to mine information for various purposes, including identifying anomalous events possibly indicating attacks of one kind or another. A technical problem is how to compute models over streaming data, which accommodate changing environments from which the data are drawn. This is the problem of "concept drift" or "environment drift." This problem is particularly hard in the context of large streaming data. How may one compute models that are accurate and useful very efficiently? For example, one cannot presume to have a great deal of computing power and resources to store a lot of data, or to pass over the data multiple times. Hence, incremental mining and effective model updating to maintain accurate modeling of the current stream are both very hard problems.

Data streams can also come from sensor networks and RFID applications. In the future, RFIDs will be a huge area, and analysis of this data is crucial to its success.

3. Mining Sequence Data and Time Series Data

Sequential and time series data mining remains an important problem. Despite progress in other related fields, how to efficiently cluster, classify and predict the trends of these data is still an important open topic.

A particularly challenging problem is the noise in time series data. It is an important open issue to tackle. Many time series used for predictions are contaminated by noise, making it difficult to do accurate short-term and long-term predictions. Examples of these applications include the predictions of financial time series and seismic time series. Although signal processing techniques, such as wavelet analysis and filtering, can be applied to remove the noise, they often introduce lags in the filtered data. Such lags reduce the accuracy of predictions because the predictor must overcome the lags before it can predict into the future. Existing data mining methods also have difficulty in handling noisy data and learning meaningful information from the data.

Some of the key issues that need to be addressed in the design of a practical data miner for noisy time series include:

- *Information/search agents to get information*: Use of wrong, too many, or too little search criteria; possibly inconsistent information from many sources; semantic analysis of (meta-) information; assimilation of information into inputs to predictor agents.
- *Learner/miner to modify information selection criteria*: apportioning of biases to feedback; developing rules for Search Agents to collect information; developing rules for Information Agents to assimilate information.
- *Predictor agents to predict trends*: Incorporation of qualitative information; multi-objective optimization not in closed form.

4. Mining Complex Knowledge from Complex Data

One important type of complex knowledge is in the form of graphs. Recent research has touched on the topic of discovering graphs and structured patterns from large data, but clearly, more needs to be done.

Another form of complexity is from data that are non-i.i.d. (independent and identically distributed). This problem can occur when mining data from multiple relations. In most domains, the objects of interest are not independent of each other, and are not of a single type. We need data mining systems that can soundly mine the rich structure of relations among objects, such as interlinked Web pages, social networks, metabolic networks in the cell, etc.

Yet another important problem is how to mine non-relational data. A great majority of most organizations' data is in *text form*, not databases, and in more complex data formats including Image, Multimedia, and Web data. Thus, there is a need to study data mining methods that go beyond classification and clustering. Some interesting questions include how to perform better automatic summarization of text and how to recognize the movement of objects and people from Web and Wireless data logs in order to discover useful spatial and temporal knowledge.

There is now a strong need for integrating data mining and knowledge inference. It is an important future topic. In particular, one important area is to incorporate background knowledge into data mining. The biggest gap between what data mining

systems can do today and what we'd like them to do is that they're unable to relate the results of mining to the real-world decisions they affect — all they can do is hand the results back to the user. Doing these inferences, and thus automating the whole data mining loop, requires representing and using world knowledge within the system. One important application of the integration is to inject domain information and business knowledge into the knowledge discovery process.

Related to mining complex knowledge, the topic of mining *interesting* knowledge remains important. In the past, several researchers have tackled this problem from different angles, but we still do not have a very good understanding of what makes discovered patterns "interesting" from the *end-user* perspective.

5. Data Mining in a Network Setting

5.1. *Community and social networks*

Today's world is interconnected through many types of links. These links include Web pages, blogs, and emails. Many respondents consider community mining and the mining of social networks as important topics. Community structures are important properties of social networks. The identification problem in itself is a challenging one. First, it's critical to have the right characterization of the notion of "community" that is to be detected. Second, the entities/nodes involved are distributed in real-life applications, and hence distributed means of identification will be desired. Third, a snapshot-based dataset may not be able to capture the real picture; what is most important lies in the local relationships (e.g. the nature and frequency of local interactions) between the entities/nodes. Under these circumstances, our challenge is to understand (1) the network's static structures (e.g. topologies and clusters) and (2) dynamic behavior (such as growth factors, robustness, and functional efficiency). A similar challenge exists in bio-informatics, as we are currently moving our attention to the dynamic studies of regulatory networks.

A questions related to this issue is what local algorithms/protocols are necessary in order to detect (or form) communities in a bottom-up fashion (as in the real world).

A concrete question is as follows. Email exchanges within an organization or in one's own mailbox over a long period of time can be mined to show how various networks of common practice or friendship start to emerge. How can we obtain and mine useful knowledge from them?

5.2. *Mining in and for computer networks — high-speed mining of high-speed streams*

Network mining problems pose a key challenge. Network links are increasing in speed, and service providers are now deploying 1 Gig Ethernet and 10 Gig Ethernet link speeds. To be able to detect anomalies (e.g. sudden traffic spikes due to a DoS (Denial of Service) attack or catastrophic event), service providers will need to be

able to capture IP packets at high link speeds and also analyze massive amounts (several hundred GB) of data each day. One will need highly scalable solutions here.

Good algorithms are, therefore, needed to detect whether DoS attacks do not exist. Also, once an attack has been detected, how does one discriminate between legitimate traffic and attack traffic so that it is possible to drop attack packets? We need techniques to

(1) detect DoS attacks,
(2) trace back to find out who the attackers are, and
(3) drop those packets that belong to attack traffic.

6. Distributed Data Mining and Mining Multi-Agent Data

The problem of distributed data mining is very important in network problems. In a distributed environment (such as a sensor or IP network), one has distributed probes placed at strategic locations within the network. The problem here is to be able to correlate the data seen at the various probes, and discover patterns in the global data seen at all the different probes. There could be different models of distributed data mining here, but one could involve a NOC that collects data from the distributed sites, and another in which all sites are treated equally. The goal here obviously would be to minimize the amount of data shipped between the various sites — essentially, to reduce the communication overhead.

In distributed mining, one problem is how to mine across multiple heterogeneous data sources: multi-database and multi-relational mining.

Another important new area is *adversary data mining*. In a growing number of domains — email spam, counter-terrorism, intrusion detection/computer security, click spam, search engine spam, surveillance, fraud detection, shopbots, file sharing, etc. — data mining systems face adversaries that deliberately manipulate the data to sabotage them (e.g. make them produce false negatives). We need to develop systems that explicitly take this into account, by combining data mining with game theory.

7. Data Mining for Biological and Environmental Problems

Many researchers that we surveyed believe that mining biological data continues to be an extremely important problem, both for data mining research and for biomedical sciences. An example of a research issue is how to apply data mining to HIV vaccine design. In molecular biology, many complex data mining tasks exist, which cannot be handled by standard data mining algorithms. These problems involve many different aspects, such as DNA, chemical properties, 3D structures, and functional properties.

There is also a need to go beyond bio-data mining. Data mining researchers should consider ecological and environmental informatics. One of the biggest concerns today, which is going to require significant data mining efforts, is the

question of how we can best understand and hence utilize our natural environment and resources — since the world today is highly "resource-driven"! Data mining will be able to make a high impact in the area of integrated data fusion and mining in ecological/environmental applications, especially when involving distributed/decentralized data sources, e.g. autonomous mobile sensor networks for monitoring climate and/or vegetation changes.

For example, how can data mining technologies be used to study and find out contributing factors in the observed doubling of the number of hurricane occurrences over the past decades, as recently reported in *Science* magazine? Most of the data sources that we are dealing with today are fast evolving, e.g. those from stock markets or city traffic. There is much interesting knowledge yet to be discovered, as far as the dynamic change regularities and/or their cross-interactions are concerned. In this regard, one of the challenges today is how to deal with the problem of dynamic temporal behavioral pattern identification and prediction in: (1) very large-scale systems (e.g. global climate changes and potential "bird flu" epidemics) and (2) human-centered systems (e.g. user-adapted human-computer interaction or P2P transactions).

Related to these questions about important applications, there is a need to focus on "killer applications" of data mining. So far three important and challenging applications for data mining have emerged: bioinformatics, CRM/personalization and security applications. However, more explorations are needed to expand these applications and extend the list of applications.

8. Data Mining Process-Related Problems

Important topics exist in improving data-mining tools and processes through automation, as suggested by several researchers. Specific issues include how to automate the composition of data mining operations and building a methodology into data mining systems to help users avoid many data mining mistakes. If we automate the different data mining process operations, it would be possible to reduce human labor as much as possible. One important issue is how to automate data cleaning. We can build models and find patterns very fast today, but 90 percent of the cost is in pre-processing (data integration, data cleaning, etc.) Reducing this cost will have a much greater payoff than further reducing the cost of model-building and pattern-finding. Another issue is how to perform systematic documentation of data cleaning. Another issue is how to combine visual interactive and automatic data mining techniques together. He observes that in many applications, data mining goals and tasks cannot be fully specified, especially in exploratory data analysis. Visualization helps to learn more about the data and define/refine the data mining tasks.

There is also a need for the development of a theory behind interactive exploration of large/complex datasets. An important question to ask is: what are the compositional approaches for multi-step mining "queries"? What is the canonical

set of data mining operators for the interactive exploration approach? For example, the data mining system *Clementine* has a nice user interface, but what is the theory behind its operations?

9. Security, Privacy, and Data Integrity

Several researchers considered privacy protection in data mining as an important topic. That is, how to ensure the users' privacy while their data are being mined. Related to this topic is data mining for protection of security and privacy. One respondent states that if we do not solve the privacy issue, data mining will become a derogatory term to the general public.

Some respondents consider the problem of knowledge integrity assessment to be important. We quote their observations: "Data mining algorithms are frequently applied to data that have been intentionally modified from their original version, in order to misinform the recipients of the data or to counter privacy and security threats. Such modifications can distort, to an unknown extent, the knowledge contained in the original data. As a result, one of the challenges facing researchers is the development of measures not only to evaluate the knowledge integrity of a collection of data, but also of measures to evaluate the knowledge integrity of individual patterns. Additionally, the problem of knowledge integrity assessment presents several challenges."

Related to the knowledge integrity assessment issue, the two most significant challenges are: (1) develop efficient algorithms for comparing the knowledge contents of the two (before and after) versions of the data, and (2) develop algorithms for estimating the impact that certain modifications of the data have on the statistical significance of individual patterns obtainable by broad classes of data mining algorithms. The first challenge requires the development of efficient algorithms and data structures to evaluate the knowledge integrity of a collection of data. The second challenge is to develop algorithms to measure the impact that the modification of data values has on a discovered pattern's statistical significance, although it might be infeasible to develop a global measure for all data mining algorithms.

10. Dealing with Non-Static, Unbalanced and Cost-Sensitive Data

An important issue is that the learned models should incorporate time because data is not static and is constantly changing in many domains. Historical actions in sampling and model building are not optimal, but they are not chosen randomly either. This gives the following challenging phenomenon for the data collection process. Suppose that we use the data collected in 2000 to learn a model. We then apply this model to select inside the 2001 population. Subsequently, we use the data about the individuals selected in 2001 to learn a new model, and then apply this model in 2002. If this process continues, then each time a new model is learned, its training set has been created using a different selection bias. Thus, a challenging problem is how to correct the bias as much as possible.

Another related issue is how to deal with unbalanced and cost-sensitive data, a major challenge in research. Charles Elkan made the observation in an invited talk at *ICML 2003 Workshop on Learning from Imbalanced Data Sets*. First, in previous studies, it has been observed that UCI datasets are small and not highly unbalanced. In a typical real-world dataset, there are at least 10^5 examples and $10^{2.5}$ features, without single well-defined target class. Interesting cases have a frequency of less than 0.01. There is much information on costs and benefits, but no overall model of profit and loss. There are different cost matrices for different examples. However, most cost matrix entries are unknown. An example of this dataset is the direct marketing DMEF data library. Furthermore, the costs of different outcomes are dependent on the examples; for example, the false negative cost of direct marketing is directly proportional to the amount of a potential donation. Traditional methods for obtaining these costs relied on sampling methods. However, sampling methods can easily give biased results.

11. Conclusions

Since its conception in the late 1980s, data mining has achieved tremendous success. Many new problems have emerged and have been solved by data mining researchers. However, there is still a lack of timely exchange of important topics in the community as a whole. This article summarizes a survey that we have conducted to rank 10 most important problems in data mining research. These problems are sampled from a small, albeit important, segment of the community. The list should obviously be a function of time for this dynamic field.

Finally, we summarize the 10 problems below:

- Developing a unifying theory of data mining
- Scaling up for high dimensional data and high speed data streams
- Mining sequence data and time series data
- Mining complex knowledge from complex data
- Data mining in a network setting
- Distributed data mining and mining multi-agent data
- Data mining for biological and environmental problems
- Data Mining process-related problems
- Security, privacy and data integrity
- Dealing with non-static, unbalanced and cost-sensitive data

Acknowledgments

We thank all who have responded to our survey requests despite their busy schedules. We wish to thank Pedro Domingos, Charles Elkan, Johannes Gehrke, Jiawei Han, David Heckerman, Daniel Keim, Jiming Liu, David Madigan, Gregory Piatetsky-Shapiro, Vijay V. Raghavan, and his associates, Rajeev Rastogi, Salvatore J. Stolfo, Alexander Tuzhilin, and Benjamin W. Wah for their kind input.