ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 1, 2025

TOP 5 CHALLENGING PROBLEMS IN DATA MINING RESEARCH

*V.Chandraprakash [1], Dr. P V Kumar [2], V Pavani [3], K Anjaneulu [4]
[1] Assistant Professor, St. Martin's Engineering College, Secunderabad, Telangana – 500100
[2]Professor, Department of CSE,Osmania Univesrsity ,Hyderabad, Telangana – 500007
[3]Assistant Professor, TKR Engineering College,Hyderabad , Telangana – 500097
[4] Assistant Professor, Department of CSE, KMIT College ,Hyderabad, Telangana – 500029
*Corresponding Author
Email: vchandraprakashit@smec.ac.in

Abstract-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, classifica- tion, association rules, etc., as well as different data mining approaches (such as statistics, machine learning, database systems, etc.), would help the field and pro- vide a basis for future research.

There is also an opportunity and need for data mining researchers to solve some long standing problems in statistical research, such as the age-old problem of avoid- ing 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?

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 1, 2025

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

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 1, 2025

search criteria; possibly inconsistent information from many sources; seman- tic 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. 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. 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 commu- nity 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

REFERENCES

[1] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, “Understanding individual human mobility patterns,” Nature, vol. 453, no. 7196, pp. 779–782, 2008.
[2] A. Monreale, F. Pinelli, R. Trasarti, and F. Giannotti, “Wherenext: a location predictor on trajectory pattern mining,” in Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009, pp. 637–646.
[3] M. Morzy, “Mining frequent trajectories of moving objects for location prediction,” Machine Learning and Data Mining in Pattern Recognition, pp. 667–680, 2007.

[4] A. Noulas, S. Scellato, N. Lathia, and C. Mascolo, "Mining user mobility features for next place prediction in location-based services," in Data mining (ICDM), 2012 IEEE 12th international conference on. IEEE, 2012, pp. 1038–1043.
[5] H. Gao, J. Tang, and H. Liu, "Mobile location prediction in spatiotemporal context," in Nokia mobile data challenge workshop, vol. 41, no. 2, 2012, pp. 1–4.
[6] P. Baumann, W. Kleiminger, and S. Santini, "The influence of temporal and spatial features on the performance of next-place prediction algorithms," in Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing. ACM, 2013, pp. 449–458.
[7] J. Ye, Z. Zhu, and H. Cheng, "What's your next move: User activity prediction in location-based social networks," in Proceedings of the 2013 SIAM International Conference on Data Mining. SIAM, 2013, pp. 171–179.