Online Mining of Changes from Data Streams: Research Problems and Preliminary Results

Guozhu Dong[1] Jiawei Han[2] Laks V.S. Lakshmanan[3]
Jian Pei[4] Haixun Wang[5] Philip S. Yu[5]

[1] Wright State University, gdong@cs.wright.edu
[2] University of Illinois at Urbana-Champaign, hanj@cs.uiuc.edu
[3] University of British Columbia, Canada, laks@cs.ubc.ca
[4] State University of New York at Buffalo, jianpei@cse.buffalo.edu (Contact author)
[5] IBM T.J. Watson Research Center, {haixun, psyu}@us.ibm.com

ABSTRACT

As data streams are gaining prominence in a growing number of emerging applications, advanced analysis and mining of data streams is becoming increasingly important. While there are some recent studies on mining data streams, we would like to ask the following essential question: *What are the distinct features of mining data streams compared to mining other kinds of data?* In this paper, we take the following position: *online mining of the* **changes** *in data streams is one of the core issues.* We propose some interesting research problems and highlight the inherent challenges. Moreover, we sketch some preliminary results.

1. INTRODUCTION

Recent research indicates that a growing number of emerging applications, such as sensor networks, networking flow analysis, and e-business and stock market online analysis, have to handle various data streams. It is demanding to conduct advanced analysis and data mining over fast and large data streams to capture the trends, patterns, and exceptions. Recently, some interesting results have been reported for modelling and handling data streams (see [1] for a comprehensive overview), such as monitoring statistics over streams and query answering (e.g., [3, 9, 4])). Furthermore, conventional OLAP and data mining models have been extended to tackle data streams, such as multi-dimensional analysis (e.g., [2]), clustering (e.g., [10]) and classification (e.g., [5, 11]).

While extending the existing data mining models to tackle data streams may provide valuable insights into the streaming data, it is high time we considered the following fundamental question: *Compared to the previous studies on mining various kinds of data, what are the distinct features/core problems of mining data streams?* In other words, *from mining data streams, do we expect something different than mining other kinds of data?*

Previous studies (e.g., [1, 8]) argue that mining data streams is challenging in the following two respects. On the one hand, random access to fast and large data streams may be impossible. Thus, multi-pass algorithms (i.e., ones that load data items into main memory multiple times) are often infeasible. On the other hand, the exact answers from data streams are often too expensive. Thus, approximate answers are acceptable. While the above two issues are critical, they are not unique to data streams. For example, online mining very large databases also requires ideally one-pass algorithms and may also accept approximations.

We argue that one of the keys to mining data streams is *online mining of changes*. For example, consider a stream of regular updates of various aircrafts' positions. An air traffic controller may be interested in the clusters of the aircrafts at each moment. However, instead of checking details for "normal" clusters, she/he may be more interested in those "abnormal" clusters, e.g., fast growing clusters indicating the forming of a traffic jam. In general, while the patterns in snapshots of data streams are important and interesting, the *changes to the patterns* may be more critical and informative. With data streams, people are often interested in mining queries like "*compared to the history, what are the distinct features of the current status?*" and "*what are the relatively stable factors over time?*" Clearly, to answer the above queries, we have to examine the changes.

Some previous works also involve change detection. For example, the emerging patterns [6] characterize the changes from one data set to the other. In [7], Ganti et al. propose methods to measure the differences of the induced models in data sets. In [5, 11], the classification of time-changing data streams is studied. The goal in the above studies is to incrementally maintain a global classifier by incorporating the changes. The description of the changes has not been studied substantially. Incremental mining studies how to

ACM SIGMOD MPDS '03 San Diego, CA, USA

update the models/patterns by factoring in the incremental part of data. However, mining data streams often requires online and dynamic detection and summarization of interesting changes.

2. PROBLEMS AND CHALLENGES

Interesting research problems on mining changes in data streams can be divided into three categories: modelling and representation of changes, mining methods, and interactive exploration of changes.

Modelling and Representation of Changes

While the term "changes" sounds general and intuitive, it is far from trivial to define and describe changes in data streams. First, it is essential to propose concise query language constructs for describing the mining queries on changes in data streams. There can be many kinds of changes in data streams, and different users may be interested in different kinds. The user should be able to specify the changes she/he wants to see. Moreover, the system should be able to rank changes based on interestingness. The operations should be integrable into the existing data mining models and languages. An "algebra" for change mining may be essential. Second, methods of summarizing and representing the changes need to be developed. In particular, effective visualization of changes is very important. Third, while the model for mining "first order" changes is common and useful, the model for mining "higher order" changes can be an important kind of knowledge in some dynamic environments. For example, a stock market analyst may feel particularly interested in the changes in the ranges of price vibration, while the range of price vibration itself is a description of changes.

Mining Methods

Efficient and scalable algorithms are needed for mining changes in data streams, at various levels. First, specific algorithms can be developed for specific change mining queries. While such query-specific approaches may not be systematic, it will provide valuable insights into the inherent properties, challenges and basic methods of change mining. Second, general evaluation methods for "change mining queries" should be developed based on the general model/query language/algebra. Third, facilities for various aspects of change mining, such as quality management, should be considered. For example, algorithms should be able to meet user's specification on level/granularities/approximation error bound of change mining.

Interactive Exploration of Changes

The results from change mining per se form data streams, which can sometimes be large and fast. It is important to develop effective approaches to support user's interactive exploration of the changes. For example, a user may want to monitor the changes at an appropriate level. Once some interesting changes are detected, she/he can closely inspect the related details.

To the best of our knowledge, the above problems have not been researched systematically so far. By no means is the above list complete. We believe that thorough studies on these issues will bring about many challenges, opportunities, and benefits to stream data processing, management, and analysis.

3. SOME PRELIMINARY RESULTS

In this section, we report two of our preliminary results on mining changes in data streams.

One fundamental problem in classifying streams with conceptual changes is how to identify in a timely manner those data in the training set that are no longer consistent with the current concepts. The impact of these data sets on the model must be excluded. A straightforward solution, which is adopted by many current approaches, discards data indiscriminately after they become old, that is, after a fixed period of time T has passed since their arrival (or a window of fixed size has been filled) [11]. While this solution is conceptually simple, the model learned from the training data, however, may fail to represent the most-up-to-date concepts because the learning method does not monitor the changes directly. It creates the following dilemma which makes it vulnerable to unpredictable conceptual changes in the data: if T is large, the training set is likely to contain outdated concepts, which reduces classification accuracy; if T is small, the training set may not have enough data, and as a result, the learned model will likely carry a large variance due to overfitting.

In light of these challenges, in [13], we propose using *weighted classifier ensembles* to mine streaming data with concept changes. Instead of continuously revising a single model, we train an ensemble of classifiers from sequential data chunks in the stream. Maintaining a most up-to-date classifier is not necessarily the ideal choice, because potentially valuable information may be wasted by discarding results of previously-trained less-accurate classifiers. We show that, in order to avoid overfitting and the problems of conflicting concepts, the expiration of old data must rely on data's distribution instead of only their arrival time. The ensemble approach offers this capability by giving each classifier a weight based on its expected prediction accuracy on the current test examples. Thus, the changes are monitored directly, and reflected into the learned model in a timely manner. Our results show that the weighted ensemble approach outperforms the single-model approach in many aspects, including prediction accuracy, model construction efficiency, and ease-of-use.

As another example, while clustering data streams provides interesting information for the analysis of streams, the *changes in the clusters over time* are often even more interesting and critical in many applications. In [12], we propose *online mining of changes of clusters in a data stream*. For example, suppose the positions of the objects at instants 1 and 2 are as shown in Figure 1. We mine the summary of changes of clusters, e.g., the cluster at instant 1 containing objects a, e, j, o, s, w, i, and h stays loosely as a cluster

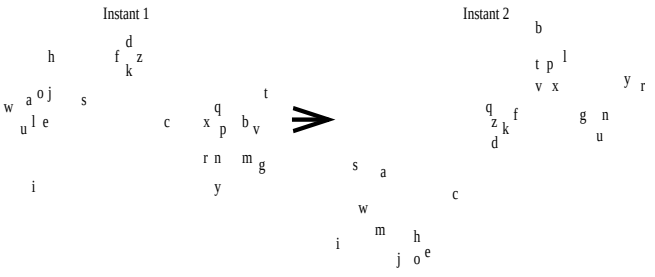

Figure 1: An example of the changes of clusters.

at instant 2 with lower density, the two clusters at instant 1, $\{d, f, k, z\}$ and $\{p, q, x, n, r, y, b, t, v, g\}$, are merging into each other at instant 2. The merged cluster at instant 2 is denser than the corresponding parts at instant 1.

We proposed two techniques to tackle this problem. (i) We developed a *density-list* notation to record the structures of clusters in the stream at any arbitrary instant. A fast and space preserving algorithm is devised to compute the density-list online from the streams. (ii) We proposed an interesting visualization technique, *density-list graph*, which can visualize online both the structures of clusters as well as their changes over time in a data stream.

4. CONCLUSIONS

In this paper, we take the position that online mining of the changes in data streams is one of the core issues of mining data streams. We identify some interesting research problems and highlight some inherent challenges. Currently, we are studying some or the problems.

5. REFERENCES

[1] B. Babcock, S. Babu, M. Datar, R. Motwani, and J. Widom. Models and issues in data stream systems. In *PODS'02*, Madison, WI, June 2002.

[2] Y. Chen, G. Dong, J. Han, B. W. Wah, and J. Wang. Multi-dimensional regression analysis of time-series data streams. In *VLDB'02*, Hong Kong, China, Aug. 2002.

[3] M. Datar, A. Gionis, P. Indyk, and R. Motwani. Maintaining stream statistics over sliding windows (extended abstract). *citeseer.nj.nec.com/491746.html*.

[4] A. Dobra, M. Garofalakis, J. Gehrke, and R. Rastogi. Processing complex aggregate queries over data streams. In *SIGMOD'02*, Madison, Wisconsin, June 2002.

[5] P. Domingos and G. Hulten. Mining high-speed data streams. In *KDD'00*, pages 71–80, Boston, MA, Aug. 2000.

[6] G. Dong and J. Li. Efficient mining of emerging patterns: Discovering trends and differences. In *KDD'99*, pages 43–52, San Diego, CA, Aug. 1999.

[7] V. Ganti, J. Gehrke, and R. Ramakrishnan. A framework for measuring changes in data characteristics. In *PODS'99*, pages 126–137, Philadelphia, PA, May/June 1999.

[8] M. Garofalakis, J. Gehrke, and R. Rastogi. Querying and mining data streams: You only get one look. In *VLDB'02*, Hong Kong, China, Aug. 2002.

[9] J. Gehrke, F. Korn, and D. Srivastava. On computing correlated aggregates over continuous data streams. In *SIGMOD'01*, pages 13–24, Santa Barbara, CA, May 2001.

[10] S. Guha, N. Mishra, R. Motwani, and L. O'Callaghan. Clustering data streams. In *FOCS'00*, pages 359–366, Redondo Beach, CA, 2000.

[11] G. Hulten, L. Spencer, and P. Domingos. Mining time-changing data streams. In *KDD'01*, San Francisco, CA, Aug. 2001.

[12] J. Pei, S.R. Ariwala, and D. Jiang. Online mining changes of clusters in data streams. In *Submitted for publication*.

[13] H. Wang, W. Fan, P.S. Yu, and J. Han. Mining concept-drifting data streams using ensemble classifiers. In *Submitted for publication*.