Incremental Pattern Discovery on Streams, Graphs and Tensors

Jimeng Sun
Carnegie Mellon University
jimeng@cs.cmu.edu

ABSTRACT

Incremental pattern discovery targets streaming applications where the data continuously arrive incrementally. The questions are how to find patterns (main trends) incrementally; or how to efficiently update the old patterns when new data arrive; or how to utilize the patterns to solve other problems such as anomaly detection?

We first investigate a powerful data model, *tensor stream* (TS), where there is one tensor per timestamp. To capture diverse data formats, we have a zero-order TS for a single time-series (e.g., the stock price over time), a first-order TS for multiple time-series (sensor measurement streams), a second-order TS for matrices (graphs), and a high-order TS for multi-arrays (Internet communication network, source-destination-port). Second, we develop different online algorithms on TS: 1) the centralized and distributed SPIRIT [7] for mining a *1st-order* TS, as well as its extensions for local correlation function and privacy preservation; 2) the compact matrix decomposition (CMD) [5] and GraphScope [4] for a *2nd-order* TS; 3) the dynamic tensor analysis (DTA) [2], streaming tensor analysis (STA) and window-based tensor analysis (WTA) for a *high-order* TS. All the techniques are extensively evaluated for real applications such as network forensics, cluster monitoring.

Keywords

Data mining, Stream mining, Graph mining, Tensor analysis.

1. MOTIVATIONS

Incremental pattern discovery targets at streaming applications where data continuously arrive incrementally. In this thesis, we want to answer the following questions: How to find patterns (main trends) incrementally? How to efficiently update the old patterns when new data arrive? How to utilize the patterns to solve other problems such as anomaly detection and clustering?

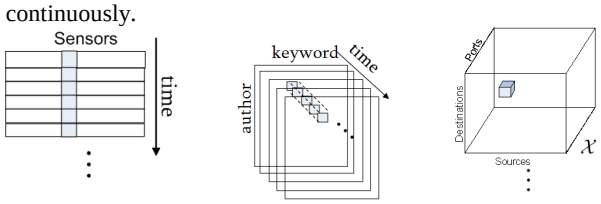
Some examples include:

- Sensor Networks [3] monitor different measurements (such as temperature and humidity) from a large number of distributed sensors. The task is to monitor correlations among different sensors over time and identify anomalies.
- Cluster Management **Error! Reference source not found.** monitors the many metrics (such as CPU and memory utilization, disk space, number of processes, etc) of a group of machines. The task is to find main trends, to identify anomalies or potential failures as well as to compress the signals for storage.
- Social Network Analysis [2] observes an evolving network of social activities (such as citation, telecommunication and corporate email networks). The task is to find communities and anomalies, and to monitor them over time.
- Network Forensics [2] monitors Internet communication in the form of source, destination, port, time, number of packets, etc. Again, the task is to summarize the communication patterns and to identify the potential attacks and anomalies.
- Financial Fraud Detection [8] examines transactional activities of a company over time and tries to identify the abnormal/fraudulent behaviors.

2. DATA MODEL

To deal with the diversity of data, we introduce an expressive data model tensor from multi-linear analysis [6]. For the Sensor Networks example, we have one measurement (e.g., temperature) from each sensor every timestamp, which forms a high dimensional vector (first order tensor) as shown in Figure 1(a). For the Social Network Analysis, we have authors publishing papers, which forms graphs represented by matrices (second order tensors). For the network forensics example, the 3rd order tensor for a given time period has three modes: source, destination and port, which can be viewed as a 3D data cube (see Figure 1(c)). An entry (*i, j, k*) in that tensor (like the small blue cube in Figure 1 (c)) has the number of packets from the corresponding source *i* to the destination *j* through port *k*, during the given time period. Figure 1 illustrates three tensor examples where the blue region indicates a single element in the tensor such as a measurement from a single sensor in (a), the number of papers that an author wrote on a given keyword in (b), the number of packets sent from a source IP to a destination IP through a certain port in (c).

Focusing on incremental applications, we propose the tensor stream (TS) which is an unbounded sequence of tensors. The streaming aspect comes from the fact that new tensors are arriving continuously.



(a) 1st order (b) 2nd order (c) 3rd order

Figure 1: Tensor Stream examples

3. INCREMENTAL PATTERN DISCOVERY

Incremental Pattern discovery is an online summarization process. In this thesis, we focus on incrementally identifying low-rank structures of the data as the patterns and monitor them over time. In another words, we consider the incremental pattern discovery as an incremental dimensionality reduction process. In this sense, the following terms are interchangeable: summaries, patterns, hidden variables, low-dimensional representations, core tensors.

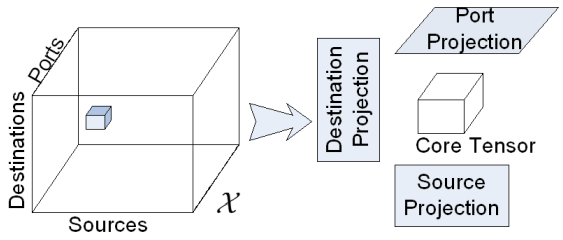


Figure 2: Pattern discovery on a 3rd order tensor

Let us illustrate the main idea through the network forensics application. In this example, the hourly communication data are represented by high dimensional (3rd order) tensors, which are summarized as low dimensional (3rd order) core tensors in a different space specified by the projection matrices (see Figure 2). Moreover, the projection matrices capture the overall correlations or hidden variables along three aspects (modes): source, destination and port. For example, the Source projection characterizes the client correlations; the Destination projection summarizes the server correlations; the Port projection monitors the port traffic correlations. The projection matrices are dynamically monitored over time. The notation of projection matrices can be very general. In particular, we explore three different subspace construction strategies:

1. *orthonormal projection*, which forms orthogonal matrices based on the data such as SPIRIT [7][3] and Dynamic and Streaming Tensor Analysis (DTA/STA) [2];
2. *example-based projection*, which judiciously select examples from data to form the subspace [120];
3. *clustering based projection*, which consists of indicator variable vectors based on cluster assignments [5].

The incremental aspect of the algorithms arrives from the fact that model needs to be constantly updated. More specifically, the problems we study are as the follows: Given a stream of tensors, how to compress them incrementally and efficiently? How to find patterns and anomalies? We plan to address two aspects of incremental pattern discovery:

- *Incremental update*: We want to update the old model efficiently, when a new tensor arrives. The key is to avoid redundant computation and storage.
- *Model efficiency*: We want an efficient method in terms of computational cost and storage consumption. The goal is to achieve linear computation and storage to the update size.

The results from incremental pattern discovery can be used for many important tasks:

Compression: The core tensor captures most of the information of the original data but in a much lower dimension. For example, our InteMon system [1] can achieve 10-100X compression gain on performance sensor monitoring data of a data center.

Anomaly detection: From the core tensor, we can approximate the original tensor and compute the reconstruction error. A large reconstruction error often indicates an anomaly. Our CMD [5] and DTA [2] can successfully detect abnormal network flows.

Clustering: We can often cluster the original data based the projection. Our GraphScope [4] successfully identifies interesting groups and time segments on several large social networks such as Enron email dataset.

4. CONCLUSION

Thesis statement *Incremental and efficient summarization of streaming data through a general and concise presentation enables many real applications in diverse domains*.

Incremental Pattern Discovery provides a unified view of several fundamental problems from the data mining perspective. A set of related tools are presented all embodying the same approach, i.e., the pursuit of efficient and effective algorithms for analyzing data streams automatically. Its success on diverse applications has confirmed the significance of all the algorithms. In this thesis, we addressed the problem from a data mining perspective, which emphasizes an algorithmic perspective and also provides a few system prototypes. In the future, all these techniques can be leveraged with a more system-oriented approach, such as deployment to de-centralized clusters and integration with existing monitoring systems.

5. REFERENCES

[1] Evan Hoke, Jimeng Sun, John D. Strunk, Gregory R. Ganger, and Christos Faloutsos. Intemon: Continuous mining of sensor data in large-scale self-* infrastructures. ACM SIGOPS Operating Systems Review, 40(3), 2003.

[2] Jimeng Sun, Dacheng Tao, and Christos Faloutsos. Beyond streams and graphs: Dynamic tensor analysis. In Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD), 2006.

[3] Jimeng Sun, Spiros Papadimitriou, and Christos Faloutsos. Distributed pattern discovery in multiple streams. In Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), 2006.

[4] Jimeng Sun, Spiros Papadimitriou, Christos Faloutsos, Philip S. Yu: GraphScope: parameter-free mining of large time-evolving graphs. KDD 2007: 687-696

[5] Jimeng Sun, Yinglian Xie, Hui Zhang, and Christos Faloutsos. Less is more: Compact matrix decomposition for large sparse graphs. In SDM, 2007.

[6] Lieven De Lathauwer. Signal Processing Based on Multilinear Algebra. PhD thesis, Katholieke, University of Leuven, Belgium, 1997.

[7] Spiros Papadimitriou, Jimeng Sun, and Christos Faloutsos. Streaming pattern discovery in multiple time-series. In VLDB, pages 697–708, 2005.

[8] Stephen Bay, Krishna Kumaraswamy, Markus G. Anderle, Rohit Kumar, and David M. Steier. Large scale detection of irregularities in accounting data. In ICDM, pages 75–86, 2006.

About the authors:

Dr. Jimeng Sun graduated from CMU on 2007 and now is a researcher at IBM TJ Watson Research Center. His research focuses on large-scale data mining in high dimensional data such as time series, data cubes and social networks. His thesis advisor is Prof. Christos Faloutsos at CMU. His PhD committee consists of Prof. Christos Faloutsos, Prof. Tom Mitchell, Prof. Hui Zhang, Dr. David Steier and Prof. Philip S. Yu.