Streaming Manifold Learning and DOE Applications

Shinjae Yoo (BNL), Hao Huang (GE), Hayan Lee (LBL)
Yangang Liu (BNL), Dmitri Zakharov (BNL), Eric Stach (BNL)

Motivation

Analyzing high velocity and large volume data streams has been one of critical challenges in various fields including business analytics, defense, energy, health, environment and science. Due to continuous improvement of sensing technology, better computing power and bigger data business models, we are in pressing need for streaming analytics than at any given time in the history. To cope with such high velocity data stream analytics challenges, we are going to focus on significantly improving streaming algorithms which read only once with limited available memory constraints. For basic statistics or supervised learning tasks, there are solid prior works such as reservoir sampling, Bloom filter, itemset mining, online adaptive learnings, margin-infused relaxed algorithm (MIRA), etc. At the same time, for unsupervised learning, there is a clear performance gap between batch algorithms and streaming algorithms. In particular, prior works have difficulties in handling high dimensional data and non-linear relationship. On the other hand, manifold learning has proven to outperform on learning high dimensional and complex data in an unsupervised way. However it demonstrates low efficiency when applied to large volume of data. Therefore, we propose to approximate manifold learning algorithms such as MCFS (Multi-cluster Feature Selection) [1] and Spectral Clustering (SC) [5] into streaming environment.

Challenges

Recognizing actionable patterns from huge volumes of high-dimensional data is a critical topic. Manifold learning is an effective technique designed to tackle this problem through advanced dimensionality reduction, and it has been widely used in a lot of the modern applications. However, approximating batch manifold learning algorithm in streaming environment poses daunting challenges. First, these manifold learning algorithms used to require building pairwise affinity matrix (a.k.a. similarity matrix), which is $O(n^2)$ space complexity. Second, such pairwise affinity may use non-linear kernel, which is hard to approximate linearly on data streams. Third, the normalization of such affinity matrix such as random walk normalization, symmetric normalization, Laplace normalization, etc. is inherently batch process. Fourth, eigenvalue decomposition (EVD) is $O(n^3)$ time complexity which is not scalable for large data streams. Finally, the streaming manifold learning should be adaptable to any change of the underlying data distribution over the stream.

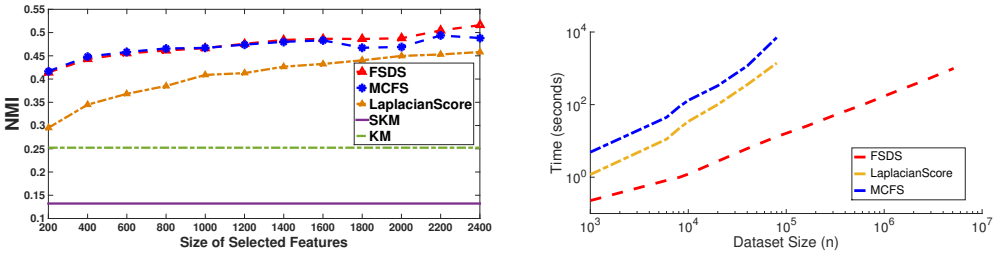
Our Approach

To overcome the above challenges, we extended frequent direction approach [4] and proposed streaming approximation of Graph Laplacian Embeddings [3][7]. The basic idea is the same to itemset mining but extended to matrix. We also adopted chunk of the stream to be input instead of one data point at a time. Laplacian normalization, which is a normalization with node degree, is one of the bottleneck in the streaming environment, since the degree distribution is continuously changed as new streams arrive. We derived streaming approximation of symmetric Laplacian normalization by maintaining the normalized dataset centroid instead of the exact degree distribution. Also for non-linear kernel approximation, we adopted Gaussian kernel linear approximation method, to approach nonlinear pattern in the dataset without sacrificing the efficiency.

We propose streaming feature selection method (FSDS) [3] based on such streaming manifold learning techniques, and maintained 98% accuracy (NMI, or normalized mutual information) of the state-of-the-art batch algorithms (MCFS, or Multi Cluster Feature Selection) [1] or Laplacian Score [2] (Fig 1a). More importantly, it achieves two or more order of speed up compared with the batch algorithm that have inherent scaling issues in larger benchmark dataset, while our proposed algorithm can be applicable to infinite data stream, as shown in Fig 1b. We also utilize this manifold learning to design a streaming spectral clustering (SSC) [7] as shown in Fig 2a and 2b. SSC achieved almost twice better accuracy (NMI) than other state-of-the-art streaming clustering algorithms such as BIRCH [8] or HDDStr [6] but demonstrated the similar scaling performance against the other state-of-the-art streaming clustering algorithms, which are quite encouraging results.

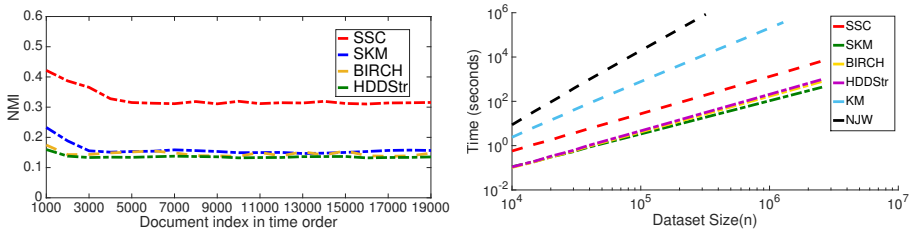
On-going work

We are working on further improvement of manifold learning algorithms and other type of learning algorithms



(a) Feature Selection on 20Newsgroup (b) Feature Selection Scalability

Figure 1: On 20 news group dataset, our proposed FSDS shows similar accuracy with state-of-the-art MCFS but shows strong scalibility. Feature selection result shows better results than both streaming and batch K-means results.



(a) Clustering on 20Newsgroup (b) Clustering Scalability

Figure 2: On 20 news group dataset, our proposed SSC shows almost two times better resultsthan baselines and shows strong scalibility.

in streaming setting to address current DOEs data challenges. We are working on the following three major DOE application areas at the moment.

- For material science, TEM (Transmission Electron Microscopy) at CFN (Center for Functional Nanomaterials) generates 3GB/s raw video streams (up to 1600 frames / sec) that are very challenging and time consuming to manage and analyse. To address these challenges we are going to apply streaming manifold learning (dimensionality reduction algorithm) that will allow us to handle and process high velocity data streams in a manageable fashion. For instance, a detection of a material morphology and structural changes over video data stream would be much easier on such manifold space than the original video stream.

- For climate science, LES (Large Eddy Simulation)-DNS(Direct Numerical Simulation), one of Exascale problem, will generate large scale of simulation output data stream. We want to analyze the simulation output on the fly to avoid storing whole intermediate output files. Analyzing such simulation on the fly can also steer simulation (parameterization) toward the interesting and meaningful simulation study.

- For biology application, clustering analysis of meta-genomics data can be applicable to various levels from assembly quality improvement to abundance profile analysis. The key challenges are due to the scale of raw data (i.e. 1TB). An intermediate analysis may generate much larger scale of data, and thus streaming analytics is a viable choice to pursue high quality analysis.

References

[1] D. Cai, C. Zhang, and X. He. Unsupervised feature selection for multi-cluster data. *SIGKDD 2010*.

[2] X. He, D. Cai, and P. Niyogi. Laplacian score for feature selection. *NIPS 2005*.

[3] H. Huang, S. Yoo, and S. P. Kasiviswanathan. Unsupervised feature selection on data streams. *CIKM 2015*.

[4] E. Liberty. Simple and deterministic matrix sketching. In *SIGKDD*, pages 581–588, 2013.

[5] A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems*, 14:846–856, 2002.

[6] I. Ntoutsi, A. Zimek, T. Palpanas, P. Kröger, and H.-P. Kriegel. Density-based projected clustering over high dimensional data streams. In *SDM*, 2012.

[7] S. Yoo, H. Huang, and S. P. Kasiviswanathan. Streaming spectral clustering. *ICDE 2016 (accepted)*.

[8] T. Zhang, R. Ramakrishnan, and M. Livny. Birch: an efficient data clustering method for very large databases. In *ACM SIGMOD Record*, volume 25, pages 103–114. ACM, 1996.