High-Dimensional Data Stream Visualization

Chanon Jenakom

8 April, 2019

1 Introduction

In recent times, there are many sources of data that can be used for performing data analysis on. Most of these data contain so many features and attributes that contribute to the meaning of each piece of data, such as churn detection in microblogs [1] and financial analysis [2]. Moreover, sources of these kinds of data are usually from social media sites such as Twitter, which is also provided as streams. Streaming data introduces more constraining characteristics to the problem: dynamic, continuous and unbounded [3].

Many high-dimensional data visualization methods have been proposed over the years to extract meaningful patterns from the data [2]. One of the earliest methods used were various types of graphs [4]. For an example, a matrix of scatter plots can be used to visualize the pairs of attributes within the data. Another technique used is a popular dimensionality reduction technique called Principal Component Analysis (PCA) [5]. While effective in creating a subset of independent features that represent the original data, this method is not capable of consistently capturing both the local and global structures of the data [6]. However, another technique called t-Distributed Stochastic Neighbor Embedding, or t-SNE, is capable of capturing various structures of the high-dimensional data considerably well.

Despite of the fact that there are many visualizing algorithms proposed in the field, only a selected few are specifically designed to visualize streaming data. Most are modified versions of pre-existing techniques adapted for streaming environments, such as a suggestion provided in [7] or GPU-accelerated version of t-SNE [8], or a new PCA implementation, named History PCA [9].

In this paper, each high-dimensional data visualization techniques are reviewed and compared in details. Each section will explore the performance of each technique on MNIST Digits dataset, and its possibility in being used in streaming contexted is discussed.

2 Algorithms

In this section, the brief history of data visualization in high dimension is discussed. Moreover, each of the techniques that is effective and popular is reviewed in terms of its mathematical and statistical foundation and its visualization quality, using the digits dataset as shown in Figure 1.

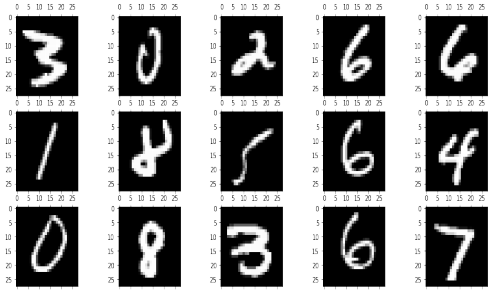


Figure 1: Examples of MNIST Digits

2.1 History

In order to perform visualization on high-dimensional data, there have been many techniques and algorithms proposed. Ranging from simple graphs such as

scatter plots or heatmaps [4] to more complex methods that rely on statistical theories such as PCA [5] and t-SNE [6]. These varying techniques offer different use cases and strengths and weaknesses when presented with different sets of data.

2.2 Principal Component Analysis

Principal Component Analysis is a statistical procedure that convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components [5]. While used mostly as a tool for making predictive models, it has also been used to visualize genetic distance and relatedness between data points.

PCA can be thought of as fitting a p-dimensional ellipsoid to the data, where each of the axis of the ellipsoid represents a principal component, as show in Figure 2. The size of the axis corresponds to the variance along that axis, which signifies to contribution the component has to the overall meaning of the dataset.

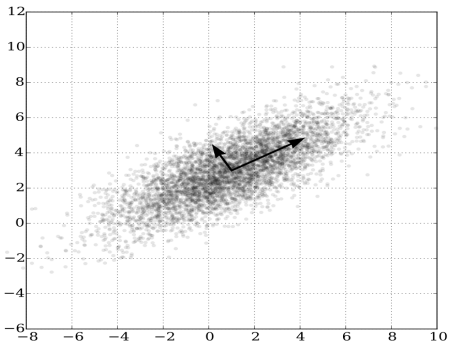


Figure 2: Principal Component Analysis on 2D Data

The first principal component is very important to this algorithm, whose variance must be maximized. Thus, the first weight vector has tp satisfy the following formula:

$$w_{(1)} = \text{argmax} \frac{w^T X^T X w}{w^T w} \tag{1}$$

and every further k^{th}-component has to satisfy the following formula:

$$\hat{X}_k = X - \sum_{s=1}^{k-1} X w_{(s)} w_{(s)}^T \tag{2}$$

When used on MNIST digits dataset, as seen in Figure 3, it can be seen that the separation between digits are not apparent as expected. There is no distinct separation between digits, and most are even placed on the same space within the xy-plane.

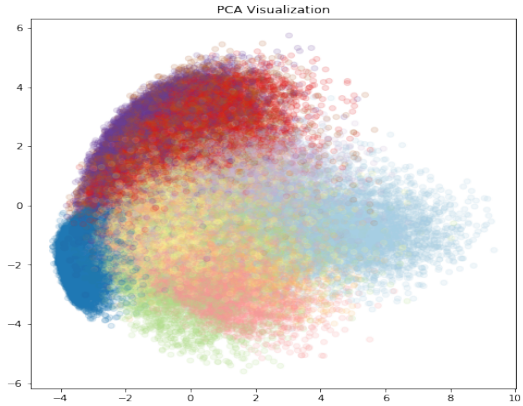


Figure 3: PCA on MNIST Digits Dataset

2.3 t-Distrubuted Stochastic Neighbor Embedding

t-Distributed Stochastic Neighbor Embedding, or t-SNE, uses the concept of probability to reduce the dimensions of high-dimensional data and is able to construct a reasonably good visualizations [6]. The main idea of t-SNE is to map distances in high-dimensional space to probabilities, then minimize the mappings of these probabilities in low-dimensional space, following the cost function:

$$C = KLP(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{3}$$

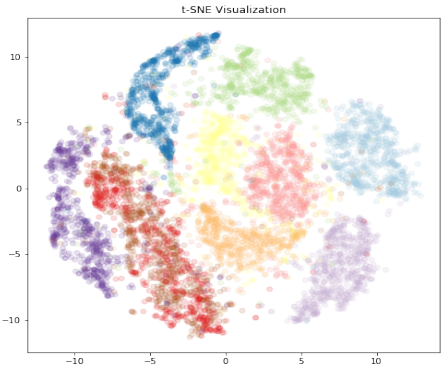
which is the sum of Kullback-Leibler divergences [10]. The algorithm is shown here:

Algorithm 1 Simple version of t-Distributed Stochastic Neighbor Embedding

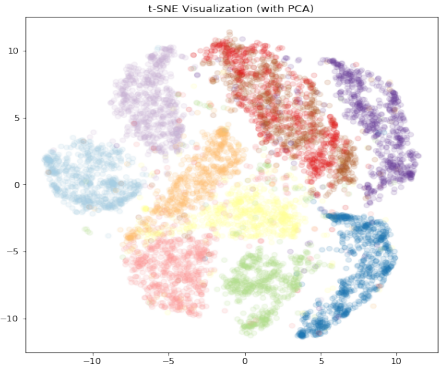
Data: data set $X = \{x_1, x_2, ..., x_n\}$,
Result: low-dimensional data representation $\gamma^{(T)} = \{y_1, y_2, ..., y_n\}$
begin

$p_{ij} \leftarrow \frac{p_{j|i} + p_{i|j}}{2n}$
$\gamma^{(0)} \leftarrow \{y_1, y_2, ..., y_n\}$
for $t \leftarrow 1$ to T **do**
    $q_{ij} \leftarrow \frac{(1+||y_i - y_j||^2)^{-1}}{\sum_{k \neq 1} (1+||y_k - y_l||^2)^{-1}}$
    $\frac{\delta C}{\delta \gamma_i} \leftarrow 4 \sum_j (p_{ij} - qij)(y_i - y_j)(1 + ||y_i - y_j||^2)^{-1}$
    $\gamma^{(t)} \leftarrow \gamma^{t-1} + \eta \frac{\delta C}{\delta \gamma_i} + \alpha(t)(\gamma^{(t-1)} - \gamma^{(t-2)})$
end for

When used on MNIST digits dataset, as seen in Figure 4, clear distinctions are shown between digits. However, the digits are still close together and indistinguishable in some areas of the 2-dimensional plane. Thus, PCA, or any other dimensionality reduction technique, can be used to extract more relevant features from the data before being fed into t-SNE algorithm.



(a) Without PCA



(b) With PCA

Figure 4: t-SNE on MNIST Digits Dataset

3 Conclusion

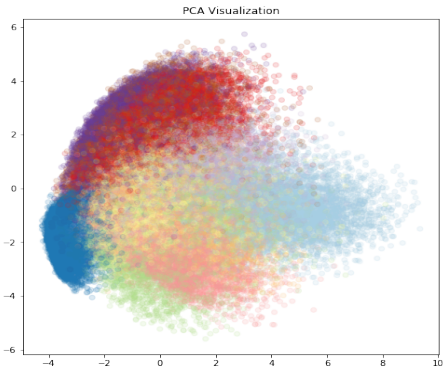
Table 1: Data Visualization Comparison

Algorithm	Digits Separation	Incremental?
PCA	Poor	Yes
t-SNE	Good	No
t-SNE & PCA	Excellent	No

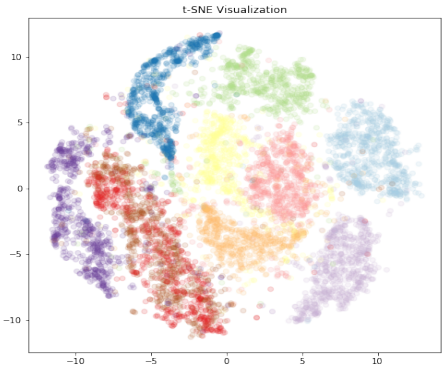
According to the resultant visualizations, it can be concluded to between Principal Component Analysis and t-Distributed Stochastic Neighbor Embedding, the latter performs remarkably well in mapping high-dimensional data into low-dimensional data for ease of visualization. Moreover, both methods can be used together to further improve the quality of the data visualization.

Upon further studies regarding the abilities and limitations of each of these data visualization algorithms, it was found that PCA and its variants are able to perform data visualization using previously

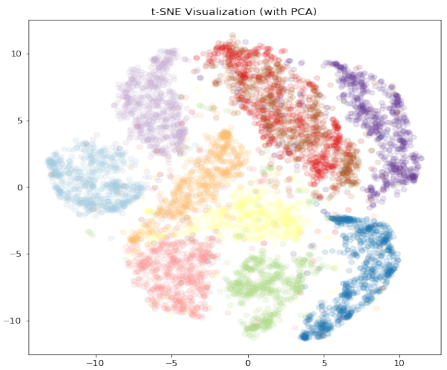
fed data, which means that the previous results are combined with new ones to create a continuous visualization of data. However, this is not true in the original implementation of t-SNE.



(a) PCA



(b) t-SNE



(c) t-SNE & PCA

Figure 5: MNIST Digits Dataset Visualizations

References

[1] Hadi Amiri and Hal Daumé III. *Short Text Representation for Detecting Churn in Microblogs*. AAAI Conference on Artificail Intelligence, 2016.

[2] Dorina Marghescu. *Multi-dimensional Data Visualization Techniques for Exploring Financial Performance Data*. Americas Conference on Information Systems (AMCIS), 2007.

[3] R. Kalaivani and Dr. S. Vijayarani. *Data Stream Mining - A Survey*. International Journal of Innovative Research in Computer and Communication Engineering, 2017.

[4] Georges Grinstein, Marjan Trutschl, and Urška Cvek. *High-Dimensional Visualizations*. ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, 2001.

[5] I.T. Jolliffe. *Principal Component Analysis*. Springer-Verlag New York, 2002.

[6] L.J.P van der Maaten and G.E. Hinton. *Visualizing High-Dimensional data Using t-SNE*. Journal of Machine Learning Research, 2008.

[7] L.J.P van der Maaten. *Accelerating t-SNE using Tree-Based Algorithms*. Journal of Machine Learning Research, 2014.

[8] David M. Chan, Roshan Rao, Forrest Huang, and john F. Canny. *t-SNE-CUDA: GPU-Accelerated t-SNE and its Applications to Modern Data*. High Performance Machine Learning Workshop, 2018.

[9] Puyudi Yang, Cho-Jui Hsieh, and Jane-Ling Wang. *History PCA: A New Algorithm for Streaming PCA*. University of California, 2018.

[10] S Kullback and R.A. Leibler. *On information and sufficiency*. Annals of Mathematcial Statistics, 1951.