Focused Stochastic Neighbor Embedding for Better Preserving Points of Interest

Rafael Baez Ramirez
dept. Computer Science
The University of Texas at El Paso
El Paso, TX, USA
baezrafael03@gmail.com

Sanuj Kumar
dept. Computer Science
New Mexico State University
Las Cruces, NM, USA
sanujkr@nmsu.edu

Tuan M. V. Le
dept. Computer Science
New Mexico State University
Las Cruces, NM, USA
tuanle@nmsu.edu

Huiping Cao
dept. Computer Science
New Mexico State University
Las Cruces, NM, USA
hcao@nmsu.edu

***Abstract*—Dimensionality reduction aims to find low-dimensional embeddings of high-dimensional data such that the low-dimensional representation preserves some meaningful properties of structures in the original data. When low-dimensional space is 2- or 3-dimensional, the low-dimensional embeddings can be visualized using a scatterplot map. Most of the existing methods try to preserve the local neighborhoods of all data points. However, in general, it is impossible to retain all such information for all data points in the low-dimensional space. As a result, there could be some data points whose neighborhoods are not faithfully displayed in the visualization due to information loss. If the information loss happens around a specific set of points of interest (e.g., specific patients, or proteins under observed), this may be problematic because the withdrawn insights may not be accurate for these observed data points. Therefore, in this paper, we introduce a problem called focused dimensionality reduction where given an original high-dimensional dataset and a set of points of interest, we want to find 2- or 3-dimensional embeddings of the original data such that the information loss in the local neighborhoods surrounding the points of interest is minimized as much as possible. In other words, if the information loss is inevitable, it should not happen around the points of interest. To solve the problem, we extend the stochastic neighbor embedding method and introduce a focused objective function where we put more weight on losses that involve points of interest. Experiments on real-world datasets show that our proposed method is better in preserving the local neighborhood structure of points of interest while the generated visualizations are as good as those generated by the stochastic neighbor embedding method.**

***Index Terms*—dimensionality reduction, stochastic neighbor embedding, focused analysis**

I. Introduction

Dimensionality reduction aims to transform complex high-dimensional data to low-dimensional embeddings so that the low-dimensional representation preserves as much of the original structural relationships as possible. One of the major applications of dimensionality reduction is data visualization. Given a dataset of N data points, $X = \{x_1, x_2, \ldots, x_N\}$, dimensionality reduction can be used to find embeddings $Y = \{y_1, y_2, \ldots, y_N\}$ of X in a 2- or 3-dimensional space that can be displayed in a scatterplot. This type of visualization has a wide range of applications such as understanding human genetic data [1]–[3], cluster analysis of geochemistry data [4], [5], network intrusion detection [6], [7], interpreting deep learning models and results [8], [9]. In these applications, visualization can help users visually discover patterns and insights in data that are difficult to detect in original space due to its complex nature.

Several dimensionality reduction techniques have been proposed including linear methods such as Principal Components Analysis (PCA), multidimensional scaling (MDS) [10], locality preserving projection [11]; and nonlinear methods such as Kernel PCA [12], UMAP [13], Stochastic Neighbor Embedding (SNE) [14], and t-SNE [15]. Most of the existing methods try to preserve the local neighborhoods of all data points. However, in general, it is impossible to retain all such information for all data points in the low-dimensional space. As a result, there could be some data points whose neighborhoods are not faithfully displayed in the visualization due to information loss. If the information loss happens around a specific set of points of interest (e.g., specific patients, or proteins under observed), this may be problematic because the withdrawn insights may not be accurate for these observed data points. Therefore, in this paper, we introduce a problem called focused dimensionality reduction where given a original high-dimensional dataset and a set of points of interest, we want to find 2- or 3-dimensional embeddings of the original data such that the information loss in the local neighborhoods surrounding the points of interest is minimized as much as possible. In other words, if the information loss is inevitable, it should not happen around the points of interest.

The main idea to solve the problem is that we will enforce the optimization of the dimensionality reduction to focus more on points of interest by putting more weights on losses that involve points of interest. To demonstrate this approach, we extend the stochastic neighbor embedding method [14] and introduce a focused objective function where pairs of points of interest will be preferentially optimized via regularization. Although we focus on implementing our idea with stochastic neighbor embedding, this approach can also be added on top of other dimensionality reduction methods to make them more focused in needed applications.

We summarize our main contributions as follows:

1) We introduce a problem called focused dimensionality reduction and propose a stochastic neighborhood embedding approach to solve the problem via regularization.
2) We conduct extensive experiments with datasets from different domains. The results show that our proposed

method is better in preserving the local neighborhood structure of points of interest while the generated visualizations are as good as those generated by the stochastic neighbor embedding method.

II. Related Work

There are two main approaches to embed high-dimensional data objects into a lower-dimensional space. The first approach includes linear techniques such as the classic method PCA [16] where a linear projection of the original data is used to capture as much variance found in the data as possible. Another linear method is multidimensional scaling (MDS) [10] that measures dissimilarities between data objects using Euclidean distance and learns the low-dimensional embeddings by keeping those dissimilar points far apart in the low-dimensional space. For data that lies on a non-linear manifold, these linear methods may not be able to discover that non-linear structure. Therefore, in the non-linear approach, several non-linear methods have been proposed such as Nonlinear PCA [17], Kernel PCA [12], Isomap [18], Self-organizing maps [19] and its probabilistic variant GTM [20], Elastic Nets [21], and LLE [22]. In non-linear approach, the most related work to our proposed method are SNE [14] and its t-distributed variant t-SNE [15]. Different from the above methods, SNE models the pairwise similarity between points i and j by using a probability defined over distances between the data points. The pairwise similarities are computed in both high-dimensional space and low-dimensional space. SNE then learns the low-dimensional embeddings by minimizing the discrepancies between the two probability distributions. t-SNE improves SNE to deal with the crowding problem by using Student-t distribution as a heavy-tailed distribution in the low-dimensional space. All above methods treat all data points as equally important and aim to preserve the structural information for all points. As a result, the inevitable information loss could happen to any data points. In contrast, our proposed method lets users specify a set of points of interest, and while trying to preserve the structural information for all points, our method will preferentially minimize the information loss around the points of interest.

III. Focused Stochastic Neighbor Embedding

We consider the problem of focused dimensionality reduction. Given a dataset of N data points in high-dimensional space, $X = \{x_1, x_2, \ldots, x_N\}$, and a set of I points of interest, $POI = \{p_1, p_2, \ldots, p_I\}$, $POI \subset X$, our proposed method aims to find embeddings $Y = \{y_1, y_2, \ldots, y_N\}$ of X in a 2- or 3-dimensional space for visualization such that the local neighborhoods of data points are preserved as much as possible in the visualization space, and the points of interest will be preferentially optimized to avoid the information loss in the local neighborhoods around them. To solve the problem, we propose a method called fSNE. fSNE extends SNE [14] with a focused objective function that allows the optimization to put more weights on modeling correctly the local neighborhoods of points of interest.

Following SNE, for two data points i and j, we parameterize the conditional probability that j is a neighbor of i, $p(j|i)$, as a function of Euclidean distances between the data points in the original high-dimensional space:

$$p(j|i) = \frac{\exp(-||x_i - x_j||^2/2\sigma_i^2)}{\sum_{k \neq i} \exp(-||x_i - x_k||^2/2\sigma_i^2)} \tag{1}$$

here σ_i is the variance of the Gaussian centered at x_i. The denominator summing over all pairs acts as a normalization ensuring that $\sum_j p(j|i) = 1$. Intuitively, $p(j|i)$ represents the similarity of i and j. The closer they are in the original space, the higher $p(j|i)$ is and the more similar they are.

Similarly, in the visualization space, the conditional probability that j is a neighbor of i, $q(j|i)$, can be computed as follows:

$$q_{j|i} = \frac{\exp(-||y_i - y_j||^2}{\sum_{k \neq i} \exp(-||y_i - y_k||^2)} \tag{2}$$

To learn the embeddings $Y = \{y_1, y_2, \ldots, y_N\}$, we will minimize the mismatch between $q(j|i)$ and $p(j|i)$. In other words, the similarity between y_i and y_j in the visualization space should reflect the similarity between x_i and x_j in the original space. More specifically, let P_i be the conditional distribution over all other data points given data point x_i in the original space, and Q_i be the conditional distribution over all other data points given data point y_i in the visualization space, we can minimize the Kullback-Leibler divergence between P_i and Q_i for all points i as follows:

$$\mathbf{L} = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p(j|i) \log \frac{p(j|i)}{q(j|i)} \tag{3}$$

The above objective function treats all points as equally important. Therefore, to let the model focus on points of interest POI and minimize the information loss around them as much as possible, we introduce a coefficient parameter λ for pairs of points that involve points of interests. More specifically, the new objective function that incorporates information on points of interest becomes:

$$\mathbf{L}^* = \lambda \left(\sum_{i \in POI} \sum_{j \notin POI} p_{j|i} log \frac{p_{j|i}}{q_{j|i}} + \sum_i \sum_{j \in POI} p_{j|i} log \frac{p_{j|i}}{q_{j|i}} \right) + \sum_{i \notin POI} \sum_{j \notin POI} p_{j|i} log \frac{p_{j|i}}{q_{j|i}} \tag{4}$$

here $\lambda \geq 1$ is the weight for pairs that involve points of interests [1]. The greater λ is, the more the model will focus on minimizing the mismatch between $q(j|i)$ and $p(j|i)$, if i or j is one of the points of interest. Note that when $\lambda = 1$, $\mathbf{L}^*$ becomes $\mathbf{L}$, i.e., there is no focus on any points of interest and fSNE will be reduced to the original SNE.

[1] λ is set to 2 in our experiments

Similar to SNE, we also implement the binary search for σ_i in Eq.1 that produces a P_i with a fixed perplexity. The perplexity can be interpreted as the effective number of neighbors. We vary the perplexity in our experiments to show its effect to the performance. To optimize $\mathbf{L}^*$, we provide a GPU-accelerated implementation of fSNE that uses Adam as the optimizing algorithm. fSNE scales well to large datasets on GPUs with enough memory.

IV. Experiments

We evaluate the visualization performance of fSNE on the following three datasets of different data types:

1) **MNIST**[2] **-** MNIST is an image dataset with handwritten digits (28x28 pixels). For our experiments, we use a sample of 5000 images (500 images for each digit).
2) **20Newsgroups**[3] **-** 20Newsgroups is a corpus of news articles categorized into 20 groups. For experiments, we use the training subset that has 11314 documents. As preprocessing, we lemmatize the words, remove stopwords and documents with length less than 5 words. The vocabulary size is 5000. We represent documents using tf-idf vectors.
3) **Wine Quality**[4] **-** Wine quality is a numerical dataset of 3918 samples of wines. Each sample has 11 wine features (e.g., volatile acidity, citric acid, density, pH, and alcohol). The labels are the quality levels of wine samples. This dataset includes wine samples that have quality scores from 3 (poor) to 9 (excellent).

For each dataset, we randomly sample three sets of points of interest; each has 100 data points. We set $\lambda = 2$ for fSNE and vary the perplexities from 10 to 50. For a direct comparison, we compare our method to SNE. Note that other methods are orthogonal to our method because we can add the focused layer on top of those methods for focused analysis. fSNE and SNE are trained for 1000 iterations with learning rate set to 0.1 and optimized using Adam algorithm. All experimental results are averaged across three independent runs.

1) Quantitative Analysis: For evaluating the quality of the local structure preservation, we adopt the Local Approximation of Preserved Structure (LAPS) [23] that calculates the local divergence for a point x_i given the dataset X and the embeddings Y. A lower local divergence score signifies a better preservation of the original local neighborhood structure. We compute the local divergence for every point in the POI set and report the averaged local divergence. Figure 1 shows the results on three datasets with varied perplexities. As we can see, fSNE is better in preserving the local neighborhood structure for points of interest across all datasets and all settings. Therefore, our method is particularly useful when users want the patterns involving the points of interest to be faithfully displayed in the visualization.

[2]https://pytorch.org/vision/stable/generated/torchvision.datasets.MNIST.html
[3]https://scikit-learn.org/0.19/datasets/twenty_newsgroups.html
[4]https://github.com/aindrila-ghosh/LAPS_and_GAPS/blob/master/data/Wine_Quality.csv

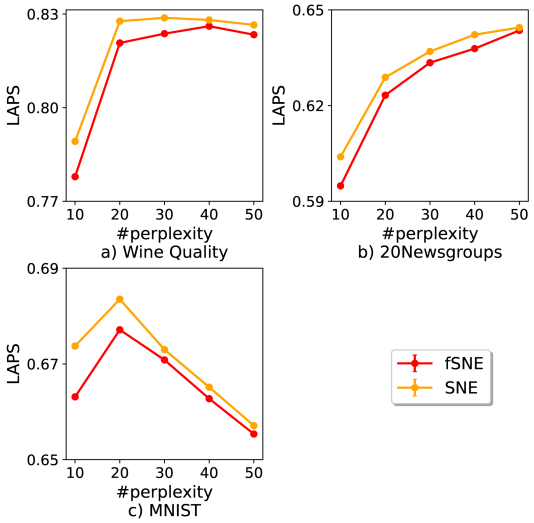

Fig. 1. Local divergence across different perplexities

As another quantitative evaluation, we want to show that while preferentially optimizing for points of interest, our method also produces a good visualization of all the points in the dataset. To measure the quality of the visualization, we use the k-nearest neighbors (kNN) accuracy in the visualization space [15]. A good visualization will achieve a high classification accuracy because it groups documents of the same label together in the visualization space. We report the averaged kNN accuracy across different k for fSNE and SNE in Figure 2. As we can see, the performance of fSNE is comparable to that of SNE, which demonstrates that the generated visualizations by fSNE are as good as those generated by SNE.

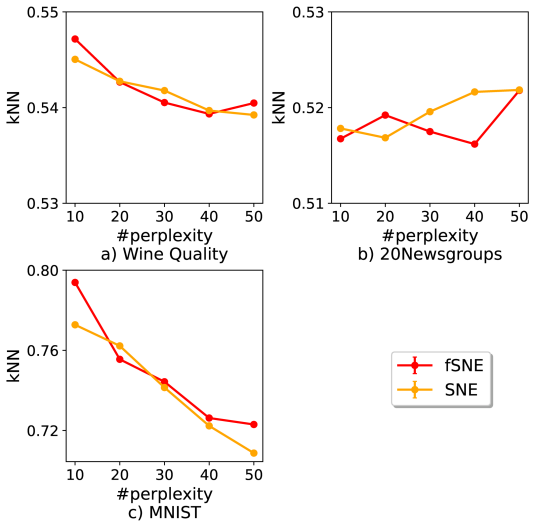

Fig. 2. kNN accuracy across different perplexities

2) Qualitative Analysis: For qualitative evaluation, we present example visualizations of fSNE and SNE in Figures 3, 4 and 5 where it can be seen that fSNE's overall visualizations are as good as those of SNE. For a deeper analysis to

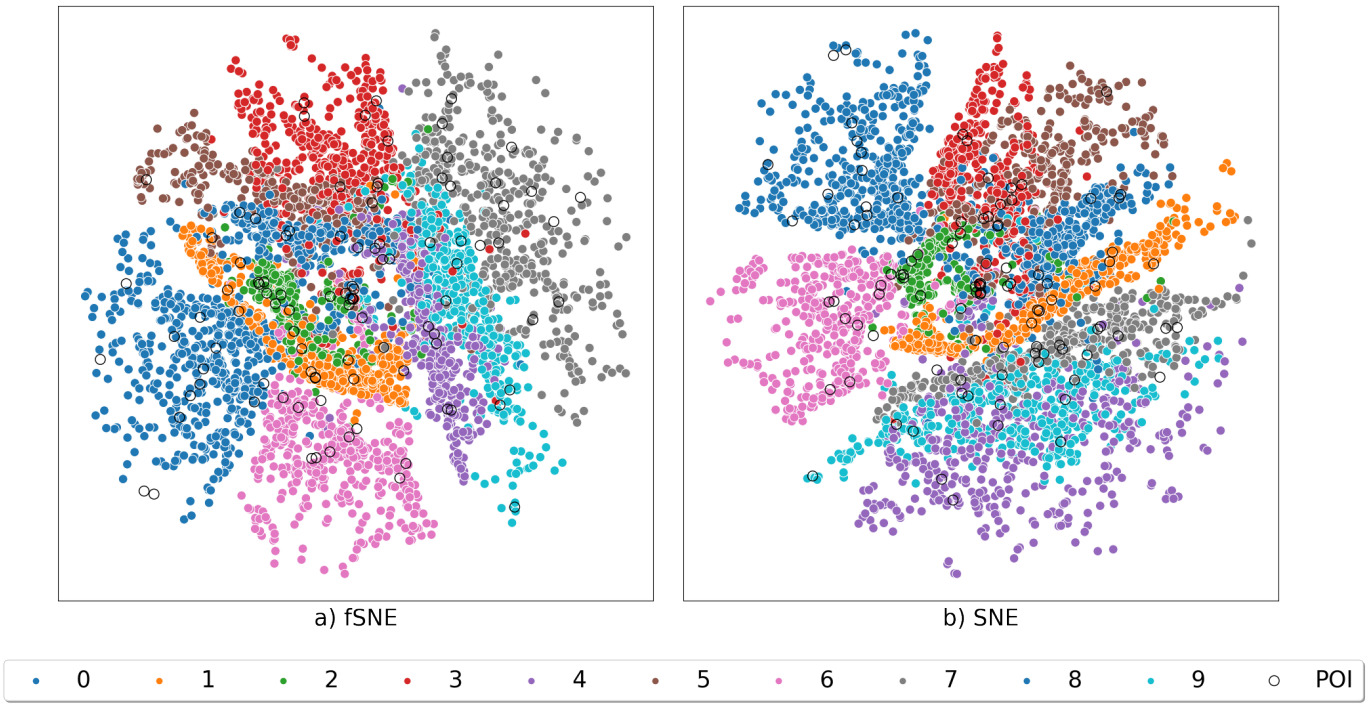

Fig. 3. Visualization of MNIST by fSNE and SNE with perplexity $= 10$

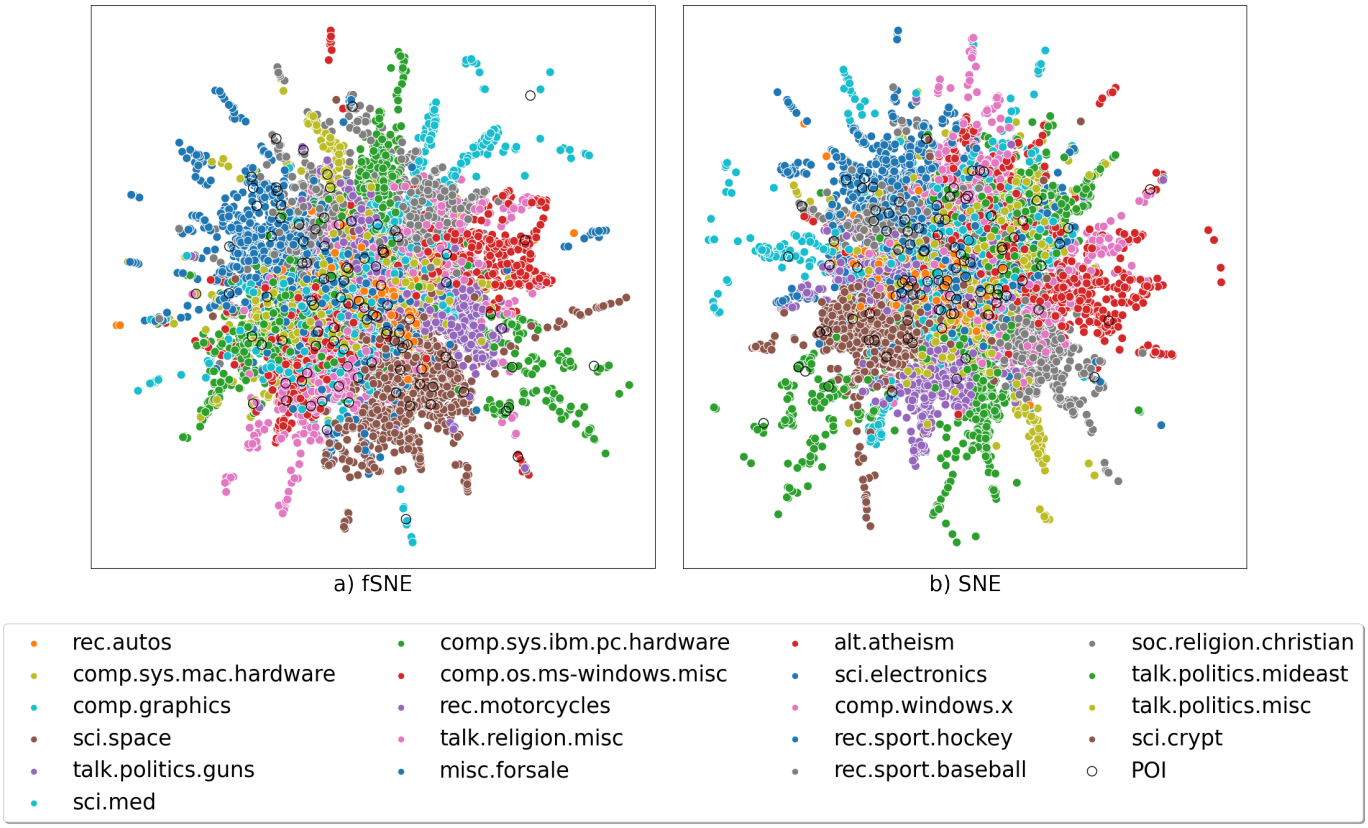

Fig. 4. Visualization of 20Newsgroups by fSNE and SNE with perplexity $= 10$

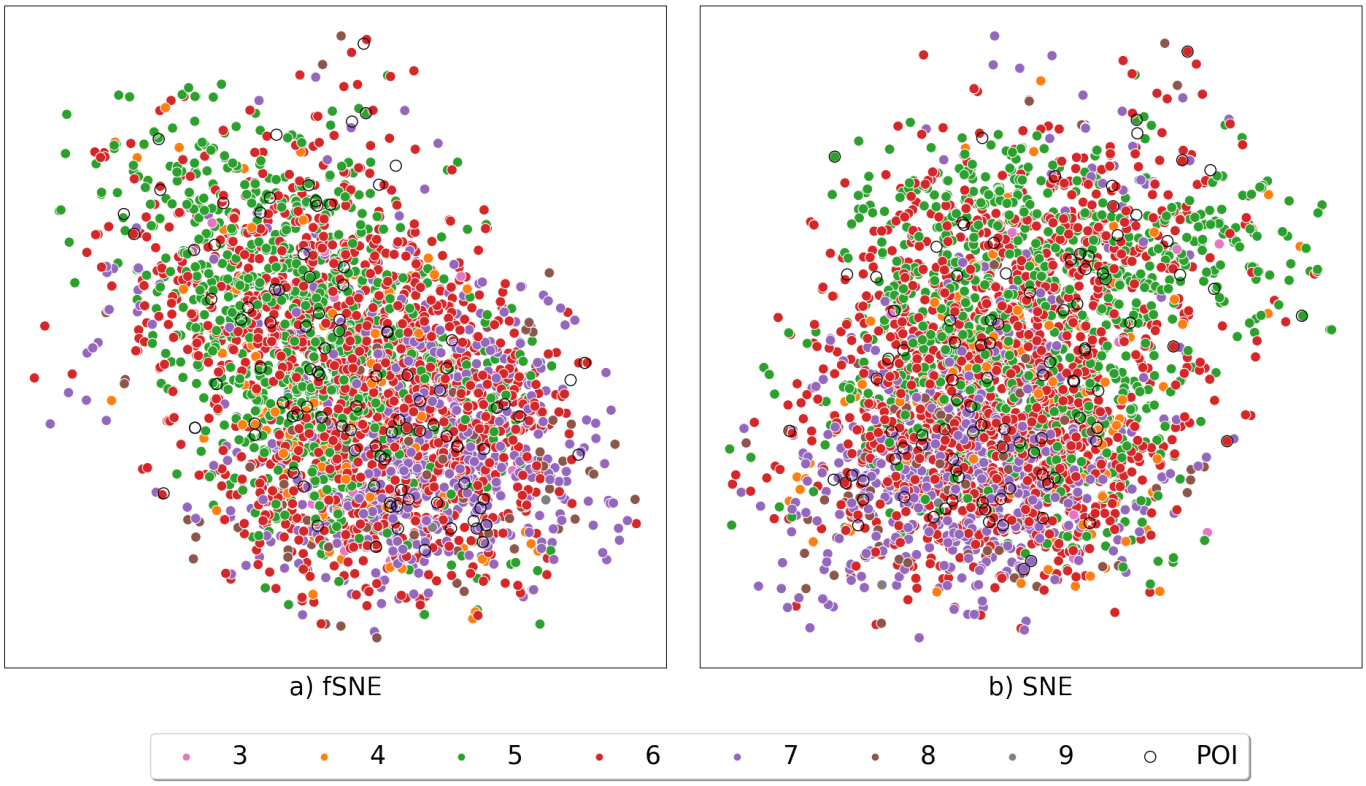

Fig. 5. Visualization of Wine Quality by fSNE and SNE with perplexity $= 10$

TABLE I
RUNNING TIME (IN SECONDS)

Name	Type	Instances	Features	SNE	fSNE
Wine Quality	Tabular	3918	11	209.9	208.3
MNIST	Image	5000	784	256.4	259.4
20Newsgroups	Text	11314	5000	994.5	1035.6

showcase how well fSNE preserves the local neighborhood structure of points of interest, Figures 6, 7 and 8 show the original local neighborhood structure surrounding a point of interest (the red point near the center of each figure). In those figures, we zoom in the visualizations and show the 50 nearest neighbors of the point of interest. The original neighborhood points are marked with an 'x'. We clearly notice that in the visualizations by fSNE, among the 50 nearest neighbors in the visualization space, there are more points that are the original neighbors of the point of interest in the high-dimensional space, which demonstrates the effectiveness of fSNE in preserving the local neighborhood structure of points of interest.

3) Running Time: fSNE scales well to large datasets on GPUs with enough memory. The running time for each dataset is reported in Table I.

V. CONCLUSION

We propose a method based on stochastic neighbor embedding for focused dimensionality reduction that aims to preferentially preserve the local neighborhood structures of points of interest. We demonstrate the effectiveness of the proposed method in visualizing high-dimensional datasets. The results show that our proposed method is better in preserving the local neighborhood structure of points of interest while the generated visualizations are as good as those generated by the stochastic neighbor embedding method. For future

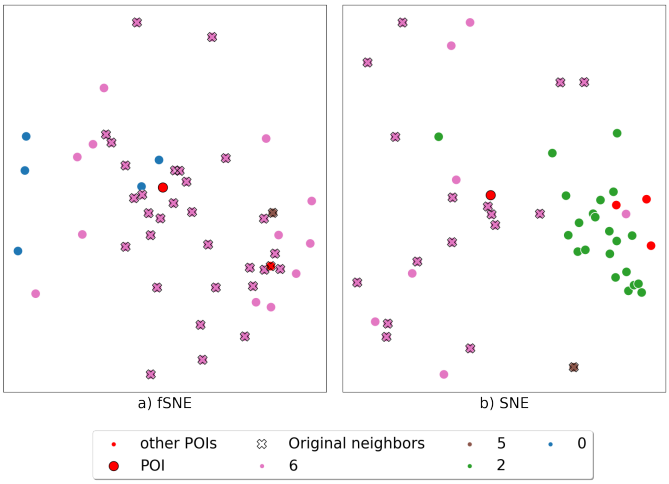

Fig. 6. Scatterplot shows the 50 nearest neighbors of a given point of interest in the visualization space for MNIST dataset. The original neighborhood points are marked with an 'x'

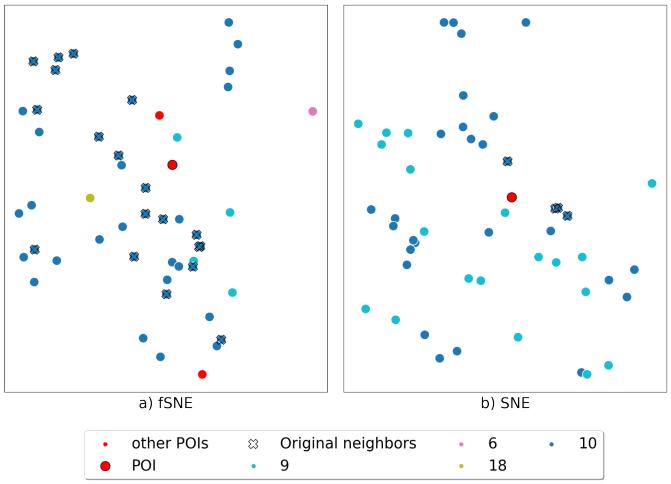

Fig. 7. Scatterplot shows the 50 nearest neighbors of a given point of interest in the visualization space for 20Newsgroups dataset. The original neighborhood points are marked with an 'x'

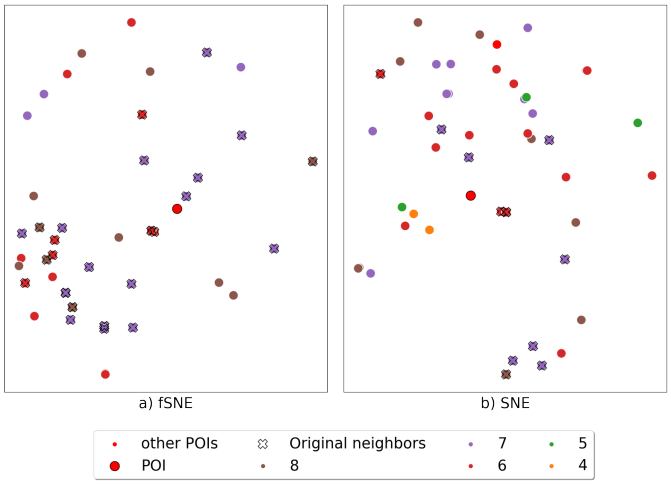

Fig. 8. Scatterplot shows the 50 nearest neighbors of a given point of interest in the visualization space for Wine Quality dataset. The original neighborhood points are marked with an 'x'

work, we plan to develop a generalized framework that allows conveniently adding a focused layer on top of non-linear dimensionality methods for focused analysis.

Acknowledgment

This work is supported by the grant "REU Site: BIGDatA - Big Data Analytics for Cyber-physical Systems" from the National Science Foundation (NSF #1950121).

References

[1] W. Li, J. E. Cerise, Y. Yang, and H. Han, "Application of t-sne to human genetic data," *Journal of bioinformatics and computational biology*, vol. 15, no. 04, p. 1750017, 2017.

[2] A. Platzer, "Visualization of snps with t-sne," *PloS one*, vol. 8, no. 2, p. e56883, 2013.

[3] G. C. Linderman, M. Rachh, J. G. Hoskins, S. Steinerberger, and Y. Kluger, "Fast interpolation-based t-sne for improved visualization of single-cell rna-seq data," *Nature methods*, vol. 16, no. 3, pp. 243–245, 2019.

[4] H. Liu, J. Yang, M. Ye, S. C. James, Z. Tang, J. Dong, and T. Xing, "Using t-distributed stochastic neighbor embedding (t-sne) for cluster analysis and spatial zone delineation of groundwater geochemistry data," *Journal of Hydrology*, vol. 597, p. 126146, 2021.

[5] M. Balamurali and A. Melkumyan, "t-sne based visualisation and clustering of geological domain," in *International Conference on Neural Information Processing*. Springer, 2016, pp. 565–572.

[6] Y. Hamid and M. Sugumaran, "A t-sne based non linear dimension reduction for network intrusion detection," *International Journal of Information Technology*, vol. 12, no. 1, pp. 125–134, 2020.

[7] H. Yao, C. Li, and P. Sun, "Using parametric t-distributed stochastic neighbor embedding combined with hierarchical neural network for network intrusion detection." *Int. J. Netw. Secur.*, vol. 22, no. 2, pp. 265–274, 2020.

[8] M. Kahng, P. Y. Andrews, A. Kalro, and D. H. Chau, "A cti v is: Visual exploration of industry-scale deep neural network models," *IEEE transactions on visualization and computer graphics*, vol. 24, no. 1, pp. 88–97, 2017.

[9] Y. Wang, H. Su, B. Zhang, and X. Hu, "Interpret neural networks by identifying critical data routing paths," in *proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8906–8914.

[10] W. S. Torgerson, "Multidimensional scaling: I. theory and method," *Psychometrika*, vol. 17, no. 4, pp. 401–419, 1952.

[11] X. He and P. Niyogi, "Locality preserving projections," *Advances in neural information processing systems*, vol. 16, 2003.

[12] B. Schölkopf, A. Smola, and K.-R. Müller, "Kernel principal component analysis," in *International conference on artificial neural networks*. Springer, 1997, pp. 583–588.

[13] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[14] G. E. Hinton and S. Roweis, "Stochastic neighbor embedding," *Advances in neural information processing systems*, vol. 15, 2002.

[15] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[16] Y. Qu, G. Ostrouchov, N. Samatova, and A. Geist, "Principal component analysis for dimension reduction in massive distributed data sets," in *Proceedings of IEEE International Conference on Data Mining (ICDM)*, vol. 1318, no. 1784, 2002, p. 1788.

[17] M. Scholz, M. Fraunholz, and J. Selbig, "Nonlinear principal component analysis: neural network models and applications," in *Principal manifolds for data visualization and dimension reduction*. Springer, 2008, pp. 44–67.

[18] J. B. Tenenbaum, V. d. Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[19] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biological cybernetics*, vol. 43, no. 1, pp. 59–69, 1982.

[20] C. M. Bishop, M. Svensén, and C. K. Williams, "Gtm: The generative topographic mapping," *Neural computation*, vol. 10, no. 1, pp. 215–234, 1998.

[21] A. N. Gorban and A. Y. Zinovyev, "Elastic maps and nets for approximating principal manifolds and their application to microarray data visualization," in *Principal manifolds for data visualization and dimension reduction*. Springer, 2008, pp. 96–130.

[22] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[23] A. Ghosh, M. Nashaat, J. Miller, and S. Quader, "Interpretation of structural preservation in low-dimensional embeddings," *IEEE Transactions on Knowledge and Data Engineering*, 2020.