m-TSNE: A Framework for Visualizing High-Dimensional Multivariate Time Series

Minh Nguyen[1], Sanjay Purushotham, PhD[1], Hien To[1], Cyrus Shahabi , PhD[1]
[1]University of Southern California, Los Angeles, CA, USA

Abstract

Multivariate time series (MTS) have become increasingly common in healthcare domains where human vital signs and laboratory results are collected for predictive diagnosis. Recently, there have been increasing efforts to visualize healthcare MTS data based on star charts or parallel coordinates. However, such techniques might not be ideal for visualizing a large MTS dataset, since it is difficult to obtain insights or interpretations due to the inherent high dimensionality of MTS. In this paper, we propose "m-TSNE": a simple and novel framework to visualize high-dimensional MTS data by projecting them into a low-dimensional (2-D or 3-D) space while capturing the underlying data properties. Our framework is easy to use and provides interpretable insights for healthcare professionals to understand MTS data. We evaluate our visualization framework on two real-world datasets and demonstrate that the results of our m-TSNE show patterns that are easy to understand while the other methods' visualization may have limitations in interpretability.

1 Introduction

Big data analytics in healthcare is emerging as a large amount of health and medical data are being generated every day. Recent development of different types of health sensors and e-health platforms has opened up great opportunities for collecting, monitoring and analyzing patients' health conditions from multiple data sources [1]. Performing analytics and extracting insights on healthcare data is challenging due to the large volume, high dimensionality, heterogeneity, and dynamic nature of the healthcare data. To address these challenges, many studies are being conducted in several fields such as machine learning, data mining, statistics, health informatics, etc. Data visualization is one such field that provides tools for visual interpretations of the underlying data patterns and trends, and helps in further data analysis.

Several techniques [2, 3, 4, 5] have been developed to visualize and analyze time series data. Since these techniques usually analyze univariate time series (UTS) data by handling only one data variable at a time (e.g., monitoring respiratory rate or heart rate over time to detect anomaly), they may not fully capture the inherent correlations of the multivariate time series (MTS) data [6]. Therefore, we believe that MTS data corresponding to multiple variables should be treated as a whole, rather than being broken into individual UTS as they can provide greater insight into data representing patients' conditions. However, multivariate data is notoriously hard to represent because of the difficulty of mentally picturing data in more than three dimensions [7].

Previous visualization efforts [8, 9, 10, 11] on multivariate data focused on displaying multiple dimensions of the data in 2-D plot, and left the interpretations to human observer. Figures 1 and 2 show the star chart and parallel coordinates visualizations that have been previously proposed for healthcare data visualization [10, 11]. These techniques might not be ideal for real-world applications which have high-dimensional MTS data points since they tend to result in complex visualization plots with less interpretability. Researchers have explored dimensionality reduction techniques [12, 13, 14] to handle high-dimensional data points by projecting them into low-dimensional space. These approaches have been applied successfully in computer vision applications where data consists of images. However, such techniques, in general, cannot be applied directly on medical MTS datasets because each data source may have different properties, e.g. time dependency structure is present in vital signs time series. Hence, there is a need for developing new techniques to visualize high-dimensional MTS data.

In this paper, we propose m-TSNE (Multivariate Time Series t-Distributed Stochastic Neighbor Embedding): a framework for visualizing MTS data in low-dimensional space that is capable of providing insights and interpretations of the high-dimensional MTS datasets. m-TSNE first calculates the similarity between
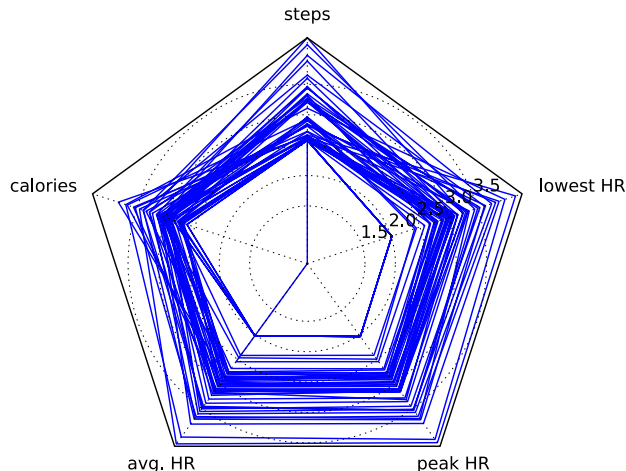
Figure 1: Star chart visualization [8] of MTS human monitoring data of a subject in ATOM-HP dataset (Section 4.1). Data has 5 variables: step counts, calories, lowest heart rate, average heart rate, and peak heart rate. Each polygon is a multivariate data point at a time instance.
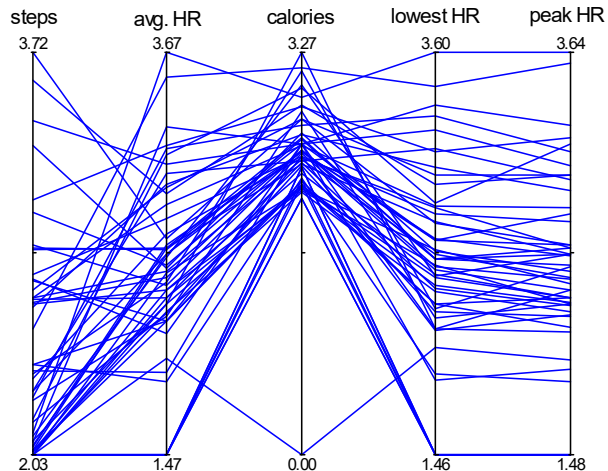

Figure 2: Parallel coordinates visualization [9] of the same data shown in Figure 1. Each polyline that connects the parallel axes represents a multivariate data point at a time instance. The number of polyline in the figure is the number of time instance in the MTS.

each MTS data points in high-dimensional space, based on Extended Frobenius norm (EROS) [6] which is a similarity metric for MTS data. Then, it computes the low-dimensional (2-D or 3-D) projection of the MTS data points using a gradient descent method by preserving the similarity relation between pairs of high-dimensional points. We conduct visualization experiments on two healthcare datasets: ATOM-HP dataset (Section 4.1), a dataset collected to study impact of chemotherapy on patient's activity; and Electroencephalogram (EEG) dataset [15], a dataset collected to study whether there is a genetic predisposition to alcoholism. We evaluate and the compare m-TSNE's visualizations with other competing approaches by conducting an user study. We show that m-TSNE's visualizations on the ATOM-HP dataset can extract patient activity level patterns and outliers (during chemotherapy cycle) which helps oncologists to study their treatment's effects on patient's fatigue. On the EEG dataset, we show that control and alcoholic subjects can be easily identified (separated) in our low-dimensional visualization which is not possible in the original high-dimensional space.

The organization of the paper is as follows: We discuss the related work in the field of data visualization in Section 2. Section 3 describes our m-TSNE visualization framework. Section 4 reports the empirical evaluation using the two aforementioned datasets and Section 5 concludes with discussion and future work.

2 Related Work

Due to the pervasiveness of time series data and its wide range of applications in healthcare monitoring, stock market analysis, traffic analysis, etc., understanding time-series data through visualization has attracted lots of attention, among which there have been efforts focusing on visualizing MTS data [8, 9, 10, 16]. These techniques visualize multivariate data by simply displaying individual variable data independently on one shared-space display [17] without treating multiple variables as a whole. Nevertheless, shared-space techniques may not be efficient to visualize high-dimensional MTS dataset since the resulting shared-space visualization may show a large amount of overlapping and clutter between different variables' time series, which is difficult to interpret. For instance, as shown in Figure 3, in [10], the authors proposed a technique named Multivariate Time Series Amalgam (MTSA), to jointly visualize multiple variables of MTS on a single display. Nevertheless, MTSA may not be suitable when the MTS dataset has a large number of measurements or has high dimensionality. The authors in [16] proposed a visualization method for multivariate data based on star chart [8] which represents each variable on an axis, with the axes arranged around a circle. An example of this technique is depicted in Figure 1. Another popular technique for visualizing MTS is parallel
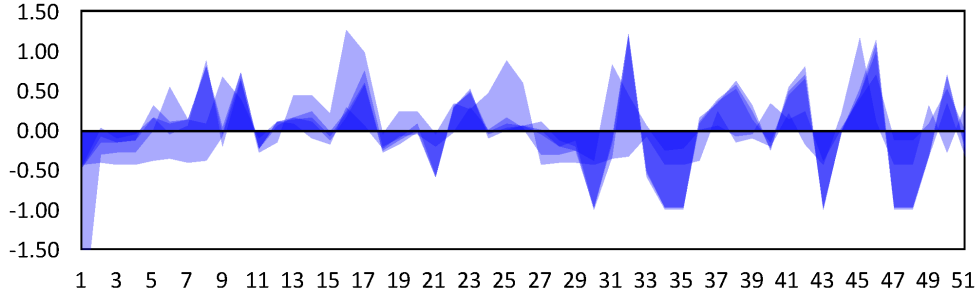
Figure 3: MTSA visualization [10] of the same raw data as Figure 1. x-axis represents the time instance (the date number), and y-axis represents the variables values.
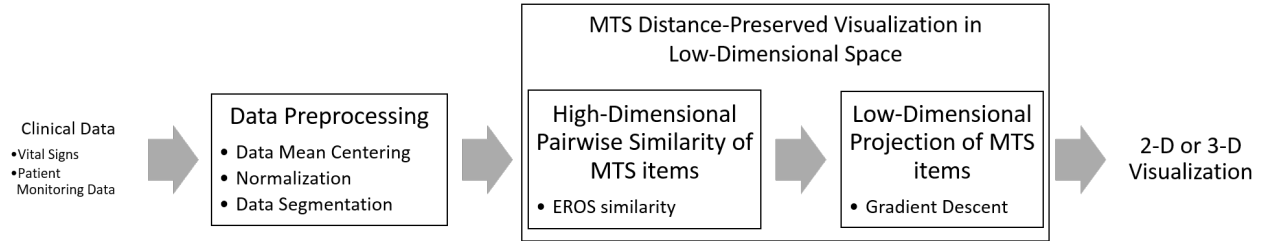

Figure 4: Pipeline of m-TSNE framework for MTS Visualization

coordinates [9] (shown in Figure 2). These methods have limitations in interpretability when the MTS has a large number of variables or a large number of time instances per variable. Unlike these studies, our proposed m-TSNE technique deals with multiple variables of MTS as a whole by applying dimensionality reduction techniques to project MTS data points to low-dimensional latent space and it also preserves the distances between the MTS in the original space.

3   MTS Visualization Framework

Figure 4 shows the pipeline of our proposed m-TSNE approach for MTS Visualization. First, the MTS data is processed by mean-centering and normalization; then it is segmented into multiple MTS items [12] (discussed in Section 3.1). EROS [6] is calculated to find the high-dimensional pairwise similarity between the MTS items (see Section 3.2). For visualization, the MTS items is projected to low-dimensional space with the guarantee that the high-dimensional pairwise similarity relation is preserved (explained in Section 3.3). That is, if two MTS items are neighbors (or far apart) in high-dimensional space, their low-dimensional projection points should also be close (or far apart) to each other. We use gradient descent method similar to [14] for dimensionality reduction and projection.

3.1   Data Preprocessing

Data preprocessing consists of data mean-centering, normalization, and segmentation. Mean-centering by subtracting the mean value of a variable, and normalization by dividing the variable with its standard deviation helps the MTS items to be in the same scale [12] which is needed for the further MTS similarity calculation step. The MTS data is also segmented into MTS items corresponding to the time instance (eg. an hour, a day). The trend of the data over the time should be visualized as the patterns of all MTS items.

3.2   MTS Pairwise High-dimensional Similarity

One of the widely used techniques for high-dimensional data similarity calculation is t-Distributed Stochastic Neighbor Embedding (t-SNE) [13]. In order to calculate the high-dimensional data points similarity, t-SNE

[13] computes the Euclidean distance between two data points, then converts the distance into conditional probabilities that represent similarities. The details of the techniques can be found in [18].

As our study concerns MTS, the similarity metric should be suitable for MTS data type and should also take into account the correlation between MTS variables. The Euclidean distance similarity metric is shown to be not suitable for time series data as acceleration and deceleration along the time axis is suboptimal for distance matching [19]. As a result, Dynamic Time Warping (DTW) is generally used to overcome the limitations of Euclidean distance metric [20]. However, both of these distance only work for UTS data matching, as they do not consider the correlation between MTS variables. Therefore, the t-SNE technique may not perform well on MTS datasets if Euclidean distance or DTW is used as a similarity metric.

We propose to use EROS similarity metric [6] to overcome the shortcomings for above similarity metrics. EROS similarity metric is a technique based on the Principal Component Analysis method in which an MTS item is treated as a whole, i.e. it is not broken in multiple UTS, to preserve the correlation between variables. Let us denote MTS data by X, where $X = \{x_1, x_2, ..., x_k\}$ and x_i corresponds to a multivariate data point (high-dimensional) at time instance i. Each x_i is termed as an MTS item and it is represented as an m x n matrix, where m is the number of observations (eg. patients), and n is the number of variables. Each variable can be a vital sign variable such as heart rate, respiration rate, etc. or an activity monitoring variable such as step count, active hour, etc. Given 2 MTS items x_i and x_j, EROS first computes the eigenvectors and eigenvalues of each item. Thereafter, it measures the cosine similarities of the corresponding eigenvectors of x_i and x_j. Finally, EROS similarity is the weighted sum of all the cosine similarities of the eigenvectors. The weight is calculated as the aggregated value based on all the eigenvalues of the MTS items in the dataset. The EROS similarity metric is described in Equation 1:

$$EROS(x_i, x_j, w) = \sum_{l=1}^{n} w_l | < v_{il}, v_{jl} > |$$ (1)

Where, $EROS(x_i, x_j, w)$ is the similarity of MTS item x_i and MTS item x_j, $v_i = [v_{i1}, \cdots, v_{in}]$ and $v_j = [v_{j1}, \cdots, v_{jn}]$ are the two sets of eigenvectors of x_i and x_j respectively and w_l is the aggregated weight computed based on the eigenvalue corresponding to the lth eigenvector in the weight vector w. The computed pairwise similarities of MTS items are used for projecting each MTS item into low-dimensional (2-D or 3-D) space. Section 3.3 describes the details of how to project MTS data points to lower dimensional space using a gradient descent method.

3.3 MTS Low-dimensional Projection

Principal Component Analysis (PCA) [12, 21] is a popular approach to map high-dimensional data into low-dimensional space. PCA is a linear mapping which focuses on preserving the low-dimensional projection of dissimilar data points far apart. However, PCA is shown not to be suitable for non-linear manifolds where preserving the low-dimensional projection of similar data points close to each other is important [13, 14]. To overcome this drawbacks of PCA, we propose to perform an optimization method: gradient descent (similar to the one used in t-SNE [14]) for m-TSNE where we minimize the mismatch between high-dimensional and low-dimensional spaces. The gradient descent method works by minimizing a cost function over all data points. The details of the cost function and performing gradient descent is explained in our technical report [18]. In the following section, we will discuss the empirical results of our m-TSNE approach, and compare it to PCA, Euclidean-based t-SNE, and DTW-based t-SNE approaches for MTS visualization.

4 Experiments

4.1 Datasets

We evaluate m-TSNE visualization using two healthcare datasets: ATOM-HP[1] dataset which is a coarse-grained time series dataset collected from our on-going study, and EEG dataset - a fine-grained time series dataset from UCI machine learning repository [22].

---

[1]ATOM-HP: Analytical Technologies to Objectively Measure Human Performance

ATOM-HP dataset : This dataset is collected to study how to quantify the activity levels of cancer patients undergoing chemotherapy treatment to complement the current clinical assessment. The patients suffer from treatment induced fatigue that affects their daily activities, which is usually measured by physicians using Eastern Cooperative Oncology Group (ECOG) scores [23]. However, this score tends to suffer from subjective bias. Therefore, a more robust objective measurement is needed to evaluate patients' activity / performance status, which is the goal of our study. In ATOM-HP, each patient carries a wearable sensor during their chemotherapy treatment cycle. The chemotherapy cycle consists of two chemotherapy visits. The date range between the two chemotherapy visits varies from two to three weeks. There are eight patients in our dataset (more patients are being enrolled in this on-going study). The daily data of each patient is an MTS which has five variables: number of steps, total calories, average heart rate, peak heart rate, and lowest heart rate. Collected data is sampled every hour and for each patient, data is collected for at least 50 consecutive days.

EEG dataset [15] : The dataset is collected to examine if EEG correlates genetic predisposition to alcoholism. There are three versions of the dataset. In our work, we use the large dataset version that has 10 control subjects, and 10 alcoholic subjects. Each subject performs 30 trials which can be classified as three trial types: exposure to a single stimulus, exposure to two matching stimulus, and exposure to two non-matching stimulus [24]. In total there are 600 (=20x30) trials. The data of one trial is an MTS of 64 variables corresponding to 64 electrodes placed on the subject's scalps. The data sample rate is 256Hz.

4.2 Experimental Setup

To evaluate our framework, we compare m-TSNE to the visualizations of PCA, Euclidean-based t-SNE, and DTW-based t-SNE methods. In the ATOM-HP dataset, given an MTS data of one subject, we are interested in visualization of MTS to show the trends and outliers in the subject's daily activity performance during the chemotherapy treatment. The subject's MTS data is represented as an m x 5 matrix with m is the total hours of monitoring, and 5 is the number of variables (Section 4.1). For instance, the size of the MTS matrix data of the subject in Figure 5 is 1224 x 5. As the study considers monitoring daily performance, the MTS is segmented into multiple MTS items of 24 x 5 matrices which represent the 24 hours of the data after mean-centering, and normalization. In Figure 5, we have a total of 51 (= 1224/24) data points corresponding to 51 MTS items data matrices collected during 51 days when the subject was involved in the study. For m-TSNE, we calculate the pairwise similarity using EROS similarity metric (Equation 1). The pairwise similarities are put through the gradient descent to compute the MTS items' projection in low-dimensional (2-D or 3-D) space (Figure 5, Figure 7). For the EEG dataset, each subject's trial is represented as an MTS matrix of 256 x 64 where 256 is the number of observations and 64 is the number of EEG variables. The pairwise EROS similarities of 600 MTS items corresponding to 600 trials is calculated. The similarities are then used to compute the projection of 600 MTS as Figure 6.

PCA: The MTS is preprocessed, and segmented into MTS items as above (Section 4.2). After preprocessing, the daily aggregated multivariate data points are computed as sum of number of steps, sum of calories, average heart rate, peak heart rate, and lowest heart rate over 24 hours. All aggregated data of MTS items are put together as a matrix of m' x 5 where 5 is the number of variables, and m' is the number of data points with each row representing an aggregated multivariate data point. PCA is used for low-dimensional projection and visualization as shown in Figure 8.

Euclidean-based t-SNE: After data preprocessing, a pairwise Euclidean distance for each data points pair is calculated. Based on the Euclidean distance, a pairwise probability similarity matrix of each pair of data points is computed as in [18].

DTW-based t-SNE: Since Euclidean distance does not work well for time series data, we can measure MTS similarity using DTW distance. This approach is similar to Euclidean-based t-SNE, but uses the pairwise Dynamic Time Warping distance instead of Euclidean distance.

m-TSNE: m-TSNE calculates EROS pairwise similarity matrix, and computes MTS projection using the gradient descent method as described in Section 3.

We implemented above four methods in Python 2.7.11 and we will release our code on GitHub.
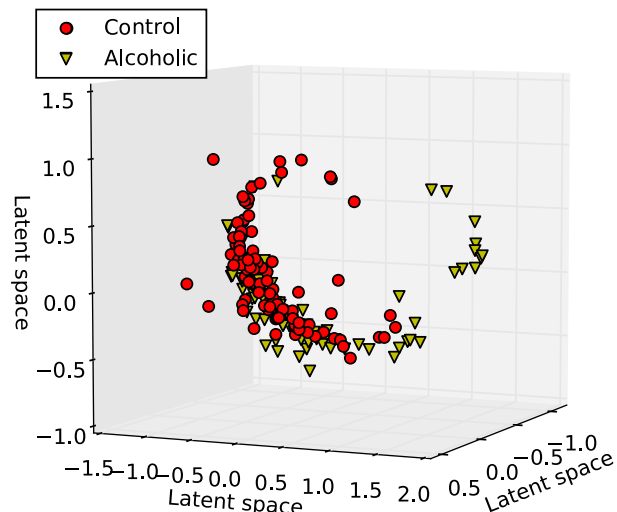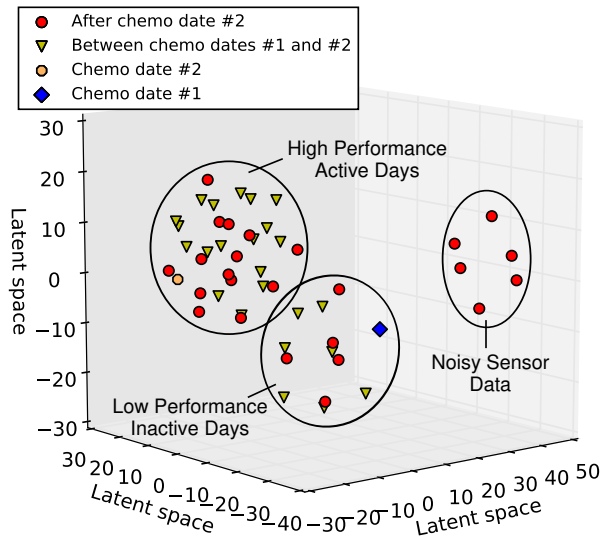
Figure 5: m-TSNE visualization of ATOM-HP dataset. **Figure 6:** m-TSNE visualization of EEG dataset.

4.3 Visualization Results

Figure 5 shows that the visualization of the MTS items of a subject undergoing treatment using m-TSNE. In this figure, there are 51 points corresponding to 51 days of chemotherapy treatment cycle. Each data point is labeled by the date order whether it is before or after a chemotherapy date in the study cycle. As can be seen, the figure shows that the 51 points form 3 clusters. To understand the three clusters, a 2-D display with annotation for each data point is provided in Figure 7. The annotation are in a format [**date number**]_[**step counts**] (step counts variable is chosen as it is easier to interpret human performance using this variable as compared to other variables). The figure was also shown to a health professional. Based on the health professionals' inputs, and the variables values, in Figure 5, these clusters can be interpreted as: (1) cluster of high performance / active days (the left-most cluster with a high number of step counts for each data point (more than 800)), (2) cluster of low performance / inactive days (the lower cluster with a low number of step counts for each data point), (3) cluster of noisy data (outliers) from the sensors (the right-most cluster) (noisy data appears to have abnormal values due to sensors). This is one of the insights obtained from our visualization approach that can help the healthcare professionals to understand the effect of the treatment using data collected from wearable sensors. Moreover, the figure also shows that the subject has low activities for a few days immediately following the two chemotherapy dates. This insight indicates that the subject may suffer from fatigue due to the chemotherapy session. We believe that these insights are quite helpful for oncologists to study their patient's activity performance, and also these might help them in designing better objective measures to quantify human performance.

For comparison, we provide the visualizations of the same subject using PCA, Euclidean-based t-SNE and DTW-based t-SNE methods in Figure 8, Figure 9, and Figure 10, respectively. Note, the star chart, parallel coordinates chart, and MTSA of the same subject's data are also shown in Figure 1, Figure 2, and Figure 3 respectively. It is clear that these visualizations do not provide clear insights about the patient's activities / performance status. Euclidean-based, and DTW-based t-SNE figures (Figure 9, Figure 10) show that the noisy sensor data points are close to each other. However, they do not give clear distinct clusters like m-TSNE. Based on these figures, it may be difficult to interpret and understand the data insights as these approaches show overlapping clusters which appear as a cloud of data points. PCA (Figure 8) might provide some insights of the subject performance along the axis of highest principle component, however as shown in the figure it does not form distinct clusters for outliers or for different levels of activity.

Figure 6 provides m-TSNE visualization for the EEG dataset. Each data point is one subject's MTS data performing a trial, and is labeled based on the subject category: *control* and *alcoholic*. The figure clearly depicts a manifold in which the points representing the control group lie inside, and are covered by alcoholic group. It also shows that all the outliers in our visualization belong to alcoholic group, providing
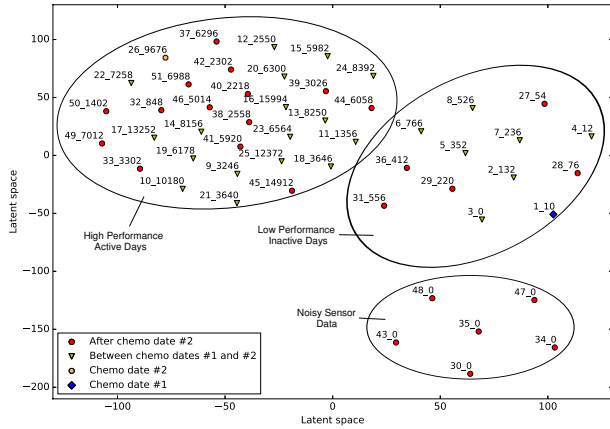
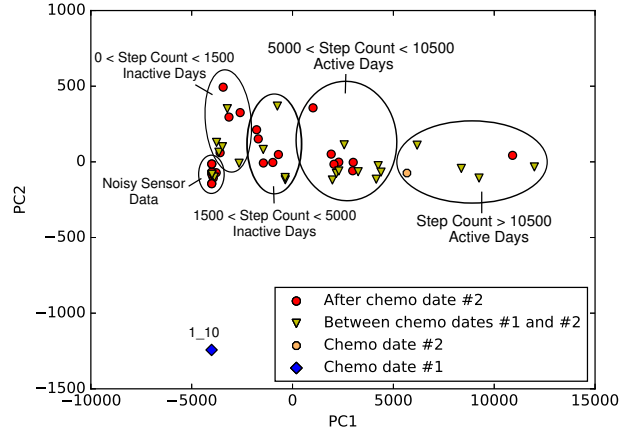**Figure 7:** m-TSNE 2-D visualization of ATOM-HP dataset with annotations.

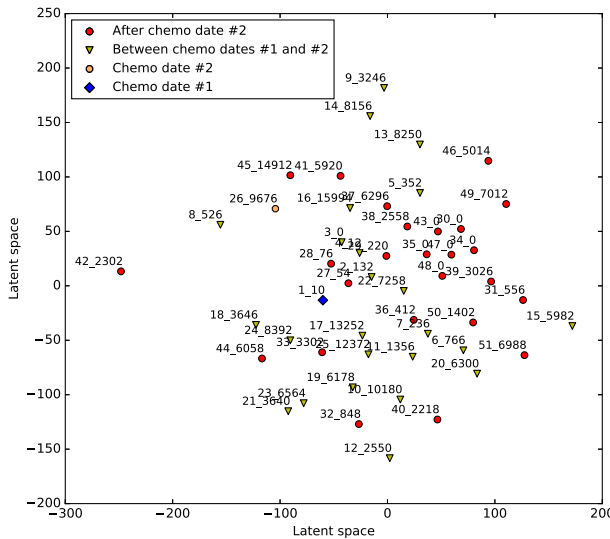**Figure 8:** PCA 2-D visualization of ATOM-HP dataset.

Figure 9: Euclidean-based t-SNE 2-D visualization of ATOM-HP dataset with annotations.
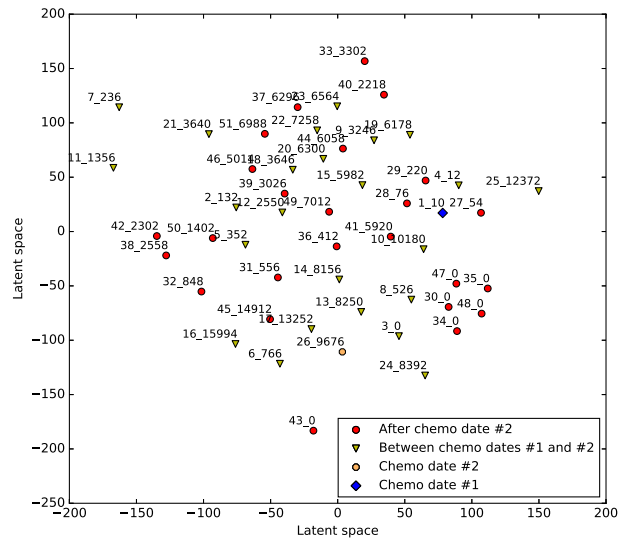

Figure 10: DTW-based t-SNE 2-D visualization of ATOM-HP dataset with annotations.

interpretable insights which are not extracted by PCA or the other MTS visualization techniques.

User Study: We evaluated the interpretability of three techniques: PCA, DTW-based t-SNE and m-TSNE by conducting a controlled user study with 6 non-healthcare professionals. We removed the names of the techniques, labels and color codes from the visualization results to avoid bias in users' interpretability. Each user was shown the visualizations of these three techniques on all subjects in ATOM-HP dataset and instructed to assign a score if they could find (interpretable) clusters, trends or outliers. For each technique, an user assigned a score (1 - the lowest score, 2 or 3 - the best score) in such a way that better techniques received higher scores. We aggregated the scores over all users and report out findings here: m-TSNE obtained the highest score of 2.48, PCA obtained a score of 1.92, and DTW-based t-SNE obtained the lowest score of 1.6. An oncologist was also included in our user study to verify our visualizations and insights. He agreed that the clusters found by our approach is very useful to study the patients' fatigue during their treatment cycle. This user study shows that m-TSNE could provide interpretability and insights when compared to the other competing methods.

5   Conclusion and Future Work

In this paper, we proposed m-TSNE: a framework to visualize high-dimensional MTS data. m-TSNE uses EROS to compute similarity between MTS data points and projects them to low-dimensional space for visualization. Empirical evaluation on two healthcare datasets showed that our approach provides interpretable insights via visualization while the other visualization methods which use PCA, Euclidean-based t-SNE, and DTW-based t-SNE, are more difficult to interpret. These insights could help healthcare professionals to evaluate their patients' performance. For future work, we plan to extend our work by building a tool for showing the visualization of the MTS dynamically.

References

[1] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: promise and potential," *Health Information Science and Systems*, vol. 2, no. 1, p. 1, 2014.

[2] J. J. Van Wijk and E. R. Van Selow, "Cluster and calendar based visualization of time series data," in *Information Visualization, 1999.(Info Vis' 99) Proceedings. 1999 IEEE Symposium on*. IEEE, 1999, pp. 4–9.

[3] A. Aris, B. Shneiderman, C. Plaisant, G. Shmueli, and W. Jank, "Representing unevenly-spaced time series data for visualization and interactive exploration," in *IFIP Conference on Human-Computer Interaction*. Springer, 2005.

[4] Y. Li, J. Lin, and T. Oates, "Visualizing variable-length time series motifs." in *SDM*. SIAM, 2012, pp. 895–906.

[5] C. Shi, W. Cui, S. Liu, P. Xu, W. Chen, and H. Qu, "Rankexplorer: Visualization of ranking changes in large time series data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2669–2678, 2012.

[6] K. Yang and C. Shahabi, "A pca-based similarity measure for multivariate time series," in *Proceedings of the 2nd ACM international workshop on Multimedia databases*. ACM, 2004, pp. 65–74.

[7] J. Heer, M. Bostock, and V. Ogievetsky, "A tour through the visualization zoo. retrieved may 17, 2013," 2010.

[8] J. M. Chambers, *Graphical methods for data analysis*, 1983.

[9] A. Inselberg, "The plane with parallel coordinates," *The visual computer*, vol. 1, no. 2, pp. 69–91, 1985.

[10] C. U. Lehmann and J. Fackler, "Visualizing multivariate time series data to detect specific medical conditions," 2008.

[11] V. West, D. Borland, and W. E. Hammond, "Visualization of ehr and health related data for information discovery," in *Workshop on Visual Analytics in Healthcare*, 2013.

[12] I. Jolliffe, *Principal component analysis*. Wiley Online Library, 2002.

[13] G. E. Hinton and S. T. Roweis, "Stochastic neighbor embedding," in *Advances in neural information processing systems*, 2002, pp. 833–840.

[14] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[15] X. L. Zhang, H. Begleiter, B. Porjesz, W. Wang, and A. Litke, "Event related potentials during object recognition tasks," *Brain Research Bulletin*, vol. 38, no. 6, pp. 531–538, 1995.

[16] D. Borland, V. West, and W. Hammond, "Multivariate visualization of system-wide national health service data using radial coordinates," in *Proc. Workshop on Visual Analytics in Healthcare*, 2014.

[17] W. Javed, B. McDonnel, and N. Elmqvist, "Graphical perception of multiple time series," *IEEE transactions on visualization and computer graphics*, vol. 16, no. 6, pp. 927–934, 2010.

[18] M. Nguyen, S. Purushotham, H. To, and C. Shahabi, "m-tsne: A framework for visualizing high-dimensional multivariate time series (technical report)," *University of Southern California, 2016. http://infolab.usc.edu/DocsDemos/mtsne.pdf*.

[19] O. Maimon and L. Rokach, *Data mining and knowledge discovery handbook*. Springer, 2005, vol. 2.

[20] M. Müller, "Dynamic time warping," *Information retrieval for music and motion*, pp. 69–84, 2007.

[21] W. Krzanowski, "Between-groups comparison of principal components," *Journal of the American Statistical Association*, vol. 74, no. 367, pp. 703–707, 1979.

[22] M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[23] M. M. Oken, R. H. Creech, D. C. Tormey, J. Horton, T. E. Davis, E. T. Mcfadden, and P. P. Carbone, "Toxicity and response criteria of the eastern cooperative oncology group." *American journal of clinical oncology*, vol. 5, no. 6, 1982.

[24] J. G. Snodgrass and M. Vanderwart, "A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity." *Journal of experimental psychology: Human learning and memory*, vol. 6, no. 2, 1980.