Human-Centred Visual Analytics

CSCM 27

Paper Summary

Joint t-SNE for Comparable Projections of

Multiple High-Dimensional Datasets



DESHAN CHATHUSANKA KOTAWILA WAGARACHCHIGE

2110572

Swansea University 23th November 2022

High-dimensional time series data visualisation has been a challenging research area. Most traditional techniques were explicitly designed for time-invariant datasets. However, time-dependent datasets can be independently projected in multiple plots; static clusters may be projected into different regions while inducing undesirable effects due to the stochastic nature of the algorithms. In contrast, the combined dataset can be visualised into a single plot; this approach is computationally expensive and visually inefficient. This research introduced high fidelity and consistent graphlet-based comparable projection algorithm to project time-dependent data evolution considering topological changes over time.

Temporal MDS projected high dimensional data into 1D and visualised it in a 2D time series plot, but 1D conversion can ignore more insightful information. T-SNE is a multivariate non-linear projection technique significantly faster than traditional MDS. The main objective of this method is to preserve the similarity between the gaussian probability distribution in higher dimensional space and the t-probability distribution in the projected space by minimising KL divergence. The main challenge of the t-SNE in time-series data visualisation is the misalignment of static clusters because of the stochastic nature and unawareness of topological changes in the higher dimensional space.

Dynamic t-SNE is a state-of-the-art time-dependent multivariate data visualisation method. It introduced an additional loss term to the KL divergence to minimise the inter-frame movement of clusters. It penalises the deviation of dynamic clusters, where the topology change in the original space, known as the smoothing. The topological nature of the current frame may affect the data embedding at a faraway time frame, known as long-range interference.

This method dynamically converts the dataset at each time frame into an undirected KNN graph. Then, the current graph and the one at the adjacent time frame are compared using Graphlet Frequency Distributions(GDF) around each node to recognise topological changes. The point similarity combines the cosine similarity of GDF and the proportion of common KNN to preserve the same topology and neighbours around similar nodes in the lower dimensional space, respectively. The algorithm introduced vector constraints to the objective function to maintain consistency between static data points based on edge similarity. This research used a uniform sampling-based graphlet-generating algorithm, MCMC, to alleviate the computational complexity of traditional approaches.

This research evaluated the algorithm qualitatively and quantitatively. Gaussian synthetic dataset and two real-world datasets were selected to assess fidelity preserving external validity. JT performed faithfully with synthetic datasets under different transformations, whereas DT suffered from smoothness under structural changes. The research found that the JT was outstanding with middle-range k values during the graph population stage. The second evaluation was done with the first five-digit images in the MNIST dataset executing two replacement transformations: zero with nine and one with three. JT positioned the original and replaced 3-digit clusters together and the original 4-digit and replaced 9-digit closer, keeping the 2-digit cluster unchanged. DT projected the original and replaced 3-digit clusters partially closed due to the smoothness effect, and the original 4-digit and replaced 9-digit clusters closer even before the transformation due to long-range interference. The proposed method preserves visual consistency between static clusters and projects topologically changed clusters faithfully. This algorithm is computationally efficient due to the size two sliding window mechanism.

Barnes-Hut t-SNE can improve the time complexity many folds. Moreover, loop-based parallelism can enhance the processing power, and the estimation framework can improve the graphlet continuing process. Finally, the GFD-based vector constraint can be integrated with other t-SNE families: MDS, UMAP, and Hierarchical Projection.

# References

[1] Wang, Y., Chen, L., Jo, J., & Wang, Y. (2021). Joint t-sne for comparable projections of multiple high-dimensional datasets. *IEEE Transactions on Visualization and Computer Graphics*, *28*(1), 623-632.

[2] Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, *9*(11).

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

[4] López, D., Oehlberg, L., Doger, C., & Isenberg, T. (2015). Towards an understanding of mobile touch navigation in a stereoscopic viewing environment for 3D data exploration. *IEEE Transactions on Visualization and Computer Graphics*, *22*(5), 1616-1629.