Human-Centred Visual Analytics

CSCM 27

Paper Summary of Joint T-SNE



DESHAN CHATHUSANKA KOTAWILA WAGARACHCHIGE

2110572

Swansea University 23th November 2022

High-dimensional time series data visualisation has been a challenging research area in this data-driven era. Most traditional techniques, such as t-SNE and UMAP, were explicitly designed for time-invariant datasets. However, such multidimensional projection techniques can visualise time-variant datasets in two straightforward ways. Datasets can be independently projected in multiple plots; time-invariant 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 with time-based colour coding. But, this approach is computationally expensive and visually inefficient(visual clutters). This research introduced high fidelity and consistent graphlet-based comparable projection algorithm to visualise 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 this technique can ignore more insightful information due to 1D conversion. One of the most popular robust non-linear multivariate projection techniques is t-SNE, which is 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 state-of-the-art of time-dependent data visualisation by exploiting conventional t-SNE and introducing an additional loss term to minimise the inter-frame movement of data points. In contrast, this loss term penalises the inter-frame deviation of data points, albeit a topology change in the original space during the time concerned, known as the smoothing effect. In addition, this technique projects whole datasets simultaneously, introducing several significant issues. 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. This algorithm evades long-range interference due to concerns about only adjacent time frames for data embedding processing. 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 preserve consistency between static edges based on edge similarity, which is the multiplication of corresponding point similarities. This research used a uniform sampling-based graphlet-generating algorithm, MCMC, to alleviate the computational complexity of traditional approaches.

(results/ pros and cons)

The algorithm was qualitatively evaluated using meaningful data transformations and quantitatively with three metrics: Local Coherence Error, KNN preservation, and KL divergence. This research selected one synthetic dataset and two real-world datasets with meaningful data transformations to evaluate fidelity preserving external validity. Regarding the synthetic dataset, JT had high fidelity under different transformations in three different time frames, whereas DT suffered from smoothness under structural changes, particularly overlapping and splitting. In addition, JT qualitatively outperformed DT with all evaluation metrics except KL divergence due to extra loss term. Moreover, the research found that the JT algorithm was outstanding with middle-range k values during the graph population stage.

# Future Works

(paper + own suggestions)

# References