Graph-based Stochastic Neighbor Embedding for Robust Data Visualisations

Tobias Wängberg (speaker), Chun-Biu Li, and Joanna Tyrcha

Stockholm University

Contact: tobias@math.su.se

April 15 2021

The t-distributed Stochastic Neighbour Embedding (tSNE) algorithm has emerged as one of the leading methods for visualising High Dimensional (HD) data in a wide variety of fields, especially for revealing cluster structure in HD single cell transcriptomics data. However, several shortcomings of the algorithm have been identified. Specifically, tSNE is often unable to correctly represent hierarchical relationships between clusters and spurious patterns may arise in the visualisations due to incorrect hyper-parameter settings, which could mislead the researcher. In this work we propose to combine tSNE with shape-aware graph distances to mitigate some of the limitations of the original method. We show the advantage of the graph based algorithm on simulated data sets, where we see a significant improvement in visualizing imbalanced and non-linear clusters, as well as preservation of hierarchical structure, based on quantitative validation indices. Moreover, we propose a particular hyper-parameter setting, different from previously suggested settings, which we find consistently optimal across all the test cases conducted in this work. Lastly, we demonstrate the superior performance in the visualisations of the MNIST image data set as well as single cell transcriptomics gene expression data.

Keywords: High dimensional statistics, tSNE, data visualisation, graph distances, single cell transcriptomics, MNIST, dimensionality reduction evaluation