A Tool for Visualizing and Analyzing High-Dimensional Clustering Performance

Justin Lin*and Julia Fukuyama†

Abstract

Technological advances have spurred an increase in data complexity and dimensionality. We are now in an era in which data sets containing thousands of features are commonplace. To digest and analyze such high-dimensional data, dimension reduction techniques have been developed and advanced along with computational power. Of these techniques, nonlinear methods are most commonly employed when working with high-dimensional data because of their ability to construct visual two-dimensional representations of high-dimensional data. These methods unevenly stretch and shrink space in a way that represents the data's structure in a fewer number of dimensions. However, attempting to capture high-dimensional structures in a significantly lower number of dimensions requires drastic manipulation of space. As such, nonlinear dimension reduction methods are known to occasionally capture false structures, especially in noisy settings. In efforts to deal with this phenomenon, we developed an interactive tool that enables researchers to better understand and diagnose their dimension reduction results. It uses various analytical plots to provide a multi-faceted perspective on captured structures in the data to determine if they're faithful to the original data or remnants of the dimension reduction process. The tool is available in an R package named *insert name here*.

1 Introduction

The potency of nonlinear dimension reduction methods lies in their flexibility, allowing them to model complex data structures. That same flexibility, however, makes them difficult to use and interpret. Each method requires a slew of hyperparameters that need to be calibrated, and even when adequately calibrated, these methods require a trained eye to interpret. For example, the two most popular nonlinear dimension reduction methods, t-SNE and UMAP, are known to generate unintuitive results ([1], [6]). The results often cluster, even when no clusters exist in the data. Moreover, cluster sizes and inter-cluster distances can be unreliable. We've developed an interactive tool that researchers may use to conduct a post-hoc analysis when applying high-dimensional clustering. The main goal of the tool is provide an additional perspective on inter-cluster relationships, making it easier to distinguish structures faithful to the signal from the noise.

2 Methods

Suppose our goal is to cluster a high-dimensional data set $Z \in \mathbb{R}^{n\times p}$, containing n points in p dimensions. To visualize the clustering, a two-dimensional representation $X \in \mathbb{R}^{n\times 2}$ can be calculated using a dimension reduction method of choice. The tool takes as input a dissimilarity matrix $D \in \mathbb{R}^{n\times n}$, representing the high-dimensional inter-point dissimilarities; the low-dimensional representation X; and the clustering. Optionally, the user may also provide a set of ID's to label the points. The high-dimensional dissimilarity matrix is used to construct a minimum spanning tree meant to model the global structure of the high-dimensional data.

2.1 The Minimum Spanning Tree

Manifold learning is the most popular approach to nonlinear dimension reduction in which the data is assumed to be uniformly drawn from a high-dimensional manifold embedded in p dimensions. These

*Department of Mathematics, Indiana University
†Department of Statistics, Indiana University

methods often make use of graphs to estimate the population manifold. For example, IsoMap, the nonlinear extension of multidimensional scaling (MDS), uses ϵ-neighborhood and k-neighborhood graphs [3]. Hartigan's MAP test for uses minimal ascending path spanning trees to detect multimodality [5]. More modern examples also make use of graphs. [2] introduces the maximum information spanning tree and [4] uses the minimum spanning tree (MST) for dimension reduction purposes.

We've opted for the MST for

References

[1] Andy Coenen and Adam Pearce for Google PAIR. Understanding UMAP. *https://pair-code.github.io/understanding-umap/.*

[2] Bracken M. King and Bruce Tidor. MIST: Maximum information spanning trees for dimension reduction of biological data sets. *Bioinformatics 25:9, 1165-1172*, 2009.

[3] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science 290:2319*, 2000.

[4] Daniel Probst and Jean-Louis Reymond. Visualization of very large high-dimensional data sets as minimum spanning trees. *Journal of Cheminformatics 12:12*, 2020.

[5] Gregory Paul M. Rozál and J.A. Hartigan. The MAP test for multimodality. *Journal of Classification 11, 5-36*, 1994.

[6] Martin Wattenberg, Fernanda Viégas, and Ian Johnson. How to Use t-SNE Effectively. *Distill*, 2016.