Paper Review (UMAP)

Title: Dimensionality reduction for visualizing single-cell data using UMAP

Author: Etienne Becht1, Leland McInnes2, John Healy2, Charles-Antoine Dutertre1, Immanuel W H Kwok1, Lai Guan Ng1, Florent Ginhoux1 & Evan W Newell1, 3

Motivation: The motivation of this paper is to evaluate the performance of a new dimensionality reduction (DR) algorithm called UMAP through rigorous testing versus other established algorithms. Given the high dimensional/complex data in today's world, scientists are racing to find better ways to understand them. UMAP will hopefully be the next stepping stone in that search for a better technique.

Summary: The research paper convers a new DR technique, UMAP, also known as uniform manifold approximation and projection. Initially the authors compare it to t-SNE, the go-to DR used by many scientists and researchers in recent times. They found both algorithms to be well suited for clustering similar cell populations. In the end, UMAP was able to better maintain continuity than t-SNE.

The paper also compares t-SNE and UMAP by applying both algorithms to a mass cytometry dataset and a scRNAseq dataset. They found UMAP was able to better represent the continuity of cell phenotypes in hematopoiesis. UMAP did this by revealing a five branched structure that was consistent with hematopoietic differentiation. While t-SNE also revealed a similar pattern, it was less clear according to the researchers. UMAP was also able to preserve the density of clusters in the reduced space more uniformly than t-SNE.

The authors decided to dive into the speed of UMAP versus t-SNE, FIt-SNE, FIt-SNE l.e., scvis, and PCA. After running tests on high dimensional datasets, the researchers found UMAP is faster than current standards for DR and can better preserve local and global structure. They also found UMAP can compete with Fit-SNE for large datasets.

Approach and contributions: The authors used data to test different algorithms. They tested how well they could display high dimensional data, how fast they were, and how well they preserved local and global structures. The main take away from the paper is that UMAP is able to better represent preserve data structure compared to current standards like t-SNE and FIt-SNE, while also being faster and more reproducible. UMAP will be extremely important to machine learning as it will allow models to deal with high dimensional data more effectively by preserving data structure and being faster than most current algorithms. The paper builds on previous DR techniques such as t-SNE. The authors applied UMAP along with many other DR algorithms to investigate the performance in ways mentioned above.

Areas for improvement: One potential weakness in the methods section of this paper is the authors using the default value as the hyper parameter in the Phenograph clustering algorithm. This could affect the results of the separability of cell populations benchmark. To further this paper, the authors could test the performance of UMAP on other types of datasets in different fields or on different applications. This would cement UMAP as one of the best current DR algorithms.