Approximative Linear t-SNE using k-Means

Sonja Biedermann,
01402891

29. März 2021

Dimensionality reduction is the transformation of high dimensional data into a lower dimensional space while keeping the interesting parts of the data intact. Ideally, the low dimensional so-called embedding reveals structure that would have been much more difficult to detect in the high dimensional space.

There are many different approaches to dimensionality reduction, and the most common taxonomy separates them into linear and non-linear methods. Linear methods transform the high dimensional data solely by applying linear transformations, such as PCA, while non-linear methods are essentially free to transform the data by any means.

In this thesis, we will be focusing on a non-linear method called t-SNE, which is a variant of Stochastic Neighborhood Embedding using the Student-t distribution. t-SNE has become popular due to the visual properties of the embeddings it produces: well-separated clusters consisting of similar items. It is, however, a stochastic method, non-parametric in its original form, possesses a difficult to optimize non-convex objective function and also has a prohibitively high runtime complexity of $\mathcal{O}(n^2)$.

Methods have been devised that target the issue of t-SNE's computational cost. Speeding it up is done by approximating the computation of various parts of the algorithm, or by formulating a different objective function that is easier to compute. The latter we will regard as methods that are similar to t-SNE, but not strictly variants. Approximations include Barnes Hut t-SNE, which uses trees to achieve a runtime complexity of $\mathcal{O}(n \log n)$, and FIt-SNE, which uses polynomial interpolation and the fast Fourier transform with a computational cost of $\mathcal{O}(n)$. A more distant linear cousin of t-SNE, UMAP, is also considered more closely in our thesis, due to its popularity and success in the biological field.

The main contribution of this thesis is developing yet another variant of t-SNE based on using k-Means to summarize data and speed up computation. We also rely on locality sensitive hashing for finding (approximate) nearest neighbors.

We thoroughly evaluate our method on multiple data sets using the average k-nearest neighbor label purity and average absolute Pearson correlation coefficient between high and low dimensional distances and compare it against the other aforementioned methods.