World Academy of Science, Engineering and Technology
International Journal of Computer and Systems Engineering
Vol:15, No:8, 2021

Index t-SNE: Tracking Dynamics of High-Dimensional Datasets with Coherent Embeddings

Authors : Gaelle Candel, David Naccache

Abstract : t-SNE is an embedding method that the data science community has widely used. It helps two main tasks: to display results by coloring items according to the item class or feature value; and for forensic, giving a first overview of the dataset distribution. Two interesting characteristics of t-SNE are the structure preservation property and the answer to the crowding problem, where all neighbors in high dimensional space cannot be represented correctly in low dimensional space. t-SNE preserves the local neighborhood, and similar items are nicely spaced by adjusting to the local density. These two characteristics produce a meaningful representation, where the cluster area is proportional to its size in number, and relationships between clusters are materialized by closeness on the embedding. This algorithm is non-parametric. The transformation from a high to low dimensional space is described but not learned. Two initializations of the algorithm would lead to two different embeddings. In a forensic approach, analysts would like to compare two or more datasets using their embedding. A naive approach would be to embed all datasets together. However, this process is costly as the complexity of t-SNE is quadratic and would be infeasible for too many datasets. Another approach would be to learn a parametric model over an embedding built with a subset of data. While this approach is highly scalable, points could be mapped at the same exact position, making them indistinguishable. This type of model would be unable to adapt to new outliers nor concept drift. This paper presents a methodology to reuse an embedding to create a new one, where cluster positions are preserved. The optimization process minimizes two costs, one relative to the embedding shape and the second relative to the support embedding' match. The embedding with the support process can be repeated more than once, with the newly obtained embedding. The successive embedding can be used to study the impact of one variable over the dataset distribution or monitor changes over time. This method has the same complexity as t-SNE per embedding, and memory requirements are only doubled. For a dataset of n elements sorted and split into k subsets, the total embedding complexity would be reduced from $O(n^2)$ to $O(n^2=k)$, and the memory requirement from n^2 to $2(n=k)^2$, which enables computation on recent laptops. The method showed promising results on a real-world dataset, allowing to observe the birth, evolution, and death of clusters. The proposed approach facilitates identifying significant trends and changes, which empowers the monitoring high dimensional datasets' dynamics.

Keywords : concept drift, data visualization, dimension reduction, embedding, monitoring, reusability, t-SNE, unsupervised learning
Conference Title : ICBDVA 2021 : International Conference on Big Data Visual Analytics
Conference Location : Venice, Italy
Conference Dates : August 12-13, 2021