Cosine Similarity based TopoMap

Devarsh Patel
dp3324@nyu.edu

Master of Science in Computer Science,
NYU Tandon School of Engineering

May 6, 2022

Track: Track 1

Github: https://github.com/dp3324/advance-topomap

Abstract

The main goal of this project is to take TopoMap[1] one step further in terms of information computation complexity and space efficiency. It will answer many unresolved bottlenecks emerged during the initial iteration in the previous work. The point of interest for designing and developing this advanced TopoMap[1] solution is to answer the training problem that arise in Machine Learning models because of High-Dimensional data. While using polynomial data with multiple regression model, It requires more computation time along with intensive space utilization. This project aims to minimize the dimension of the data while preserving the model accuracy and precision. The advancement of this technology can help data engineers to identify and modify the data cloud as per the requirement so that they can focus on whats more relevant that something that is unimportant.

1 Introduction

TopoMap[1] is an advanced Multidimensional Projection (MDP) technique used for high-dimensional data, that ensures to maintain the topographical definition of data. Dealing with high-dimensional data create complex computation intensity with the increase size in dataset. To over-come this bottleneck, it is always preferred to only consider the dimension, i.e. parameters, of the dataset which has the most impact on the predicted output. This means it is more suitable to remove inert parameters or dimension of the dataset to minimize the computation time. Here where TopoMap[1] comes, It provides better visualizations of introduced data by decreasing d-dimensional data to 2D Space vector. TopoMap[1] is an advanced implementation of sophisticated projection techniques like Classical MDS, IsoMap, tSNE and UMAP. Most of the Multidimensional projection technique uses geographical distance as a base to define the relationship between the data points, which often results into false clustering as many times the data points in same coordinate region have significant topographical variance. Multidimensional scaling(MDS) is term coined for techniques which reduces data dimension by significant proportion. In context of projection, if a technique reduces data dimension to a point where it can be represented in 2D or 3D cartesian space, then that is called as Multidimensional Projection (MDP).

This research paper focuses on two main aspects of original TopoMap alogrithm: Euclidian Minimum Spanning Tree, and Convex Hull Alignment Algorithm. The proposed technique here uses Cosine Similarity based Minimum Spanning Tree instead of Euclidian distance. To accomodate the changes and the data type, Tree based hull alignment algorithm is proposed here. This algorithm varies from the convex hull alignment algorithm in terms of point placement in the component.

2 Literature Survey

There are many data dimensionality reduction algorithm designed to preserve the relationship between the points but mos to them are based on the parameters provided by

2.1 Topological data analysis

2.2 Topology-based Multidimensional Projection

3 Methodology

4 Results & Discussion

5 Future Scope

6 Conclusion

References

[1] In: ().