Cosine Similarity based TopoMap

Devarsh Patel
dp3324@nyu.edu

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

May 7, 2022

Track: Track 1

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

Abstract

The main goal of this project is to take TopoMap[**harishd**] 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[**harishd**] 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[**harishd**] 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[**harishd**] comes, It provides better visualizations of introduced data by decreasing d-dimensional data to 2D Space vector. TopoMap[**harishd**] 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 most to them are based on the parameters dependents on the geographical terms. They utlizes distance based matrices like Euclidian distance to define the co-relation. They all are

efficient in their own terms but comes with many limitations like leaving n-level simplexes. Many of this algorithms certainty decreases as the size of the data increases.

2.1 Multidimensional Scaling

Multidimensional Scaling or MDS is the simplest means of visualizing the similarities between the cases in the dataset. MDS uses the distance between n objects and define the cluster based on them. MDS places the n-dimensional data in 2D cartesian space such that it maintains the distance between the datapoints. There are mainly 4 type of Multidimensional Scaling algorithms.

2.1.1 Classicial Multidimensional scaling

Classical multidimension scaling also known as Principle Component Analysis (PCA), Torgerson Scaling or Torgerson–Gower scaling, is an algorithm which takes the input matrix giving the dissimilarities between the pairs of items and place them such that the loss function or strain is minimized. The strain can be defines as follows:

$$Strain_D(x_1, x_2....x_n) = \left(\frac{\sum_{i,j}(b_{ij} - x_i^T x_j)^2}{\sum_{i,j} b_{ij}^2} \right)^{1/2}$$

where x_i are the N-dimensional space vectors, $x_i^T x_j$ defines the scalar product between x_i and x_j and b_{ij} are the element of matrix B.

2.1.2 Metric Multidimensional scaling

Metric Multidimensional scaling is similar to the Classical Multidimensional scaling but it tries to minimize the Strain in the space instead of the stress.

$$Stress_D(x_1, x_2....x_n) = \sqrt{\sum_{i \neq j = 1,..,N} \left(d_{ij} - \|x_i - x_j\|\right)^2}$$

2.1.3 Non-metric multidimensional scaling

In contrast to non classical and metric multidimensional scaling, Non-metric MDS finds both the non-parametric monotonic between the data points and the Euclidian distance between them. This data dimensionality reduction algorithm is based on isotonic regression.

For x denote the vector of proximities, $f(x)$ a monotonic transformation of x, and d as the point distance, then the coordinate of the points in 2D cartesian are found by minimizing the stress as follows:

$$Stress = \sqrt{\frac{\sum \left(f(x) - d\right)^2}{\sum d^2}}$$

2.1.4 Generalized multidimensional matrix

This method is extension of metric multidimensional scaling which utilizes the non-Euclidian space. It allows finding the minimum-distortion embedding of one surface into another.

2.2 Topology-based Multidimensional Projection

Topology-based methods exist from a very long time when it comes to visualizing data. It has played an important role in many visualization prone domains like astrophysics, medical imaging, robotics and material science. The described TopoMap is based on the Rips[**rips**] filtration technique. Rips filtration technique is used to topographically analyse the high-dimensional data point which also efficiently capture the homology of the manifold sampled by this data points. It also correlates to the Reeb Graph[**reedgraph**], which contract the cluster of similar topological definition to a single point, which result into skeleton like visualization. When it comes to discrete point cloud visualization, the Mapper[**mapper**] is an better approach which is based on Reed[**reedgraph**] approximation for the nearest neighbour graph.

2.3 Topology-based Multidimensional Projection

TopoMap[**harishd**] provides an in-depth comparison to multiple Multidimensional Scaling(MDP)/Multidimensional Projection(MDP) techniques. While most of them works on geodesic distance between the data points, especially conventional Isomap, TopoMap[**harishd**] works on homological difference between the points. IsoMap, tSNE and UMAP have many improved variation but they all does not preserve 0-cyclic groups. An Isomap variant proposed by Lee and Verleysen[**leeverleysen**], tears the high-dimension data into non-contractable loops which preserves manifold unfolding.

The work represented here, by considering Rips Filtration[**rips**], is focused on analysing topographical distance. The TopoMap is compared to the terrain metaphors by the separation point of preserving the topological angle. It is guarranted that TopoMap preseves the connected component during filtration and result into identical 0-cyclic homology presistance diagram from the produced result and original data.