Topological Data Analysis

Deepak Choudhary(11234) & Samarth Bansal(11630)

Abstract

Data Scientists need to make sense out of and classify data which is complex either in terms of the sheer size of data, or due to the involvement of a large number of characteristics. There are various methods of Machine Learning for this purpose. Most of these methods make some assumption about the data. In this project we try to study and implement Topological Data Analysis which tries to extract meaning from data by creating and studying the shape of the data without making assumptions about the data.

1. Introduction

In our daily life we come across various data sets in the form of huge array of numbers in almost every field. How do we understand these complex data sets. Earlier methods started by asking very specific questions related to the data. But these methods lack to capture some kinds of inherent information.

1.1. Problem Definition

Given a large/complex data set, classify it or get insight into the data and its behaviour. And if applicable compare the results with existing methods such as PCA and MDS.

1.2. Motivation

Topological Data Analysis is sensitive to both large and small scale patterns that often fail to be detected by other analysis methods, such as principal component analysis, (PCA), multidimensional scaling, (MDS), and cluster analysis. PCA and MDS produce unstructured scatterplots and clustering methods produce distinct, unrelated groups. These methodologies sometimes obscure geometric features that topological methods capture.

1.3. Key Ideas

There are three key ideas of topology that make extraction of patterns via shape possible.

1.3.1. It studies shapes in a coordinate free way.

1.3.2. It studies the properties of shapes that are invariant under ‘‘small’’ deformations.

1.3.3. Third idea is that of compressed representations of shapes. Topology deals with finite representations of shapes called triangulations, which means identifying a shape using a finite combinatorial object called a simplicial complex or a network.

2. Approach

2.1. Given a data set and three inputs:

2.1.1. a distance metric

2.1.2. one or more filter functions(real valued quantities associated to the data points)

2.1.3. and two resolution parameters(“resolution” and “percent overlap”)

we construct a network of nodes with edges between them.

2.2. The nodes represent similar data and the connection between them is represented by edges and instead of a unorganised massive structure we get a shape or network. We can now use our visual techniques to observe patterns in the data. These patterns are what we mean when extracting knowledge.

3. Resources

3.1. References

3.1.1. Lum, P.Y.et al. Extracting insights from the shape of complex data using topology. Sci. Rep. 3, 1236; DOI:10.1038/srep01236 (2013).

3.1.2. Gunnar Carlsson,2009, Bulletin (New Series) of The American Mathematical Society, Volume 46, Number 2, April 2009, Pages 255–308

3.2. Data Sets

3.2.1. G.M. Reaven and R.G. Miller, Diabetologica 16:17-24 (1979)

3.2.2. Breast Cancer Data Set (Supplementary Table 1 http://www.nature.com/nature/journal/v415/n6871/suppinfo/415530a.html)