Background Reading for Topological Data Analysis

Topological data analysis (TDA) is a methodology for applying the mathematical study of shape (topology) to the study of large and complex data sets. Its development was initiated around the year 2000. It has now seen applications in a large number of varied domains. The purpose of this document is to give a brief summary of the literature and outline of the directions in which the methodology has been applied.

There are two distinct threads in the subject. One is the mapping thread, which produces shapes encoded as graphs or simplicial complexes on which one can operate directly. The second thread is the shape measurement thread, in which one can construct invariants of shapes that can be used to obtain understanding of data as well as to coordinatize various kinds of unstructured data. Both themes appear prominently in applications.


Mapping

This area is concentrated around one construction, given in [1] and [11], which is a generalization of the notion of Reeb graphs in computational geometry. It takes as input a point cloud (a data set equipped with a dissimilarity measure), and produces as output a network or graph in the computer science sense, which we refer to as the topological model. These graphs can be taken as maps describing the “similarity landscape” of the data. This kind of topological summary has been extremely useful in various applications. There is a theory developing around the stability of the construction, which will simplify and strengthen the ability to perform inference with it. Examples of this kind of work are [31] and [32]. In addition, the topological model as constructed by Ayasdi includes a great deal of functionality beyond the simple display of the map, which permits the development of models and applications deriving from the topological analysis, as well as inference of various kinds. Explicit comparisons with some other methods for unsupervised analysis have been performed in [21] in the context of hyperspectral imaging. The published research has been concentrated in the biomedical realm, although the range of applications is growing beyond that. The applications include work in numerous areas, enumerated below with the corresponding references.
Shape Measurement

The work here is concentrated around persistent homology. This method is an adaptation of the homology signatures in standard algebraic topology, that are able to capture the presence of various kinds of patterns in shapes. A survey of the method is given in [10]. The output of this method is a signature called a barcode (or equivalently, a persistence diagram), which is somewhat analogous to the dendrograms produced by hierarchical clustering, but which capture higher order properties of a shape. There are two directions of applications of these signatures. One is the measurement of the overall shape of given data sets. This kind of application is carried out in [13] (viral evolution), [22] and [42] (image processing), and [35] and [36] (neuroscience). Applications of this kind of analysis includes the construction of compression schemes [26] and coordinatization of texture data [28]. The second direction is the coordinatization of unstructured data, such as databases of molecules or images. This kind of application is exemplified by the work described in [33] and [43], and has been shown to be quite effective in drug discovery and in the materials science of glasses and other materials. The technical background for both methods are given in [1], [10], [29], [30], and [40]. There is considerable theoretical work concerning stability properties of persistent homology, for example in [23] and [24]. The work describing coordinatization methods based on persistent homology is exemplified in [38] and [39].


