Mapper : A Topological Mapping Tool for Point Cloud Data

Gurjeet Singh[1] Facundo Mémoli[2] Gunnar Carlsson[2]

[1]ICME, Stanford University, California, USA
[2]Department of Mathematics , Stanford University, California, USA

We introduce a new method for the qualitative analysis, simplification and visualization of high dimensional data sets, as well as the qualitative analysis of functions on these datasets. In many cases, data coming from real applications is massive and it is not possible to visualize and discern structure even in low dimensional projections. We propose a method which can be used to reduce high dimensional datasets into simplicial complexes with far fewer points which can capture topological and geometric information at a specified resolution. We refer to our method as `Mapper` in the sequel. Our construction provides a coordinatization of the data by providing a discrete and combinatorial object, a simplicial complex, to which the dataset maps and which can represent the dataset in a useful way.

In the simplest case, `Mapper` begins with a dataset X, which is regarded as a discrete metric space, and a continuous real valued function $f : X \to \mathbb{R}$, to produce a graph. This function can be a function which reflects geometric properties of the dataset, such as the result of a density estimator, or can be a user defined function, which reflects properties of the data being studied. In the first case, one is attempting to obtain information about the qualitative properties of the dataset itself, and in the second case one is trying to understand how these properties interact with interesting functions on the dataset. Now, the basic idea behind `Mapper` can be referred to as *partial clustering*, in that a key step is to apply standard clustering algorithms to subsets of the original dataset, and then to understand the interaction of the partial clusters formed in this way with each other. That is, if U and V are subsets of the dataset, and $U \cap V$ is non-empty, then the clusters obtained from U and V respectively may have non-empty intersections, and these intersections are used in building a simplicial complex.

In more detail, assume $f : X \to \mathbb{R}$ is given together with an open covering $\{U_\alpha\}_{\alpha \in A}$ of its range, for a finite index set A. Then $\{f^{-1}(U_\alpha)\}_{\alpha \in A}$ forms an open covering of X. For each α we consider the decomposition of $f^{-1}(U_\alpha)$ into its path connected components: $f^{-1}(U_\alpha) = \cup_{i=1}^{j_\alpha} V(\alpha, i)$ where j_α is the number of connected components of U_α.

Let $W = \{W_\beta\}_{\beta \in B}$ be the set of all $V(\alpha, i)$ for all $1 \leq i \leq j_\alpha$ and $\alpha \in A$. Note that $|B| = \sum_{\alpha \in A} j_\alpha$.

Next we construct the *nerve* $\mathcal{N}(W)$ of the covering W, i.e. the simplicial complex with vertex set B and where a family $\{\beta_0, \ldots, \beta_k\}$ spans a k-simplex in $\mathcal{N}(W)$ if and only if $W_{\beta_0} \cap W_{\beta_1} \cap \ldots \cap W_{\beta_k} \neq \emptyset$. The practical counterpart of the operation of *decomposing a set into its path connected components* is of course clustering.

We show two practical examples of application of `Mapper` in the figure (refer to the caption for details).

Instead of using only a single function to explore the data, one can use multiple functions. The functions determine the space to which we produce a map. The method can easily be modified to deal with maps to parameter spaces other than $\mathbb{R}$, such as $\mathbb{R}^2$ or the unit circle S^1 in the plane. In the first of these cases, one produces a two dimensional simplicial complex, together with a natural map from the dataset to it. In the second case, one constructs a graph with a map from the graph to a circle. In the case where the target parameter space is $\mathbb{R}$, our construction amounts to a generalization of the *Reeb graph* (see [2]) associated with the filter function. If the covering of $\mathbb{R}$ is too coarse, we will be constructing an image of the Reeb graph of the function, while if it is fine enough we will recover the Reeb graph precisely.

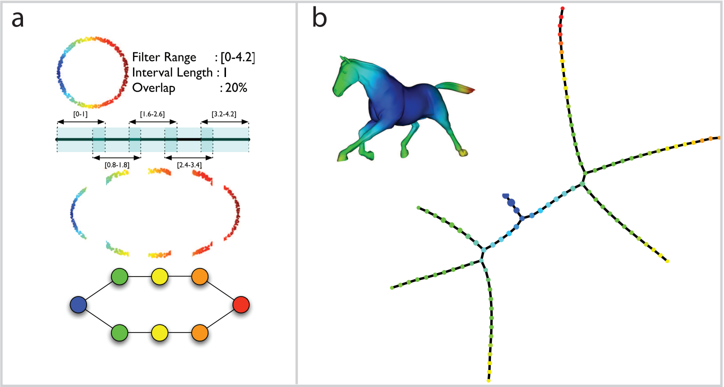

Figure 1: (a) The data is sampled from a noisy circle, and the function used is $f(x) = ||x - p||_2$, where p is the leftmost point in the data. The dataset is shown on the top left, colored by the value of the function. We divide the range of the function into 5 intervals which have length 1 and a 20% overlap. For each interval we compute the clustering of the points lying within the domain of the filter restricted to the interval, and connect the clusters whenever they have non empty intersection. At the bottom is the simplicial complex which we recover whose vertices are colored by the average function value. (b) Application of `Mapper` to shape simplification. The horse shape is colored with the function f which we choose to be the average geodesic distance to all other points on the shape (blue is low and red is high). On the right we show the output of `Mapper`, which in this case is a graph, its vertices being colored by the average function value of the points in the corresponding cluster.

This construction produces a “multiresolution” or “multiscale” image of the dataset. One can actually construct a family of simplicial complexes (graphs in the case of a one-dimensional parameter space), which are viewed as images at varying levels of coarseness, and maps between them moving from a complex at one resolution to one of coarser resolution. This fact allows one to assess the extent to which features are “real” as opposed to “artifacts”, since features which persist over a range of values of the coarseness would be viewed as being less likely to be artifacts.

We do not attempt to obtain a fully accurate representation of a dataset, but rather a low-dimensional image which is easy to understand, and which can point to areas of interest. Note that it is implicit in the method that one fixes a parameter space, and its dimension will be an upper bound on the dimension of the simplicial complex one studies. As such, it is in a certain way analogous to the idea of a Postnikov tower or the coskeletal filtration in algebraic topology [1].

In the practical implementation of the procedure one must deal with some clustering algorithm. Our procedure is not tied to any particular choice. In our experiments, however, we choose to work with single linkage clustering. Regardless of the clustering algorithm one picks, it is always necessary to estimate certain parameters (thresholds) that will ultimately tell us the number of clusters present in the data. We propose a procedure that exploits the spatial coherence of the scales in the partial datasets $f^{-1}(U_\alpha)$ in order to obtain a consistent choice of parameters.

References

[1] Allen Hatcher. *Algebraic topology.* Cambridge University Press, Cambridge, 2002.

[2] Georges Reeb. Sur les points singuliers d’une forme de Pfaff complètement intégrable ou d’une fonction numérique. *C. R. Acad. Sci. Paris*, 222:847–849, 1946.