LOCALITY PRESERVATION IN MANIFOLDS TO REDUCE DIMENSIONALITY

Kenneth Heafield
Mentor: Steven Low

Consider a finite set of points in an n-dimensional metric space (i.e. $\mathbb{R}^n$) that have some form to them. We suppose that the points were sampled from a lower dimensional manifold with some error. The problem is to identify a manifold that fits the points without imposing a particular functional form. For any finite set of points there exist infinitely many manifolds fitting perfectly so it is useful to impose restrictions based on the problem.

For example, an astronomical catalog lists n properties of stars such as magnitudes at different wavelengths. Suppose we are given an n-tuple of magnitudes and want to find similar stars (i.e. those within some distance in our n-dimensional space). If points near each other are also nearby on the manifold then the search for nearby points is reduced from n dimensions to that of the lower dimensional manifold. Such a manifold is said to preserve locality. A related problem is to determine an underlying relation between the magnitudes if such a relation exists. In this case, the manifold should not be too specific to the supplied points because it will be applied to stars not in the catalog.

Formally, given a set of points $\{p_i\}_{i=1}^k$ in n-dimensional metric space N we want to identify an l-dimensional ($l < n$) metric space L and homeomorphism

$$h : L \mapsto N$$

which induces a manifold

$$M = h(L)$$

and a function to treat deviation from the manifold

$$e : N \mapsto M$$

minimizing error

$$\sum_{i=1}^{k} d_N^2\left(p_i, e(p_i)\right)$$

subject to some constraints on h and e discussed below. Note that since h is a continuous function, it obtains minimal distances from each p_i and e necessarily maps p_i to one of these minimally distant points of M. Finally, for convenience and noting that h is a bijection, define

$$t = h^{-1} \circ e$$

which translates into underlying coordinates of the manifold.

One constraint from our example is locality. This is not a constraint per se, but a property to optimize and the trade-off between locality and error is a common theme. Suppose we want to find all p_i near a given point p. Informally, we map p into the manifold and search for our mapped p_i in the manifold. Then we verify that they are indeed near p and return these points found. Formally, we are given a closed ball B in N centered at p with radius ρ and want to find $\{p_i\} \cap B$. To do this, we construct a closed ball C in L centered at $t(p)$ with minimum radius r so that

$$t(\{p_i\}) \cap C \supseteq t(\{p_i\} \cap B)$$

We therefore have a function

$$s : N \times \mathbb{R} \mapsto \mathbb{R} \text{ by } (p, \rho) \mapsto r$$

which gives the radius needed as a function of the query. In practice, it is costly to evaluate s directly so a guaranteed bound is used. Further, the $q_i = t(p_i)$ are precomputed and we have an algorithm for identifying points of $\{q_i\}$ within a ball (i.e. an R tree) which is more efficient in lower dimensions. Output is a set

$$I = \{i : q_i \in C\} \supseteq \{i : p_i \in B\}$$

which is filtered by checking that the points are in B:

$$\{p_i : i \in I\} \supseteq \{p_i\} \cap B \Rightarrow \{p_i : i \in I, p_i \in B\} = \{p_i\} \cap B$$

The form of locality presented is strict: only one ball in the manifold is searched. In some cases it makes sense to search multiple balls or regions that are not a finite union of balls. Suppose the manifold is a Euclidean sphere in $\mathbb{R}^3$ and h maps (θ, ϕ) to the sphere according to canonical spherical coordinates with constant radius. Further, we want to find points near $\theta = 0$. Then two regions of $\mathbb{R}^2$ are searched: $\theta < 2\pi$ and $\theta \geq 0$.

Two methods of fitting manifolds are Principal Component Analysis (PCA) and Locally Linear Embedding (LLE). Principal Component Analysis purely minimizes error (not locality) and is fundamentally linear (that is the homeomorphism h is a linear transformation). Variations on PCA such as the Gifi approach allow for non-linearity. However, they still perform poorly in locality [J]. LLE is a newer approach that fits linear transformations in small regions before combining them into a global manifold. LLE performs well at finding non-linear manifolds. It has problems with spheres (a neighborhood of one pole must be excluded) and locality can be improved [S].

We seek to create an algorithm to find manifolds with better locality. This algorithm can be evaluated in its guaranteed locality, test cases, and complexity. Ideally, new data can be added without rerunning the entire algorithm. Further, searches around points not in the data set should not be too complex. This means that the manifold found should not be too specific to the data.

Taylor series can be used to approximate the function h allowing it to have many functional forms. This of course introduces some restrictions on h (namely it has a convergent Taylor series), however many systems will have this property. Polynomials, fit locally or globally by the algorithm, can therefore be used to find h in many cases. One can also limit how specific the manifold is by limiting the polynomials. Consider limiting the number of non-zero coefficients to a fraction of the number of data points. Under this scheme, higher order terms vanish. However, the uncertainty in their coefficients is sufficiently large to justify this removal. Finally, complexity is reduced because the polynomial has less terms.

The largest part of the project is to identify an algorithm itself. Most likely, we will use the existing algorithms for ideas and possibly a Taylor series method. A formal description and implementation in software are expected to come out of the process.

Finally, the algorithm will be tested. This includes proving theorems about its locality, error, and complexity. Test cases will be prepared so that it is compared with other algorithms. Spurious fits (those specific to the data points) can by observing the effects of adding data points one by one. We hope to show the algorithm performs well in locality and that it can be applied to data similarity problems.

Of particular interest is to apply such an algorithm in computer networks. The points represent network traffic. With good locality, it becomes possible to guess that an activity is similar to a previous one. The protocol parameters can be adjusted accordingly. Further, strong deviation from the manifold or similarity to flagged points (i.e. worm traffic) indicates abnormal behavior. This is the group's larger project. I, possibly along with a few others, will be contributing the algorithm.

References

[J] I.T. Jolliffe, *Principal Component Analysis*, Springer, New York, 2002.

[S] L. Saul and T. Roweis, *Think Globally, Fit Locally: Unsupervised Learning of Low Dimensional Manifolds*, Machine Learning Research **4** (2003), 119-155.