Fast and Scalable Manifold Learning by Semi-Definite Programming

Babak Alipanahi* Nathan Krislock† Ali Ghodsi‡

The problem of nonlinear dimensionality reduction is most often formulated as an instance of semi-definite programming (SDP). The effectiveness of the SDP-based algorithms is limited by the computational complexity of SDP solvers. We propose a novel SDP formulation for dimensionality reduction based on Euclidean distance matrix completion (EDMC), which leads to a stable, fast, and scalable algorithm. The key observation is that in many cases, the appealing local structure lies beyond mere neighbor relationships. In fact, in most cases the structure of a reasonably large cluster of the data should be preserved as a whole, rather than being divided into very small neighborhoods. This observation leads to a new formulation that significantly reduces the size and the number of constraints of the SDP problem. Unlike existing large-scale variations of SDP-based methods, the algorithm is convex and is not prone to local minima or dependent on an initial value. We extend the framework of the proposed algorithm to suggest new formulations for some existing SDP-based algorithms in order to reduce their complexity. In addition, the formulation of the proposed algorithm provides a natural way for exploiting side-information (e.g. class labels) in order to improve the embedding.

our experimental results demonstrates the effectiveness of the proposed method on a variety of very large datasets.

*David R. Cheriton School of Computer Science,University of Waterloo, 200 University Ave. W., Waterloo, Ontario, Canada N2L 3G1, balipana@cs.uwaterloo.ca.

†Department of Combinatorics and Optimization ,University of Waterloo, 200 University Ave. W., Waterloo, Ontario, Canada N2L 3G1, ngbkrislock@math.uwaterloo.ca.

‡Department of Statistics & David R. Cheriton School of Computer Science, University of Waterloo, 200 University Ave. W., Waterloo, Ontario, Canada N2L 3G1, aghodsib@uwaterloo.ca.

Overview of the Proposed Method

Existing SDP-based methods are all instances of Euclidean distance matrix completion in which only the pairwise distances between the k-nearest neighbors are specified, and the goal is to find appropriate values for the unspecified distances.

Suppose a given data set with n data points can be divided into q clusters. We assume these clusters have good local properties and that their structure should be preserved. Let D_ℓ denote the Euclidean Distance Matrix of the points in cluster ℓ. Clearly the Euclidean Distance Matrix , of all data points can be represented as:

$$D = \begin{bmatrix} D_1 & . & \dots & . \\ . & D_2 & \dots & . \\ \vdots & \vdots & \ddots & \vdots \\ . & . & \dots & D_q \end{bmatrix},$$

where only block diagonal elements are known. The goal is to determine the unspecified elements of D, which can be cast as an instance of SDP. Crucially, we can formulate the optimization problem such that the size of the unknown positive semi-definite matrix is proportional to the number of clusters q. This is in contrast with existing SDP-based methods such as Maximum Variance Unfolding and Minimum Volume Embedding that need to compute a positive semi-definite matrix of size n.

Topic: Learning Algorithms
Preference: Oral