Large-scale approximate ISOMAP

Charles Y. Zheng and Jinshu Wang
Department of Statistics
Stanford University
Stanford, CA 94305
{snarles, jinshuw}@stanford.edu

Arzav Jain
Department of Computer Science
Stanford University
Stanford, CA 94305
arzavj@stanford.edu

Abstract

Isometric feature mapping (ISOMAP) is an unsupervised method for nonlinear dimensionality reduction (Tenenbaum et al. 2000), intended for exploratory data analysis or feature extraction of complex data with hidden low-dimensional structure, such as text, images, or neurological data. ISOMAP requires three computationally intensive subroutines: formation of a nearest-neighbors graph, computation of all-pairs shortest-paths for the nearest-neighbors graph and computation of the largest eigenvectors of the centered all-pairs distance matrix. Traditionally, computation of the all-pairs-distance matrix is $O(n^3)$, and computation of the top eigenvectors is $O(n^2 d)$, hence application of single-core ISOMAP scales poorly for large datasets. Here we propose two approaches to approximating ISOMAP within a distributed framework. The first is to compute a random subset of the columns of the all-pairs-distance matrix, thus approximating the eigenvectors of the full distance matrix by the left singular vectors of the submatrix. This results in substantial computational savings because single-source shortest paths algorithms can be used to compute the submatrix rather than all-pairs algorithms, and also because the SVD of the submatrix is much cheaper to compute than the eigenvectors of the full matrix. Meanwhile, the approximation error can be bounded using results on column-sampling SVD (Frieze 2004). Our second approach is to compute an approximation of the full all-pairs shortest-paths using a blocked version of Floyd-Warhsall algorithm for all-pairs shortest paths. Our blocked Floyd-Warshall sacrifices accuracy for improved latency; assuming that the data has been pre-shuffled, we show that its error is bounded given appropriate block sizes for small diameter graphs (such as the nearest-neighbor graphs required for ISOMAP). We implement these approaches in Apache Spark and demonstrate their application on synthetically generated data as well application in fMRI data.

1 Introduction

Isometric feature mapping (ISOMAP) is an unsupervised method for nonlinear dimensionality reduction [4], intended for exploratory data analysis or feature extraction of complex data with hidden low-dimensional structure, such as text, images, or neurological data. ISOMAP requires three computationally intensive subroutines: formation of a nearest-neighbors graph, computation of all-pairs shortest-paths for the nearest-neighbors graph and computation of the largest eigenvectors of the centered all-pairs distance matrix. Traditionally, computation of the all-pairs-distance matrix is $O(n^3)$, and computation of the top eigenvectors is $O(n^2 d)$, hence application of single-core ISOMAP scales poorly for large datasets.

While a large body of work exists for implementing all-pairs shortest-paths and eigendecomposition in supercomputers, our work is the first to leverage specific properties of the ISOMAP dimension

reduction problem to efficiently compute an approximation to the ISOMAP coordinates. A secondary difference between our work and the previous literatue is that we are motivated specifically by the Spark framework. While our results could be applicable to distributed systems in general and even to single-core systems, our algorithms should be especially well-suited for an advanced cluster computing framework like Apache Spark [5].

Here we propose two approaches to approximating ISOMAP within a distributed framework. The first is to compute a random subset of the columns of the all-pairs-distance matrix, thus approximating the eigenvectors of the full distance matrix by the left singular vectors of the submatrix. This results in substantial computational savings because single-source shortest paths algorithms can be used to compute the submatrix rather than all-pairs algorithms, and also because the SVD of the submatrix is much cheaper to compute than the eigenvectors of the full matrix. Meanwhile, the approximation error can be bounded using results on column-sampling SVD[1].

A second approach could be start by computing the full all-pairs shortest-paths matrix using a distributed algorithm. There exists a large body of work for all-pairs shortest-paths in the context of supercomputing, of these, the most suitable for implementation in the [3] Our second approach is to compute an approximation of the full all-pairs shortest-paths using a blocked version of Floyd-Warhsall algorithm for all-pairs shortest paths. Our blocked Floyd-Warshall sacrifices accuracy for improved latency; assuming that the data has been pre-shuffled, we show that its error is bounded given appropriate block sizes for small diameter graphs (such as the nearest-neighbor graphs required for ISOMAP). We rely on classical results on the generalized birthday problem [2] for these error bounds.

References

[1] A. Frieze, R. Kannan, and S. Vempala. Fast Monte-Carlo algorithms for finding low-rank approximations. *Journal of the ACM (JACM)*, pages 1–17, 2004.

[2] V. F. Kolchin, B. A. Sevastianov, and V. P. Chistiakov. *Random allocations*. Vh Winston New York, 1978.

[3] V. Kumar and V. Singh. Scalability of parallel algorithms for the all-pairs shortest-path problem. *Journal of Parallel and Distributed Computing*, 13(2):124–138, Oct. 1991.

[4] J. Tenenbaum, V. D. Silva, and J. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(December):2319–2323, 2000.

[5] M. Zaharia and M. Chowdhury. Spark: cluster computing with working sets. *Proceedings of the . . .* , 2010.