S592 Literature Report 1

Tzeng et. al (2008): MDS for large genomic datasets

Nick Sun

February 3, 2019

Main Contributions

Genomics researchers are often interested in using dimensionality reduction techniques to uncover hidden structure in data and make analysis more tractable. A popular technique for reducing high dimenional data while preserving information is multidimentional scaling (MDS). However, classical MDS is computational intensive and infeasible for datasets with even moderately large numbers of samples.

Tzeng et al. propose a new MDS algorithm referred to as "split-combine MDS" or SC-MDS. The principle benefits of this new algorithm are increased computational efficiency and usability for large datasets $O(N^3) vs. O(p^2N)$ for SC-MDS.

Algorithm overview

- The data is first split into K intersecting groups.
  - The authors illustrate some basic restrictions on how big the number of intersection N_i need to be. It must be at least as large as p, the number of significant features in the group.
- Apply classical MDS to each of these K subgroups
- For the overlapping points, apply a QR transformation to get the *affine transformation*
- Using the affine transformation, the K groups are re-combined to get the original dataset in reduced dimensions

Examples

The authors explored a few examples in their paper. In order to validate their new technique, when possible they compared the performance of SC-MDS to classic MDS using the Kruskal STRESS metric (smaller values of STRESS indicate that there are lower values of error between the distance matrices).

The first example involved simulated data in a spiral shape. I believe the intent was to demonstrate MDS' effectiveness in working in non-Euclidean space, but it did beg the question of how often biologists and other researchers deal with data in polar coordinates. The purpose of this example was to show that if SC-MDS is performed correctly, it performs very closely to classical MDS for small datasets. The calculated STRESS metrics were very low, indicating that there is little performance difference between the two algorithms.

The second example is where SC-MDS begins to shine. It focuses on a gene correlation map, represented as a 16502 x 16502 matrix which is impractical for classical MDS to handle. SC-MDS managed to complete in a reasonable time, and although the researchers were unable to compare the results to classical MDS, they were able to demonstrate with repetitions that the results had stabilized.

The final example presented was a demonstration of SC-MDS working to reduce the computation time of K means clustering. The main gist was that applying K means on SC-MDS reduced data performed about as well on the original dataset, with the benefit of increased computational efficiency since K means was being applied in lower dimensions.

Overall, the paper was interesting and made a strong case for why SC-MDS is a valid and valuable research tool.

Caveats and critical remarks

While the paper covers a lot of ground and the authors strive to validate their new algorithm, there are still lingering questions. The authors do not go over an example or describe a good process for selecting an appropriate K to subset the data into. One can surmise that a trying a lot of values of K and checking where the variance of SC-MDS stabilizes, but that's a detail that should have been left in the paper. The authors also barely touch on other weaknesses of classical MDS, such as the requirement of having no missing data. They do briefly mention using imputation to fill in missing values, but there is no further discussion beyond that. Some background in kNN imputation and how it's appropriate for these types of genomic datasets would have been useful for all readers.

The authors also provide relatively little detail on specifying a number N_i for the number of intersecting points between all the pairs of the K groups. In the paper, they discuss an example with a simulated dataset where the number of significant dimensions is known. However, in real applications this number of dimensions p will not be known *a priori*. This begs the question of how to properly decide an N_i, which they do not elaborate on further. More detail on this would have greatly clarified the actual implementation of SC-MDS.

Additionally, the authors introduce other established MDS techniques besides classical MDS, specifically Chalmer's Linear Iteration Method and the Anchor Point method. These methods are known to have greater computational efficiency than classical MDS, so naturally the question of why classical MDS was utilized on the K subgroups when the overarching goal of this new algorithm is computational tractability. It would have been interesting to see the performance of SC-MDS when the MDS algorithm of choice was switched up. Certainly it is an interesting area for future exploration and research.