Data Clustering with Spectral Algorithms

Elliot Bobrow - elliotgreybobrow@gmail.com
Windward School, Class of 2023
USC Viterbi Department of Computer Science, SHINE 2022

USC Viterbi School of Engineering

SHINE Summer High School Intensive in Next-Generation Engineering

Introduction

There is much that isn't known about the function of different proteins in biological organisms. Scientists are working on gaining a better understanding of protein functions by collecting single-cell RNA sequencing data (scRNA-seq), datasets that contain information about many individual cells and the RNA sequences that each one expresses. These datasets can offer key insights into protein function but are far too big to interpret by hand. This is where Professor Zhang's research comes in.

Objective & Impact of Professor's Research

Professor Zhang's research focuses on clustering algorithms to understand scRNA-seq data. For example, a dataset containing cancer cells at different stages of development can be clustered by their development stage, providing insight into the growth of cancer cells. This can help doctors to determine the progression of cancer in patients, or potentially to diagnose it earlier.

```
Finished in 23 iterations
15  144 175 34
21  95  59  92
36  54  47  165
228 7   19  9
```

***Figure 1**: Example of clustering SBM data with K-means. PC: Elliot Bobrow*

```
Finished in 8 iterations
7   2   12  287
5   3   276 8
6   291 6   5
282 4   6   0
```

***Figure 2**: Example of clustering SBM data with PCA and K-means. PC: Elliot Bobrow*

In a recent publication, Professor Zhang and our mentor Chandra Sekhar Mukherjee proved that applying the algorithm PCA to scRNA-seq data makes clustering more accurate by bringing cells from the same cluster closer together at a higher rate than cells from different clusters.

Results

After seeing that applying PCA before K-means on SBM data improved clustering (**Figures 1** and **2**), we tested on an experimental dataset containing mouse embryo cells at 4 different stages of development. To show that PCA helped, we observed the average intra- and inter-cluster distances before and after PCA. The power of PCA is that it compresses intra-cluster distances much more than inter-cluster distances, not that intra-cluster distances will all be equal.

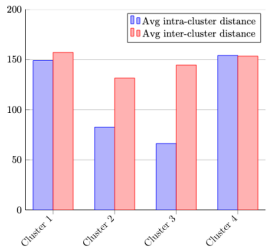

***Figure 3**: Cluster distances before PCA. Especially in clusters 1 and 4, there is no significant difference between intra- and inter-cluster distances. PC: Elliot Bobrow via Overleaf*

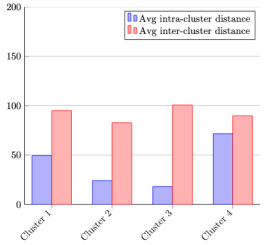

***Figure 4**: Cluster distances after PCA are much better, except for in cluster 4. PC: Elliot Bobrow via Overleaf*

Although the dataset had 4 clusters, we got better results by clustering into 10 clusters (**Figure 6**) than clustering into 4 (**Figure 5**). Using a larger number of clusters allowed cluster 4 to be treated as many smaller clusters.

```
Finished in 24 iterations
[[ 10 302 683 146]
 [458   0   0 219]
 [298   1   0 222]
 [167   0   0 211]]

Finished in 42 iterations
[[195   0   0 230]
 [ 10 300 683 142]
 [263   3   0 205]
 [465   0   0 221]]
```

***Figure 5**: Clustering results with k=4 pre-PCA (top) and post-PCA (bottom). PC: Elliot Bobrow*

```
Finished in 20 iterations
[[ 12   0   0  27]
 [  1   0   0 107]
 [  0   0   0  10]
 [  0   0   0 196]
 [  1   0   0 182]
 [193   0   0   0]
 [  0 300 683  84]
 [307   0   0   5]
 [ 97   3   0 185]
 [322   0   0   2]]

Finished in 10 iterations
[[231   0   0   0]
 [  1   0   0 173]
 [  0  74 647  52]
 [  0 229  36   0]
 [348   0   0   0]
 [  1   0   0 187]
 [ 17   0   0  54]
 [ 53   0   0 179]
 [282   0   0   3]
 [  0   0   0 150]]
```

***Figure 6**: Clustering results with k=10 pre-PCA (top) and post-PCA (bottom). PC: Elliot Bobrow*

Tools Used

SBM (Stochastic Block Model) - An algorithm to generate datasets. It creates *n* nodes and groups them into *k* random clusters. If two nodes are in the same cluster, they are more likely to share an edge (represented by a 1 in the data).

K-means - A clustering algorithm. K-means is greedy, meaning it sacrifices accuracy for speed and will usually not find the best possible clusters. It can be improved by using K-means++, which is what we used on the experimental data.

PCA (Principal Component Analysis) - A tool for re-representing data in a smaller matrix. Instead of each cell having one value for each RNA sequence, it has fewer values while still retaining most of the original information.

Acknowledgements

I would like to thank Professor Jiapeng Zhang for accepting me into the program, my mentor Chandra Sekhar Mukherjee for all his help, my lab partner Eden Cahill for being an amazing partner, and Dr. Katie Mills and Michelle Emelle for making SHINE what it is.

References

[1] Chandra Sekhar Mukherjee and Jiapeng Zhang. *Compressibility: Power of PCA in Clustering Problems Beyond Dimensionality Reduction*. 2022. arXiv: 2204.10888 [cs.LG].

[2] Vladimir Kiselev, Tallulah Andrews, and Martin Hemberg. "Challenges in unsupervised clustering of single-cell RNA-seq data". In: *Nature Reviews Genetics* 20 (Jan. 2019). DOI: 10.1038/s41576-018-0088-9.