Iterative Consensus Clustering: An Algorithm We Can All Agree On
Mindy Hong\textsuperscript{1}, Robert Pearce\textsuperscript{2}, Kevin Valakuzhy\textsuperscript{3}, Carl Meyer\textsuperscript{2}, Shaina Race\textsuperscript{2}

Methods
Iterative Consensus Clustering: An Algorithm We Can All Agree On
Mindy Hong\textsuperscript{1}, Robert Pearce\textsuperscript{2}, Kevin Valakuzhy\textsuperscript{3}, Carl Meyer\textsuperscript{2}, Shaina Race\textsuperscript{2}

Background Information

Clustering: Grouping data based on a predefined metric of similarity.

Why Cluster?
First step in interpreting large amounts of data
- Physical Observations
- Gene Expression
- Term Frequencies

Example Algorithm: K-Means
1. Randomize centroids for each cluster
2. Cluster each point with its nearest centroid
3. Move centroid to mean of its cluster
4. Repeat steps 2 and 3 until equilibrium

Problems
- Fundamental Problem of Clustering “There does not exist a best method, that is, one which is superior to all other methods” (Kogan).
- Determining the number of clusters, also known as $k$

Objectives

- Determining an accurate value for $k$
- Develop a technique that uses multiple algorithms to reach a consensus on a final clustering

Consensus Clustering

1. Each row and column represent a point
2. Each matrix entry is the number of times its corresponding row and column are clustered together

Eigengap Method
- Eigengap: the largest difference between consecutive eigenvalues
- Create a special “P Matrix” using the consensus matrix
- Sort the P Matrix eigenvalues
- The index of the eigenvalue before the eigengap is an approximation for $k$

Methods

Consensus Clustering

1. Determine an accurate value for $k$
2. Develop a technique that uses multiple algorithms to reach a consensus on a final clustering

Results / Conclusion

Number of Clusters

- Traditional Consensus Clustering
- Iterated Consensus Clustering

- Errors are weeded out through iteration
- Most algorithms come to a consensus on the final clustering
- Final clustering improves upon many individual algorithms

Conclusion

- Iterated Consensus Clustering offers better results than traditional consensus clustering in:
  - Finding the number of clusters
  - Returning an appropriate clustering
  - Calculations were done using the following techniques: Expectation Maximization, PDDP, k-Means, NMF, PCA, SVD

Acknowledgements

We would like to thank Dr. Carl Meyer and Shaina Race for their guidance through this project, as well as NSF and NSA for making this research possible.