Super Resolution Cluster Analysis

NSERC USRA Report

Da Wei (David) Zheng
Summer 2016

INTRODUCTION

Stochastic Optical Reconstrction Microscopy (STORM) is a relatively new biological tool for producing high resolution fluorescence images. The clustering behaviour of these receptors are important in B-cell activation and thus important to understand the immune system. As part of the project, we created a graph based clustering algorithm that we compared with other algorithms like ClusterViSu, and DBSCAN. A benchmark set of clustered data was created with two different methods and used to compare the algorithms.

CLUSTERING ALGORITHMS

STORMGRAPH

StormGraph uses a graph based agglomerative clustering algorithm that minimizes the map equation [2] created by me and Joshua Scurll. The algorithm runs 5 Monte Carlo simulations of randomly spatially uniform data through our algorithm to determine the minimum number of points for a meaningful cluster and r_0, the maximum search distance for creating edges. Then the algorithm run 50 iterations of the algorithm to find the minimum quantity of the map equation.

CLUSTERVISU

ClusterViSu is a recently published clustering method for Super Resolution data that uses Voronoi tesselations [1]. The method uses Monte Carlo simulations to find a area threshhold and then groups the Voronoi cells less then this threshhold to find clusters, corresponding to high density regions.

DBSCAN

DBSCAN is a density based clustering algorithm that finds regions of locally higher density [3]. The algorithm considers all points that have k neighbors within some radius ϵ, and joins all such points together. The algorithm is not sensitive to a choice of k, the authors of the orignal paper recommended $k = 4$ to be a good value. ϵ is recommended to be at the knee of the sorted k-dist graph, and we chose such a value for testing.

Simulated Data

Method 1

One set of benchmark data was generated by generating 30 clusters that were gaussians with fixed average density radius and eccentricity. Overlaying these clusters was a uniform random background of a known density proportion to the clustered. Background that overlapped the clusters were assigned to be part of the clusters. From a base set of parameters that were reasonably challenging for all clustering programs we adjusted the following parameters: the number of clusters, the radius of clusters, the variance in the radius of the clusters, the eccentricity of the clusters, the density of the clusters, and the relative background density.

Method 2

A Dirichlet process was also used to simulate clusters, as we believe this serves as a good model for BCR clustering. Three parameters were used as input: the number of points, N, the concentration parameter α, and the radius of the underlying Guassian distribution r_0. For the ith point, it is placed uniformly at random on the region of interest with probability $\frac{alpha}{alpha+i-1}$. If $i > 1$ it is added a Gaussian away from an existing point $j \in \{1 \cdots i-1\}$ with probability $\frac{1}{alpha+i-1}$. Ten samples of varying α, r_0 and N were chosen.

Results

To evaluate the results of the clustering algorithm, a metric called the mean F measure was employed. The mean F measure is the harmonic mean between precision, the percentage of correct predictions, and recall, the fraction of relavent points clustered. This measure ranges from a perfect score of 1, down to 0. As can be seen from figure 2.1, our graph based algorithms far outpreforms DBSCAN and preforms better then ClusterViSu does. While this may be in part due to the way the type of data sets generated, we believe the data sets are representative of the biological data this algorithm can be effectively used on.

Acknowledgements

I'd like to thank Josh Scurll for his supervision and guidance, Daniel Coombs for his supervision, Libin Abraham and Henry Lu for help with the biology, and the rest of Daniel Coombs lab and the other UBC USRA students for their support.

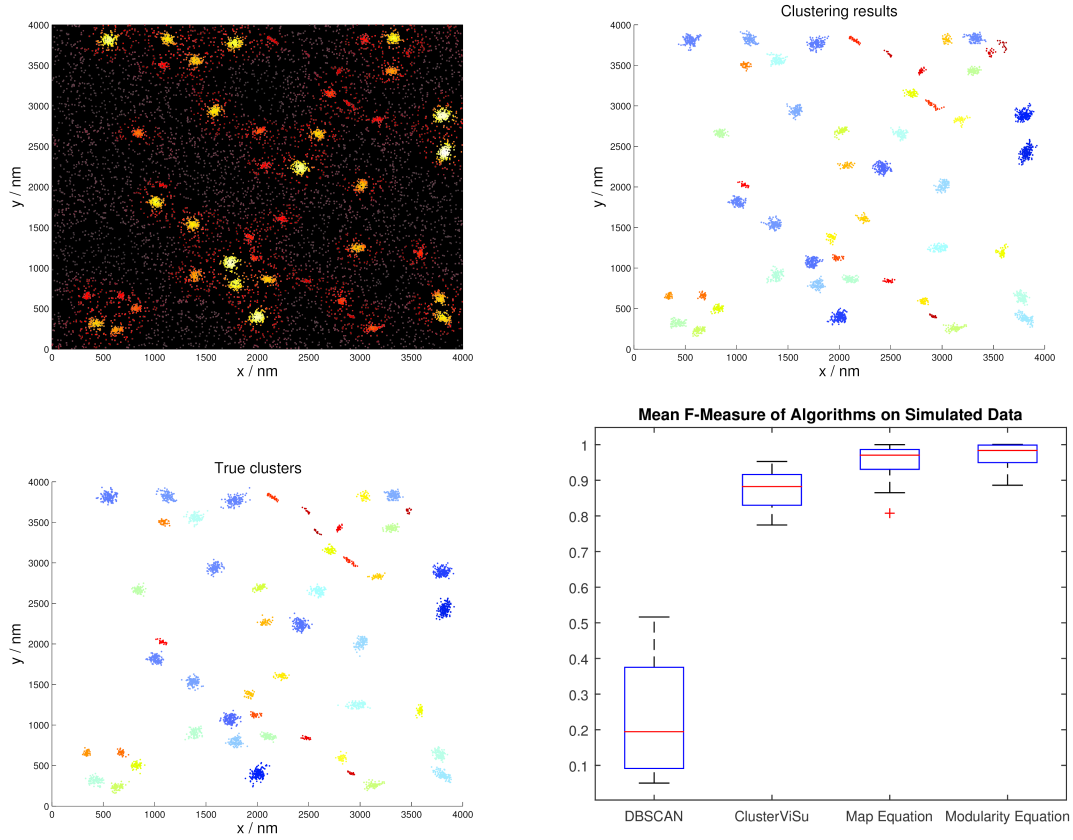

Figure 0.1: Raw data, data clustered with modularity equation, actual labels on data, and mean F-measure across 36 simulated data sets.

REFERENCES

[1] L. Andronov et al. “ClusterViSu, a method for clustering of protein complexes by Voronoi tessellation in super-resolution microscopy”. In: *Sci Rep* 6 (2016), p. 24084.

[2] Ludvig Bohlin et al. “Community Detection and Visualization of Networks with the Map Equation Framework”. In: *Measuring Scholarly Impact: Methods and Practice*. Ed. by Ying Ding, Ronald Rousseau, and Dietmar Wolfram. Cham: Springer International Publishing, 2014, pp. 3–34. ISBN: 978-3-319-10377-8. DOI: 10.1007/978-3-319-10377-8_1. URL: http://dx.doi.org/10.1007/978-3-319-10377-8_1.

[3] Martin Ester et al. “A density-based algorithm for discovering clusters in large spatial databases with noise”. In: AAAI Press, 1996, pp. 226–231.