Cell Cluster Analysis and Neighbour Detection

NSERC USRA Report - Summer 2017

Cindy Tan

August 31, 2017

1 Introduction and Motivation

This summer, I worked under the supervision of Dr. Leah Keshet and Dhananjay Bhaskar on two separate projects. The first one was a large project aiming to automatically identify all cells in a microscopy image and classify them by morphology [2]. The methodology was tested on images of pancreatic cancer cells provided by Dr. Calvin Roskelley. My contribution to the project was to find groups of cells with distinctive morphology using cluster analysis. This involved generating synthetic data on which to test algorithms, implementing clustering algorithms and assistive heuristics, and documenting the results.

In the second project, I analyzed the movement of cells in some of Dr. Roskelley's experiments with cancer cells (Figure 1). These experiments involve cells with two types of surface adhesion molecules, coloured red and green by fluorescent protein transfection. Green cells tend to be more mobile than red cells, and one of the hypotheses of the experiments is that certain cells tend to emerge as "leaders" that pull other cells along. I worked with the data extracted from the experiment videos to analyze and visualize cell neighbour relations over time (types of relations, how many neighbours on average, and so on).

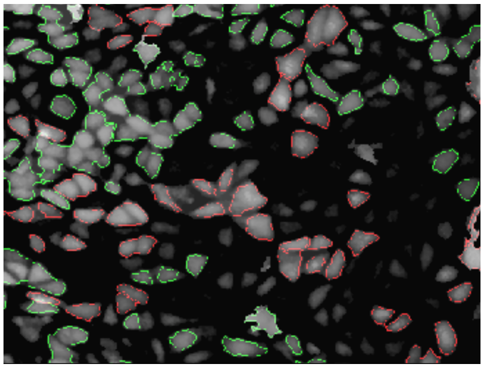

Figure 1. A snapshot of one of the original microscopy images.

2 Clustering

There are many existing algorithms to find clusters in a data set. In order to compare the performance of different methods, I generated many different sets of synthetic data containing clusters. We started with two simple clustering algorithms: k-means [4] and DBSCAN [3]. Since both algorithms require user input, we also explored the usage of silhouette scores [5] and the gap statistic [6] to find the algorithm inputs that produced the best clustering result.

For a more advanced method, we also implemented the OPTICS clustering algorithm. OPTICS is built on the ideas behind DBSCAN, but is an improvement on DBSCAN because it is much less sensitive to input (results will not vary drastically if given input parameters are modified) and can find clusters of varying densities. The OPTICS algorithm produces a hierarchical clustering from which clusters can be extracted (Figure 2) [1]. Our testing showed that OPTICS performed better than both k-means and DBSCAN with a variety of synthetic data sets.

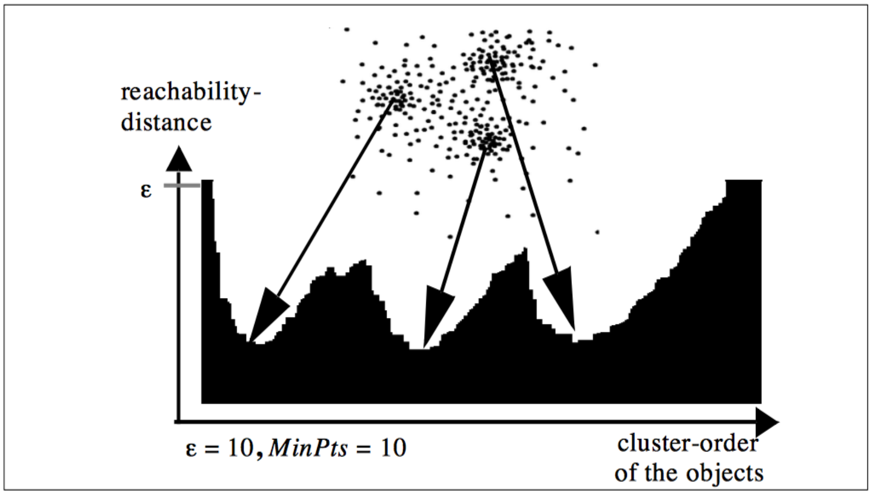

Figure 2. The output of OPTICS is a reachability distance plot. The valleys of the plot correspond to clusters. Image from [1].

3 Cell Neighbour Analysis

I worked in collaboration with Ceylin Özdemir and Alannah Wilson, undergraduate students in Dr. Roskelley's research group, to identify and investigate cell neighbour relations. Once we determined which cells were neighbours, we generated graphs (Figure 3) in hopes of finding underlying structure.

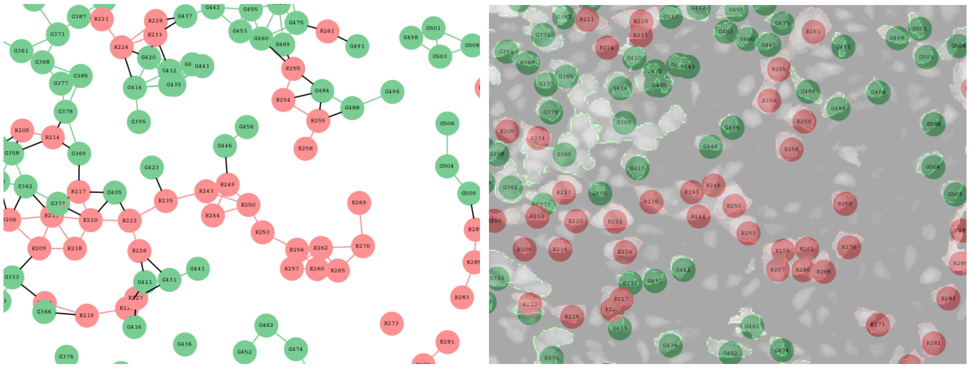

Figure 3. (Left) Graph where nodes represent cells and edges represent a neighbour relation. (Right) Same graph overlaid on the microscopy image it corresponds to (Figure 1).

We looked at patterns in homotypic (same type: red-red or green-green) vs. heterotypic (different type: red-green) neighbours (Figure 4, left) as well as cell clustering within red or green cells to see if a group of red cells were being pulled by green cells (Figure 4, right). Cluster aspect ratio and orientation were also of interest to identify a “leading edge” of green cells.

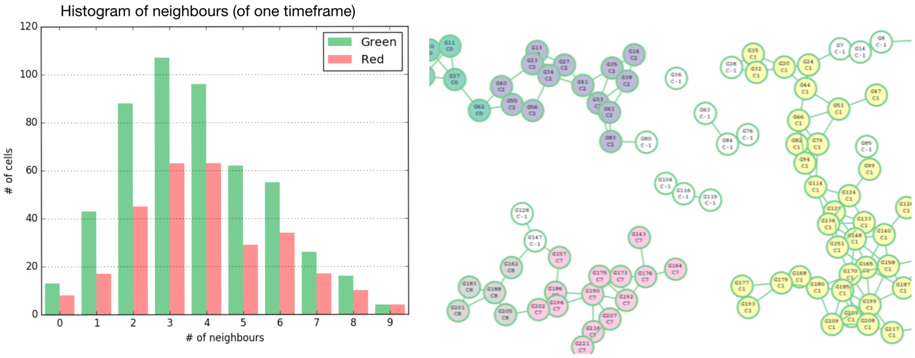

Figure 4. (Left) A histogram of the number of cells with a specific number of neighbours shows that red cells tend to have more neighbours on average. (Right) A snapshot of the clustering run on green cells only. The large yellow cluster is an example of a leading edge cluster.

4 References

[1] Ankerst, M., Breunig, M. M., Kriegel, H., & Sander, J. (1999). OPTICS: Ordering Points To Identify the Clustering Structure.

[2] Bhaskar, D. (2017). Morphology Based Cell Classification: Unsupervised Machine Learning Approach. Master?s Thesis, University of British Columbia.

[3] Ester, M., Kriegel, H., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise.

[4] Lloyd, S. P. (1982). Least squares quantization in PCM. IEEE Transactions on Information Theory, 28.2, 129-137.

[5] Rousseeuw, P. J. (1987). Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. Computational and Applied Mathematics, 20, 53-65.

[6] Tibshirani, R., Walther, G. & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B, 63, Part 2, 411-423.