Exploring KNN Clustering

Chris Bell • Catherine Peña (Mentor: Dr. Byron Gao)

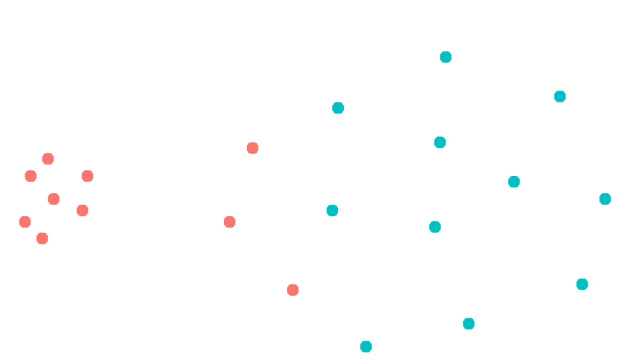
Objective

- Existing clustering objectives such as compactness in the *k*-means problem are capable of finding well-separated, spherical-shaped clusters.
- For this project we set out to provide novel clustering formulations that can overcome this problem and identify meaningful clusters of arbitrary shape.

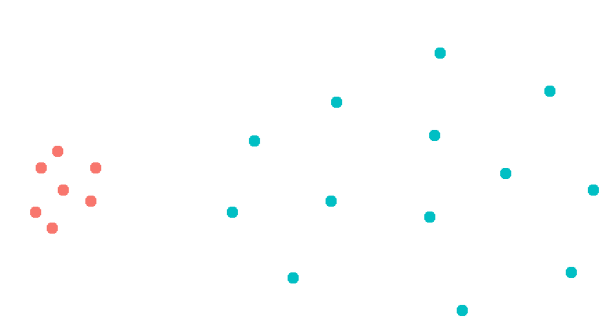
Contributions

- Identify novel KNN-based clustering formulations
- Design heuristic algorithms that solve the formulated clustering problems

Illustration



K-means result



This is an optimal configuration with survivability = 10

Clustering Formulation

Clustering – given a dataset *D* and number of clusters *c*, a clustering of *D* partitions *D* into *c* number of groups, each being a cluster, that exhibit internal cohesion and external isolation.

Neighborhood – given a neighborhood size *k*, the neighborhood of a data point *p* is the set of *k* points (excluding *p*) that are nearest to *p*.

Violation – A data point *p* incurs a violation under *k* if its cluster membership differs from the most-dominating cluster within *p*'s neighborhood.

Violation number – Given clustering *P* and neighborhood size *k*, the violation number of *P* under *k* is the total number of violations.

Clustering configuration – is a pair (*P*, *v*), where *P* is a clustering, and *v* is a violation number

Promising configuration – is a configuration for which *v* is the least.

- There may be more than one **promising configuration**. A meaningful clustering is always contained in a promising configuration but not vice versa.

Survivability – A configuration (*P*, *v*) survives a neighborhood size *k* if, under *k*, *P* has a violation number no greater than *v*. The **survivability** of (*P*, *v*) is the number of *k*'s under which (*P*, *v*) survives.

- We are most interested in counting survivability for promising configurations as a way to select optimal configurations.

Optimal configuration – is a promising configuration with the most survivability

Optimal clustering – if (*P*, *v*) is an optimal configuration, the *P* is an optimal clustering

- Each optimal clustering represents a meaningful clustering.
- There may be more than one optimal configuration and thus more than one optimal clustering.

Alternative Formulation

Moving forward, we will explore and refine an alternative clustering formulation:

Survivability of a point – given a clustering, given a set K of *k* values, the survivability of point *p* is the number of *k* values in K where *p* does not incur a violation.

Volatility of a point – volatility of *p* = |K| - (survivability of *p*); this is the opposite of survivability.

Survivability of a clustering – sum of the survivability of each point.

Volatility of a clustering – sum of the volatility of each point.

Clustering problem formulation – given a set of points, given a set K of *k* values, find a clustering with its survivability w.r.t K maximized; or, equivalently, find a clustering with its volatility w.r.t K minimized.

Algorithms

Exhaustive search: We used exhaustive search to perform experiments on toy datasets to help verify and shape our clustering formulations.

Efficient heuristic algorithm: We propose an efficient O(n2) heuristic algorithm that works in two phases:

- Phase I (initial clustering): Form an initial clustering using an existing algorithm such as *k*-means
- Phase II (iterative correction): At each step, find a point *p* whose cluster membership switch will lead to the largest decrease of volatility of the clustering. Then perform this switch on *p*. The iteration terminates when there's no such point whose membership switch would lead to decrease of volatility of the clustering.

Future Work

- Refine the clustering formulations
- Refine and Implement the heuristic algorithms
- Perform empirical evaluation by comparing our heuristics with existing methods such as *k*-means and spectral clustering
- Write up a technical report and make conference submissions