Representative Traffic Management Initiatives

Alex Estes, David Lovell, Michael Ball

University of Maryland-College Park

June 22nd, 2016

Finding Representative Data Points

- We take a large set of data
- Produce a small set of representative data points
- We may attach some information to each representative
- Representatives should be a good description of original data points
  - Data exploration
  - Allow users to better understand data
- Our application requires that each representative is a member of the original dataset

Application: Traffic Management Initiatives

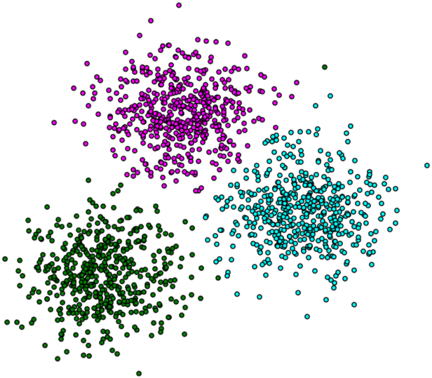
- Traffic Management Initiatives (TMIs) are used by the Federal Aviation Administration to balance the availability of resources with their demand
- TMIs often have parameters such as:
  - Number of flights allowed to access resource during restrictions
  - Time interval that access to a resource is restricted
  - Geographic region of flights included in TMI
- We wish to produce representative TMIs so that TMI decision-makers can more easily review historical TMIs.

Literature Review

- In air traffic management literature, k-means cluster centroids have been used for a similar purpose:
  - To create capacity scenarios (Liu, Hansen, and Mukherjee 2008)
  - To find types of days based on Ground Delay Program occurrence (Grabbe, Sridhar, and Mukherjee 2013)
  - To generate TMI scenarios for use in simulations (Delgado, Prats, and Sridhar 2013)
- For general problem: clustering is the nearest methodological analogue.
  - General resources: Jain, Murty, and Flynn 1999; Tan, Steinbach, Kumar, et al. 2006, Chapter 8

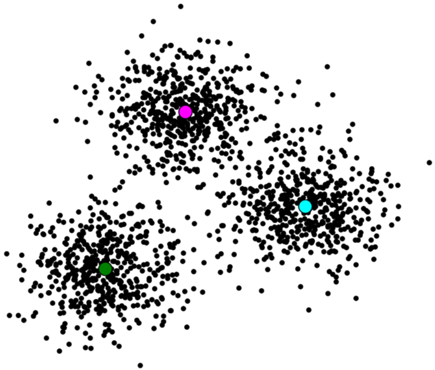
Clustering vs. Finding Representatives

- Goal of clustering: partition data into sets
- Cluster representatives may be produced but this is an intermediate step
- Clustering algorithms that produce representatives can also be applied to find representatives



Clustering vs. Finding Representatives

- Goal of representatives: find representatives
- A clustering may be generated from the representatives, but that is not the intent.
- If you removed all data except representatives, should still have reasonable idea of how data falls



Clustering vs. Finding Representatives

One example: data falls in single cluster of irregular shape.

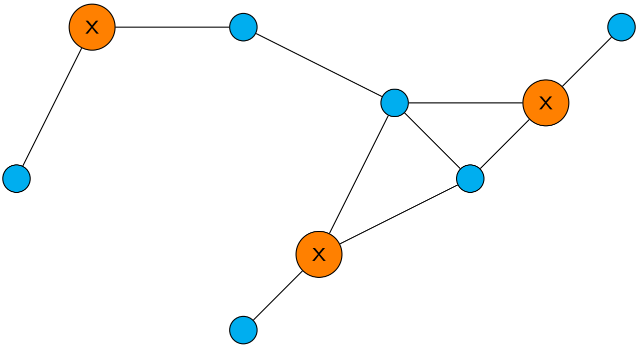
- Can accurately describe this data as a single-cluster, but this does not provide much information about the distribution of the data
- Representatives provide a way to describe this dataset that does not rely on cluster structure.

Finding Representatives with Similarity Information

- Assume we have something which tells us whether or not two observations are similar.
- We will choose a set R of representative observations to satisfy the following:
  1. If an observation is not a representative then it is similar to at least one representative in R
  2. R is of minimum size

Graph theoretic formulation

- Similarity graph G: node for each observation, edges between similar observation.
- We are looking for a set of nodes $R \subseteq V(G)$ such that
  - Every node of $V(G) \setminus R$ is adjacent to a node of R.
  - R is of minimum size.
- Note: not necessarily unique
- A solution to this problem is a *minimum dominating set* (MDS)



Minimum Dominating Set Solution

- Minimum dominating set is a well-known NP-hard problem
- IP formulation for exact solution:

$$\min \sum_{v \in V(G)} x_v$$

s.t.

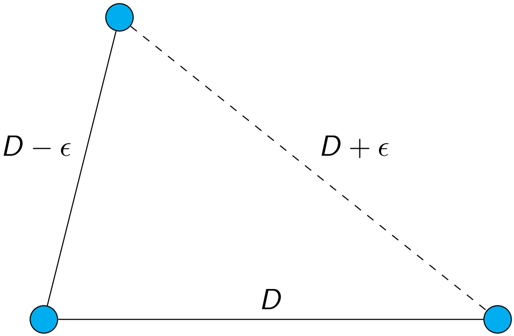
$$x_v + \sum_{u \in N(v)} x_u \geq 1$$

$$x_v \in \{0, 1\}$$

- Approximate methods that produce small dominating sets have also been proposed (Parekh 1991, Sanchis 2002, Ho, Singh, and Ewe 2006)

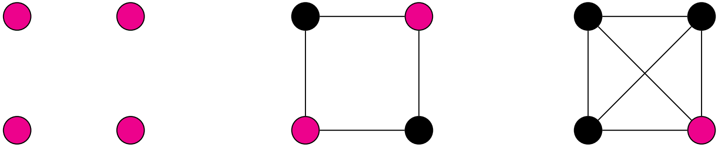
Representatives from Distance Information

- Consider if we have distance $d(x, y)$ between points x and y instead of similarity information.
- We can choose some distance threshold D and say that two points are similar if their distance is at most D.
- The we could apply our MDS method.



Representatives from Distance Information

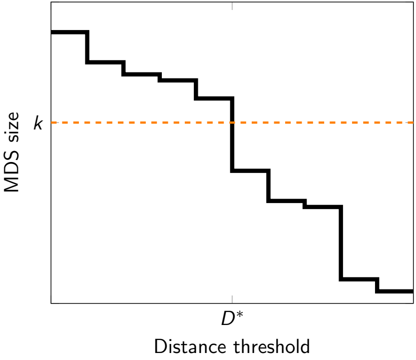
- If D is small enough, then no pair would be similar, and every point is a representative.
- As D increases, more pairs become similar, and the number of representatives in the MDS method decreases.
- If D is large enough, any pair will be similar, and the MDS method will produce a single representative.



(a) D small (b) D intermediate (c) D large

Representatives from Distance Information

- Given a number k, we can find the smallest distance threshold D^* such that there are at most k representatives.
- This is equivalent to the k-center problem, where we place k facilities to minimize the maximum distance that any customer must travel.
- A clustering method has been proposed where this problem is solved approximately to find cluster centers (Gonzalez 1985)

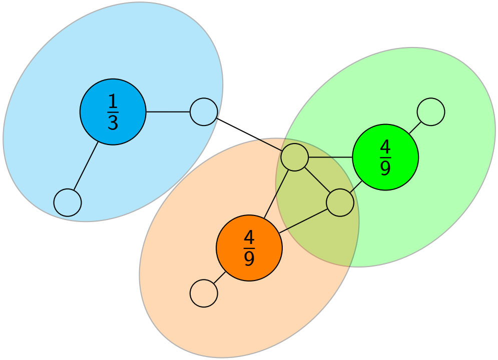


Solving the k-center problem

- The k-center facility location problem is also NP-Hard.
- There are existing exact methods (Chen and Chen 2009; Elloumi, Labbé, and Pochet 2004; Caruso, Colorni, and Aloi 2003; Ilhan, Özsoy, and Pinar 2002; Minieka 1970).
- There are existing approximate methods and heuristics (Davidović et al. 2011; Robič and Mihelič 2005; "Lexicographic local search and the p-center problem"; Caruso, Colorni, and Aloi 2003; Mladenović, Labbé, and Hansen 2003; Mihelič and Robič 2003; Hochbaum and Shmoys 1985; Gonzalez 1985).

Prevalence of Representatives

- Some representative will represent common data points, while others will represent outliers or unusual points.
- We define a measure called *prevalence* to reflect this.
- Prevalence is the proportion of observations that are similar to the representative.
- If using k-center method, use similarity from final distance threshold



Prevalence of Representatives

Notes:

- Since every observation is similar to a representative, the prevalences sum to at least one.
- An observation can be similar to multiple representatives, so the prevalences can sum to a number greater than one.
- The prevalence of a representative is an estimator of the probability that an observation will be similar to that representative
- Under some regularity conditions, this can be shown to be a consistent estimator.

k-center vs. k-means?

Disadvantages of k-means compared to proposed method:

- Does not produce centroids that are members of original dataset
- Can only be applied with Euclidean distance

Advantages of k-means compared to proposed method:

- Fast
- Has precedence in the literature

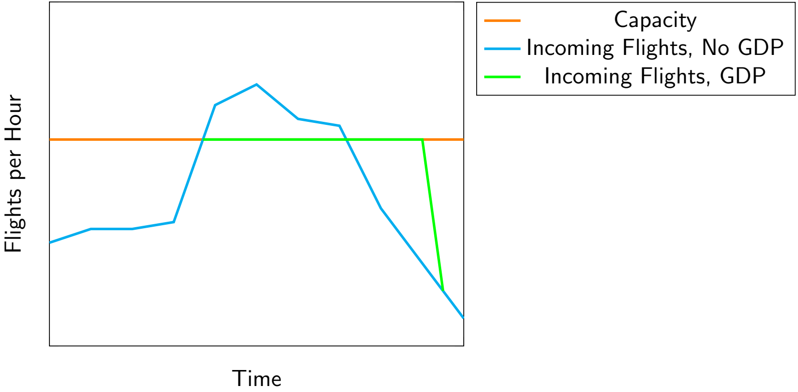
k-center vs. k-means

Other differences:

- k-center (our method): minimize maximum distance from any point to its nearest representative
- k-means: approximately minimize sum of squared distance from any point to its nearest representative
- Expect k-means to place some priority of high-density regions over low density regions
- Expect k-center to give more even coverage of data, and less affected by local variations in density.
- Similarly, expect k-center to be better for outlier detection.
- Could consider variations on our method with other objectives: e.g. minimize sum of absolute distance for even more priority on high-density regions.

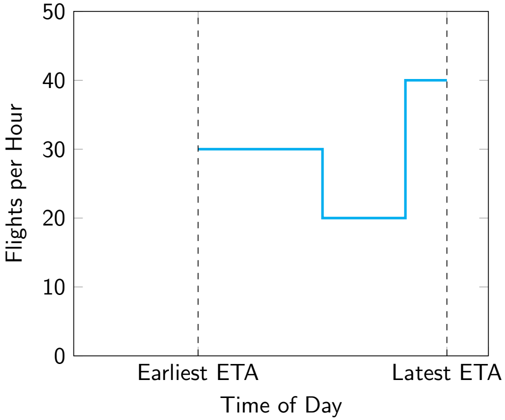
Example: TMI planning

- Consider one specific type of TMIs: Ground Delay Programs.
- Scheduled traffic can exceed capacity of airport to handle arrivals
- To prevent unsafe situations, the FAA will delay flights on the ground.
- This is a ground delay program (GDP)



GDP Features

- File Time
- Earliest/Latest Affected ETA
- Rates:



GDP Features:

Scope:

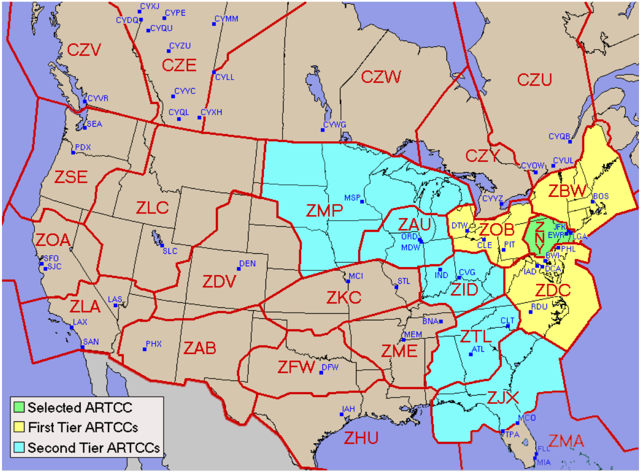


Image from FAA: http://www.fly.faa.gov/ois/tier/themap.htm

- Sometimes specified with numeric radius instead

Defining Distance

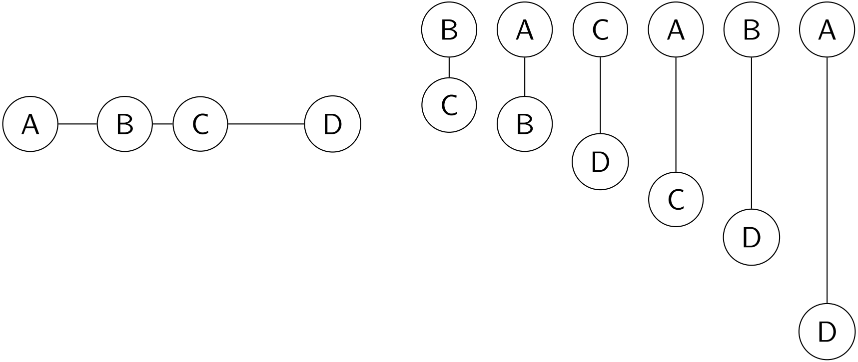
Difficulty: features have vastly different units.

- We want a measure of distance that remains interpretable, but includes all features

Defining Distance

Idea:

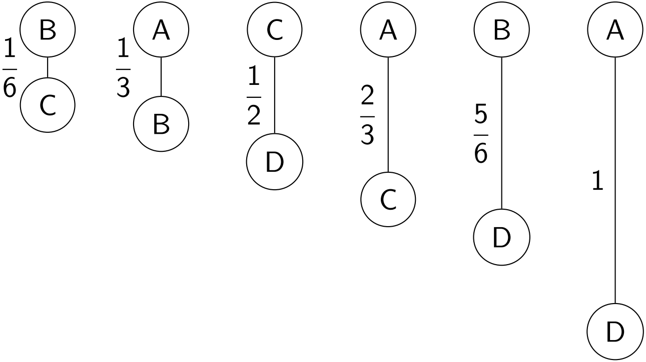
- Assume that for an individual feature f, we can find the distance between any pair of points in that feature.
- The distance between two observations x and y in feature f is relatively large if most pairs of observations have a smaller distance in that feature.



Defining Distance

Feature-wise quantile distance:

- For each feature f, we can use the proportion of pairs of whose distance is at most that of the pair (x, y) as a normalized measure of distance for the pair (x, y) in the feature f.
- We can then define the overall distance between (x, y) as the maximum of this normalized distance in each feature.



Defining Distance

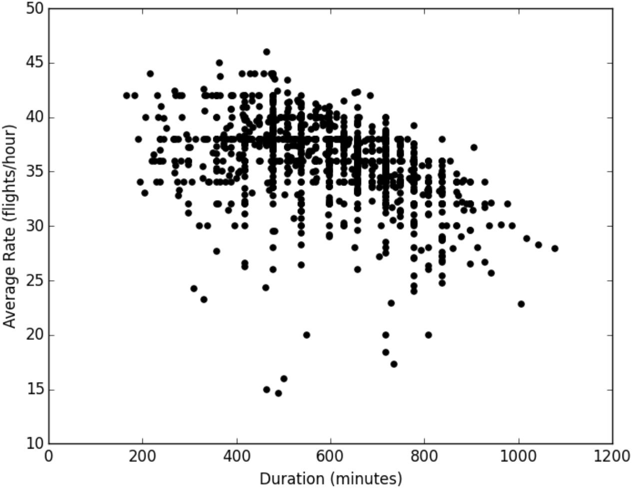
- Can express distance thresholds in terms of feature thresholds.
- Distance between a pair is less than D if and only if normalized distance of each feature is less than D.
- Can convert normalized distances back to distances in original features.
- This allows us to present our results in a manner that is easily interpreted.

Ground Delay Program Example:

- Dataset: every GDP at Newark Liberty International Airport from January 1st, 2007 to December 31st, 2014
- Data taken from FAA Advisory Database (www.fly.faa.gov)
- In total: 1302 GDPs
- In order to examine the data visually, we focus on two features: average rate, and duration.
- We compare our method with three other clustering methods that produce representatives.
- k-center problem solved exactly using method from Elloumi, Labbé, and Pochet 2004.

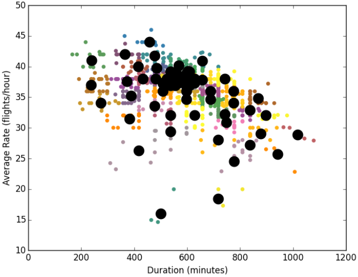
GDP Example:

Scatter plot of observations:

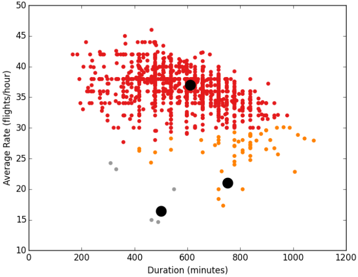


GDP Example:

Results from affinity propagation and mean shift:



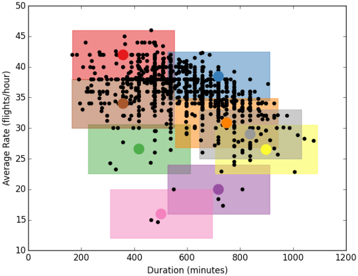
(a) Affinity Propagation



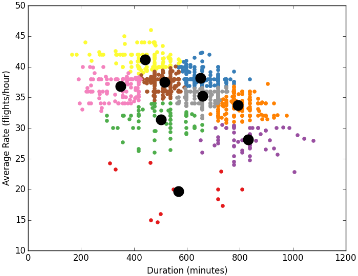
(b) Mean Shift

GDP Example:

Results from k-center method and k-means clustering:



(a) k-center (proposed method)



(b) k-mean

Further work

- Incorporate this into decision-support tool
- Visualize results from higher-dimensions
- Use representatives in further analysis, e.g. estimating the effectiveness of TMIs that are similar to a representative
- Heuristics for k-center in very large datasets

End of Presentation

Thank you for your attention. Questions?

Caruso, C, A Colorni, and L Aloi. 2003. Dominant, an algorithm for the p-center problem. *European Journal of Operational Research* **149**(1) 53–64.

Chen, Doron and Reuven Chen. 2009. New relaxation-based algorithms for the optimal solution of the continuous and discrete p-center problems. *Computers & Operations Research* **36**(5) 1646–1655.

Davidović, Tatjana et al. 2011. Bee colony optimization for the p-center problem. *Computers & Operations Research* **38**(10) 1367–1376.

Delgado, Luis, Xavier Prats, and Banavar Sridhar. 2013. Cruise speed reduction for ground delay programs: A case study for San Francisco International Airport arrivals. *Transportation Research Part C: Emerging Technologies* **36** 83–96.

Elloumi, Sourour, Martine Labbé, and Yves Pochet. 2004. A new formulation and resolution method for the p-center problem. *INFORMS Journal on Computing* **16**(1) 84–94.

Bibliography II

Gonzalez, Teofilo F. 1985. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science* **38** 293–306.

Grabbe, Shon, Banavar Sridhar, and Avijit Mukherjee. 2013. Similar Days in the NAS: an Airport Perspective. *AIAA Aviation Technology, Integration, and Operations Conference*.

Hassin, Refael, Asaf Levin, and Dana Morad. Lexicographic local search and the p-center problem. *European Journal of Operational Research* **151**.

Ho, Chin Kuan, Yashwant Prasad Singh, and Hong Tat Ewe. 2006. An enhanced ant colony optimization metaheuristic for the minimum dominating set problem. *Applied Artificial Intelligence* **20**(10) 881–903.

Hochbaum, Dorit S and David B Shmoys. 1985. A best possible heuristic for the k-center problem. *Mathematics of operations research* **10**(2) 180–184.

Bibliography III

- Ilhan, T, FA Özsoy, and MC Pinar. 2002. An efficient exact algorithm for the vertex p-center problem and computational experiments for different set covering subproblems. *Bilkent University, Department of Industrial Engineering, Technical Report*.
- Jain, Anil K, M Narasimha Murty, and Patrick J Flynn. 1999. Data clustering: a review. *ACM computing surveys (CSUR)* **31**(3) 264–323.
- Liu, Pei-chen Barry, Mark Hansen, and Avijit Mukherjee. 2008. Scenario-based air traffic flow management: From theory to practice. *Transportation Research Part B: Methodological* **42**(7) 685–702.
- Mihelič, Jurij and Borut Robič. 2003. *Approximation algorithms for the k-center problem: an experimental evaluation*. Springer.
- Minieka, Edward. 1970. The m-center problem. *Siam Review* **12**(1) 138–139.
- Mladenović, Nenad, Martine Labbé, and Pierre Hansen. 2003. Solving the p-Center problem with Tabu Search and Variable Neighborhood Search. *Networks* **42**(1) 48–64.

- Parekh, Abhay K. 1991. Analysis of a greedy heuristic for finding small dominating sets in graphs. *Information processing letters* **39**(5) 237–240.
- Robič, Borut and Jurij Mihelič. 2005. Solving the k-center problem efficiently with a dominating set algorithm. *CIT. Journal of computing and information technology* **13**(3) 225–234.
- Sanchis, Laura A. 2002. Experimental analysis of heuristic algorithms for the dominating set problem. *Algorithmica* **33**(1) 3–18.
- Tan, Pang-Ning, Michael Steinbach, Vipin Kumar, et al. 2006. *Introduction to data mining*. Vol. 1. Pearson Addison Wesley Boston.