Power Restoration Prioritization with Clustering Algorithms

Machine Learning

Paul Kim

Abstract:

Electricity companies must be prepared to handle power outages of varying severity and scale. To plan power restoration in affected areas, these companies must take into account both context dependent factors and location-based factors. For instance, the former could include flooding or downed power lines caused by a hurricane while the latter could include population density or the number of hospitals in a specific area. This project aims to use location-based factors to determine which areas to prioritize after a major outage. As such, three clustering algorithms have been tested on a dataset compiled from New York State statistics to analyze how counties are separated into priority based clusters. Although this project uses a small set of features and a rather simple metric for calculating priority, it nevertheless provides an cursory overview of how clustering algorithms can aid in power restoration planning.

Introduction:

Power companies follow a general hierarchy for power restoration that begins with fundamental systems and gradually moves down to the local level. High priority systems include power plants, substations, and transmission lines, all of which are vital for generating and distributing power to numerous customers. Following these critical systems are public facilities such as hospitals, communication stations, police stations, supermarkets, and so forth. Focus shifts to neighborhoods and individual households once these major restorations are completed.

The motivation behind this project stems from an interest in applying Machine Learning algorithms to support power restoration efforts. Given the importance of maintaining the electrical grid to keep public services functioning, it is necessary to focus efforts on locations with greater concentrations of such facilities before progressing to less urgent areas. By categorizing locations based on this metric, clustering algorithms can help create priority maps that are useful for a variety of power outage situations.

Data and Methods:

The provided file, *countyData.arff*, is a compilation of New York State statistics from 2010 and 2016. Sources include the United States Energy Information Administration, Community Health

Advocates, the New York State Department of Health, and the New York State Division of Homeland Security and Emergency Services. The dataset contains 4 numeric features that are organized by county: *population*, *number of fire departments*, *number of fire departments*, and *number of power plants*. The fifth feature, *county*, is included to label each set of numeric attributes; it is ignored when running algorithms.

These features were chosen not only because they represent a subset of location-based factors, but also due the fact that for each feature larger values imply higher priority. For example, a county with many power plants and fire departments would likely be placed into a high priority cluster. By regarding each centroid as a vector in 4-dimensional space, priority can be quantified by calculating the magnitude of each vector, then sorting clusters from highest to lowest magnitude. This method will be used to determine the priorities of each algorithm's cluster assignments.

The feature values were first normalized between 0 and 1 to balance the dataset. Three clustering algorithms were utilized to partition the counties into priority based clusters. Details for the algorithms are described below:

a) K-means

- As one of the simplest clustering algorithms, K-means provides a straightforward approach for determining cluster assignments based on the distance of points from centroids. Two distance functions were used: Euclidean distance and Manhattan distance. For reference, the Euclidean distance of two points is the magnitude of the line segment formed between them, while Manhattan distance is based on the sum of horizontal and vertical components between the points within a grid structure. Since the latter is often used for clustering with high dimensional data, it is worthwhile to compare the results of the two distance functions.
- When choosing the number of centers (K) to test on the dataset, the distribution of the feature values was taken into consideration. For example, since locations such as Kings County and Queens County contain large proportions of New York State's population, they are much more likely to be placed in high priority clusters that emphasize large population values. Given the skewed nature of the data, a small range of K values ($K = 2, 3, 4$) was chosen to ensure counties were not placed into clusters with minimal differences in priority.

b) Expectation Maximization

- The Expectation Maximization (EM) algorithm focuses on finding local maximum likelihood parameters for some statistical model or probability distribution. Put simply, it is an iterative probability based clustering algorithm that calculates the likelihood of each set of features belonging to particular clusters. Unlike K-means, it does not specify the number of centers prior to being run on the dataset. As EM is considered a robust algorithm that is widely used for problems such as clustering or pattern recognition, comparing its cluster assignments to those of K-means and the Hierarchical Clusterer will allow for comparisons between the probability and distance based algorithms.

c) Hierarchical Clusterer

- Rather than grouping counties together into clusters based on arbitrarily initialized centroids as in K-means, the Hierarchical Clusterer organizes clusters into a stratified tree structure. This ordering offers a straightforward interpretation of priority since counties that are closer to the root of the tree are considered higher priority than those near the base.
- Because the link type – the technique used to calculate the proximity between two clusters – greatly influences the results of agglomerative clustering, the data was analyzed to determine the appropriate link method. As mentioned above, the feature values are unbalanced due to the presence of counties with large proportions of New York State's population, power plants, and so forth. To account for this, the Hierarchical Clusterer was run using group average, or the mean pairwise distance between counties of two clusters, which is generally less susceptible to outliers than other link types.
- Like K-means, the Hierarchical Clusterer was tested with Euclidean and Manhattan distance functions. It also uses the same number of centers ($K = 2, 3 ,4$).

Results:

*Note: The maps for each algorithm's clustering results are in **Report Files/Maps**. The result files with cluster labels are also provided in **Report Files/Result Datasets**.*

a) Similarities:

- One notable similarity among all algorithms is the fact that the highest priority cluster tends to include counties in the New York Metropolitan area. This is to be expected, since these counties hold a large proportion of the state population. Suffolk County in particular appears in the highest priority cluster for nearly every algorithm due to the fact that it has both the highest number of fire departments (108) and highest number of power plants (40) across all counties. In general, higher priority clusters seem to contain less counties than lower priority clusters, which is reasonable given the unbalanced nature of the feature distribution.

- The maps for Expectation Maximization and K-means for *K* = 4 with both Euclidean and Manhattan distance have similar cluster assignments. The highest priority for both algorithms contain counties in the New York Metropolitan area along with Erie County, which has two relatively significant feature values: number of fire departments (91) and number of hospitals (10). The fact that these counties have been placed into the highest priority cluster for both EM and K-means using two distance functions suggests they take precedence over other locations in power outage situations.
- The differences between the clustering results of K-means using Euclidean distance and Manhattan distance are very subtle. For example, in the *K* = 3 results, there are only two more counties in the lowest priority cluster using Euclidean distance than with Manhattan distance. A similar trend is observed in the Hierarchical Clusterer's results, implying the distance functions do not significantly affect clustering assignments for either algorithm.

b) Differences:

- The most significant difference between the Hierarchical Clusterer and K-means is the number of counties in their respective clusters. For example, in the results of *K* = 4 using Euclidean distance, the former's higher priority clusters have dramatically less counties than the latter's. This should not be considered a concrete trend, however, since factors such as the random initialization of centroids in K-means may cause cluster assignments of additional tests to vary.
- Unlike EM and K-means, the Hierarchical Clusterer seems to place less emphasis on the general New York Metropolitan area. With increasing *K* values, it gradually decreases the priority levels of commonly high priority locations such as Kings County or Nassau County. In fact, by K = 4, it assigns only Suffolk County has the top priority location while placing most other counties in the lowest priority cluster. As mentioned above, Suffolk County has the highest values for two features, meaning it will almost always be included in the top priority cluster for each algorithm. However, this trend might be due to the fact that the Hierarchical Clusterer merges clusters based on proximity, which is determined by pairwise averages between points. Even though average link tends to be less susceptible to outliers, the extremely unbalanced nature of the dataset may be influencing the algorithm's outputs.

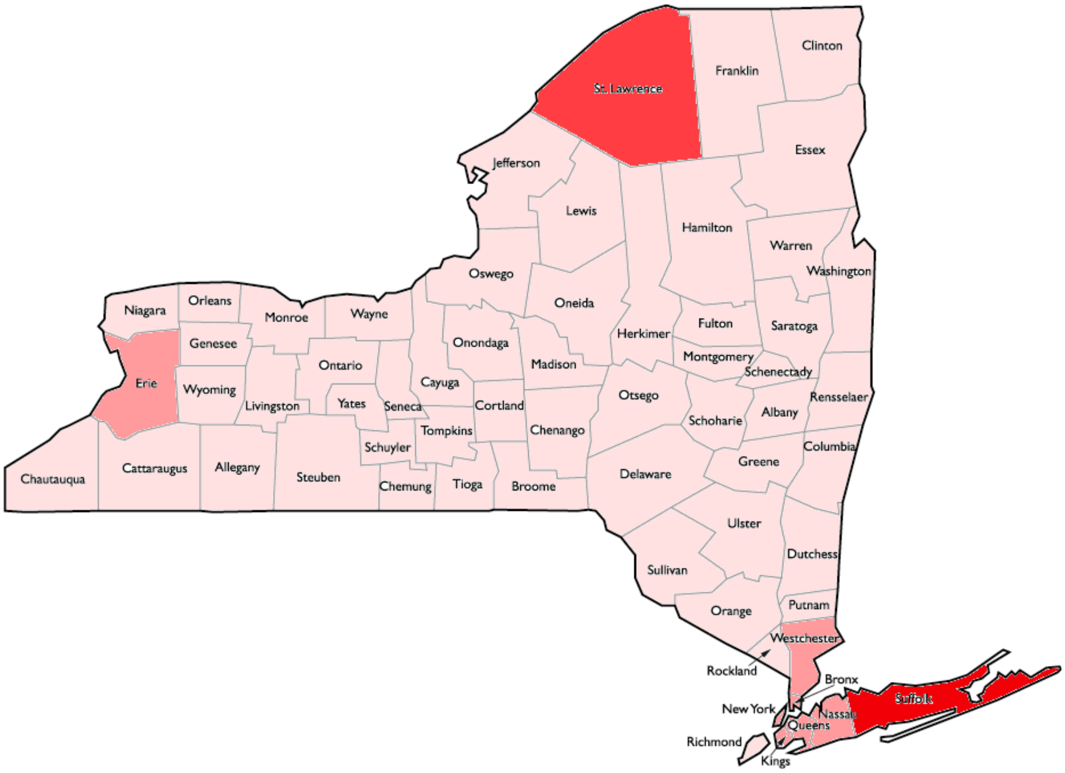

Fig 1. Hierarchical Clusterer with K = 4 using Euclidean distance

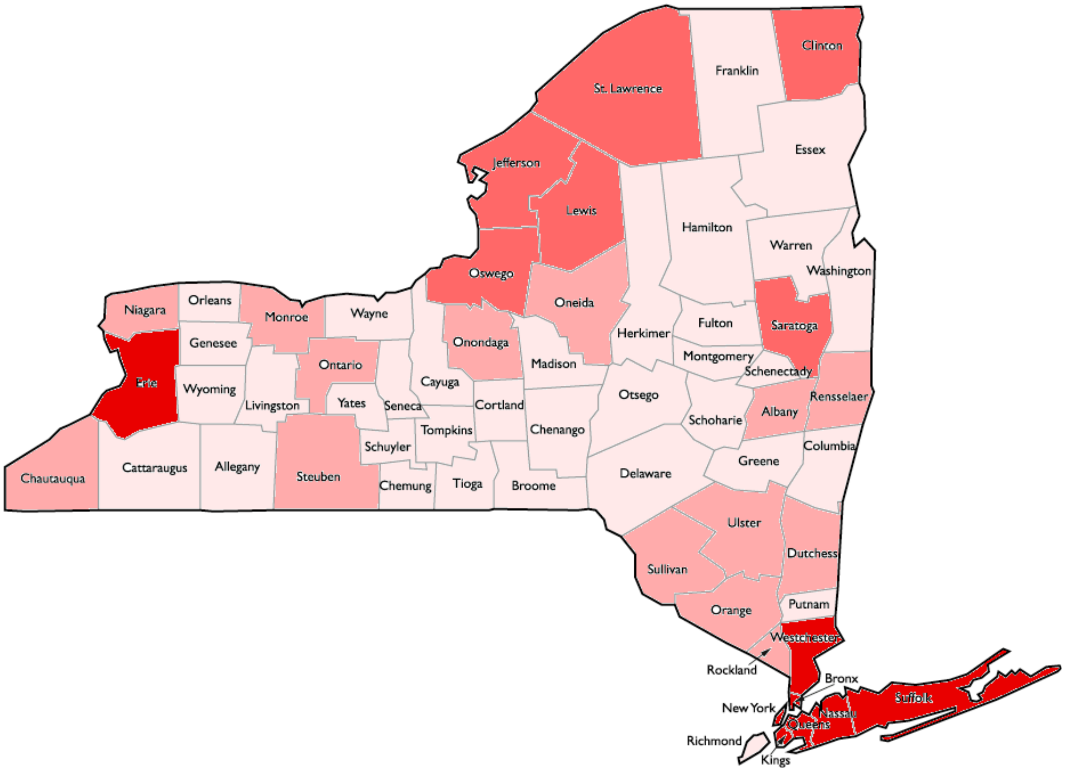

Fig 2. K-means with K = 4 using Euclidean distance

Note the tendency of the Hierarchical Clusterer to have less counties in higher priority clusters, along with its emphasis on Suffolk County.

Conclusion:

There are several limitations in the data and methods. The choice to include only four features in the dataset was based mainly on two factors: avoiding the curse of dimensionality and difficulty in finding certain statistics organized by county. For instance, although it was originally planned to include factors such as transmission line distribution and areas serviced by different power companies, these features are often represented with latitudinal and longitudinal coordinates rather than discrete locations. Given the timespan of this project, it was not feasible to write a script to arrange geospatial data by county. Furthermore, taking service areas of various power companies into account would have invalidated the method for determining priority since it requires all features to be numeric.

In terms of the methods, the decision to limit *K* to 2, 3, 4 for K-means and the Hierarchical Clusterer may have been too restrictive. Although the results show that high priority clusters tend to have only a few counties, the lower priority clusters often contain a disproportionately larger number of counties. Whether or not this is an accurate representation of areas to focus on during power restoration is difficult to determine, but additional features, namely context dependent factors, may endorse the use of higher *K* values. Similarly, representing priority by the magnitude of a cluster was also restrictive, as it prevents non-numeric attributes and numeric features in which larger values do not necessarily imply higher priority from being included in the dataset.

Despite these issues, the project can be expanded upon in certain ways to provide more robust results. For example, real time data can provide context dependent factors in situations such as hurricanes or blizzards that are currently absent in the dataset. The use of more sophisticated algorithms, along with a better heuristic for deciding attributes that are relevant in power outages, could also lead to better clustering assignments. In any case, the project serves to demonstrate how Machine Learning algorithms can analyze location-based data to create priority maps for generalized power outage scenarios. Given the importance of keeping the power grid active, these algorithms can help power restoration efforts by offering insight into which areas might require more attention.

Resources:

How power companies restore power:

https://www.nationalgridus.com/Upstate-NY-Business/Storms-Outages/FAQs

https://www.fpl.com/storm/restoration/restoration-priorities.html

Power plant data:

https://www.eia.gov/electricity/data/eia860/index.html

Hospital data:

http://communityhealthadvocates.org/health-care-options/help-paying-for-care/hospital/list

Population data:

https://data.ny.gov/Government-Finance/Census-2000-and-2010-Population-Towns/fqf5-9nc2/data

Fire department data:

https://data.ny.gov/Public-Safety/Fire-Department-Directory-for-New-York-State/qfsu-zcpv

Euclidean distance vs Manhattan distance:

http://www.ijorcs.org/uploads/archive/distance-measuring-approaches-for-clustering.pdf

Expectation Maximization:

https://docs.rapidminer.com/latest/studio/operators/modeling/segmentation/expectation_maximization_clustering.html