Genetic Algorithms for Cluster Analysis: Health Insurance Coverage and Pollution

Delaney Scheiern – Colgate University

Project Description

This project introduces Genetic Algorithms for **Cluster Analysis**, aided by a discussion of more simplistic clustering techniques such as **k-Means** and **Gaussian Mixture Model** clustering. The genetic algorithm implemented in this project focuses on a density-based initialization of potential solutions, followed by distance-based optimization. The k-Means, Gaussian Mixture Model, and Genetic Algorithm methods are demonstrated in an application of **environmental conditions** in the United States for clusters created from **health care coverage** statistics. The project explores the performance of these methods on data with **non-distinct clusters**.

Introduction

Cluster analysis is a subset of Machine Learning that arranges observations into different groups without user input. This technique has been applied in a variety of areas, from identifying someone's asthma type [4] to predicting which soccer position someone will excel at based on their skills [5].

Clustering Techniques

Clustering Tendency

Not all data can be grouped into meaningful clusters. The **Hopkins' statistic** can be used to determine how much the dataset diverges from a uniform dataset that cannot be clustered:

$$H = \frac{\sum U_i}{\sum U_i + \sum W_i}$$

where U is the distance from a real point to its nearest neighbor and W is the distance from a randomly chosen point within the data space to the nearest real data point [2]. A Hopkins' value near 1 indicates high clustering tendency, while 0.5 indicates overlapping clusters.

k-Means

k-Means is a distance-based clustering method where observations are divided into k clusters. The algorithm finds optimal cluster centroids by minimizing the **Sum of Squared Errors (SSE)**,

$$SSE(C) = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - c_k\|$$

between each observation and their cluster centroid until the centers of each cluster converge [1]. Although k-Means is the most common clustering method, its major downfall is that it is restricted to hyperspherical clusters.

Gaussian Mixture Models

Gaussian Mixture Models (GMMs) assign specific observations to Gaussian distributions, and optimize the parameters for those distributions. A GMM with k distributions, or clusters, is defined by a **linear combination of normal distributions**,

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

where $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$ [2]. Then, the **Expectation-Maximization Algorithm** is used to optimize the distribution parameters [2]. GMMs can more accurately define clusters that are in close proximity and oddly shaped than k-Means.

Genetic Algorithms

Genetic Algorithms are highly customizable and use Charles Darwin's concept of **natural selection** and survival of the fittest to solve **optimization problems**. We begin with a group of chromosomes that represent the first generation, then undergo crossover and mutation to simulate the evolutionary process. Each new generation further explores the solution space and prioritizes the fittest genes.

Genetic Algorithms for Clustering

Chromosome Representation and Assignment

Chromosomes are each **potential solutions** of the optimization problem, and they are defined by a set of cluster centers. In Figure 1, each row is a chromosome and each purple box contains a cluster center. Chromosomes can be various lengths, signifiying solutions with various numbers of clusters. This representation allows the algorithm to find the best number of clusters rather than it being supplied a priori. The **density** of observations is used to initialize the cluster centers.

Fitness Function

The fitness function is the objective that needs to be optimized. This algorithm calculates

$$f = \frac{1}{SSE + \left(\frac{\text{number of clusters in chromosome}}{2}\right)^2}$$

for each chromosome in a given generation. By maximizing this equation, the distance of observations from their cluster centers is minimized while penalizing a large number of clusters.

Selection: Crossover and Mutation

Selection consists of crossover, mutation, fitness computations, and elitism, and it decides future generations. Crossover and mutation add **variability to the gene pool** so that more possible solutions are explored. Figure 1 demonstrates the single-point crossover and mutation procedures used in this algorithm. The crossover procedure switches the ends of a pair of chromosomes after certain cutoff points. Mutation adds a random scalar to a random cluster center in a chromosome. Both occur with **adaptive probabilities**, so crossover is more likely for the best chromosomes and mutation is more likely for the worst chromosomes. Thus, the best chromosomes' genes are spread throughout the population while bad genes are changed in hopes of finding a better solution.

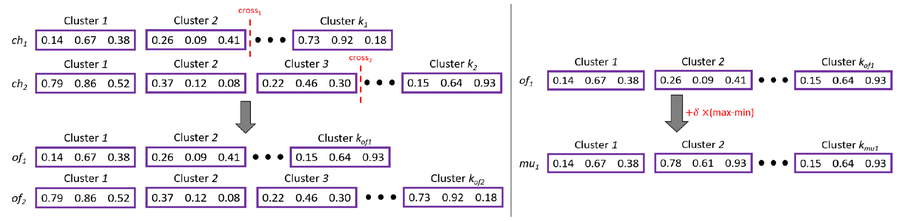

Fig. 1: Crossover and Mutation procedures

References

[1] Charu C. Aggarwal and Chandan K. Reddy, eds. *Data Clustering : Algorithms and Applications*. CRC Press LLC, 2013.

[2] Marc Peter Deisenroth, A. Aldo Faisal, and Cheng Soon Ong. *Mathematics for Machine Learning*. Cambridge University Press, 2020.

[3] U.S. Center for Disease Control and Prevention. *National Environmental Public Health Tracking Network*. Retrieved through online Database. URL: `https://ephtracking.cdc.gov/DataExplorer/`.

[4] Pranab Haldar et al. "Cluster Analysis and Clinical Asthma Phenotypes". In: *American journal of respiratory and critical care medicine* 178 (June 2008), pp. 218–24. DOI: `10.1164/rccm.200711-1754OC`.

[5] César Soto-Valero. "A Gaussian mixture clustering model for characterizing football players using the EA Sports' FIFA video game system". In: *RICYDE. Revista internacional de ciencias del deporte* 13 (July 2017), pp. 244–259. DOI: `10.5232/ricyde2017.04904`.

[6] U.S. Census Bureau. *2019 American Community Survey 1-Year Data*. Retrieved through Census API.

Data

Health Insurance Coverage data was collected from the American Community Survey (ACS) from the Census Bureau, including totals for employer-based, Veterans Affairs, Medicare, Medicaid, private, and no health insurance coverage for each county in the U.S. [6]. Environmental pollution data was collected from the CDC National Environmental Public Health Tracking Network for each U.S. county, including average concentrations of PM 2.5, Benzene, Acetaldehyde, and Formaldehyde [3].

Results

k-Means and Gaussian Mixture Model Results

The Davies-Bouldin Index (DBI) and Silhouette Index (SI) aim to minimize the intra-cluster dispersion and maximize inter-cluster distances. A value near 0 for DBI and near 1 for SI indicate well-defined clusters. Figure 2 suggests the optimal number of clusters for the healthcare data is near 10.

Fig. 2: DBI and SI for k-Means and GMMs with different numbers of clusters ranging from 2 to 50.

Genetic Algorithm Results

The GA converged after 15 generations to a solution with 10 clusters, as was predicted from Figure 2. Violin plots of the four environmental pollutants across clusters are shown in Figure 3. A one-way ANOVA was conducted for each pollutant, and each test found a statistically significant difference in means across clusters.

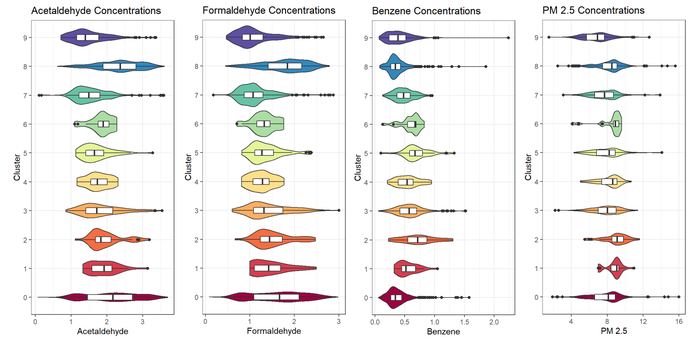

Fig. 3: Violin plots of environmental pollution

GAs are highly maleable and can be applied to a variety of applications, each warranting its own detailed alterations of each step in the algorithm. The algorithm was able to create clusters from health insurance coverage data that had distinct environmental pollutant values.