Clustering Using K-Means

Nov 14th, 2024

Project Overview: This project analyzed electricity consumption data to identify patterns among clients, with a focus on uncovering behavior that could support energy management and forecasting. The dataset included 15-minute interval readings from January 2011 to December 2014, capturing the demand of 370 clients. The primary task was to categorize clients into clusters based on their typical usage patterns using KMeans clustering.

Data Preparation and Preprocessing:

1. **Data Source:** The data was sourced from the UCI Machine Learning Repository.
2. **Cleaning and Formatting:** Data cleaning involved handling missing values, converting data formats, and normalizing usage levels to focus on behavioral patterns rather than absolute consumption.
3. **Focus Years:** For consistency, only data from 2013-2014 was used due to fewer clients with zero demand, providing a cleaner dataset.

Clustering Methodology:

1. **Normalization:** Each curve was normalized to mean 1 to eliminate scale differences.
2. **Cluster Determination:** The optimal number of clusters was determined using both inertia (elbow method) and silhouette scores, resulting in an initial recommendation of 5 clusters. Upon further refinement, 4 clusters were selected to avoid outliers.
3. **Modeling Techniques:** KMeans and Agglomerative Clustering were applied. For evaluation, both silhouette and inertia metrics were plotted against cluster numbers.

Findings:

1. **Cluster Analysis:** The final clusters revealed distinct patterns, such as weekday consumption versus weekend/holiday consumption, with some clusters showing morning peaks while others had evening peaks.
2. **Daily Analysis:** For individual clients, clustering based on daily curves identified two main patterns, likely linked to weekday and weekend behaviors.
3. **Implications:** These patterns offer insights for personalized energy management, demand forecasting, and targeted energy-saving recommendations.

Technical Skills Utilized:

- **Python (Pandas, NumPy):** For data processing and cleaning.
- **Scikit-Learn:** For clustering and silhouette analysis.
- **Matplotlib:** For plotting consumption curves and cluster centers.

Outcome and Impact: The clustering results enable better understanding of energy usage, supporting predictive models for demand forecasting. This can benefit energy providers in delivering customized recommendations, enhancing energy efficiency, and optimizing demand response strategies. The project also highlights the potential for further segmentation based on seasonal or special event impacts on consumption.