Methodology

Kathy Mo, Riyaadh Bukhsh, William He

Overall summary

This paper discusses the implementation of the Gap statistic, which is a method of finding the estimated optimal number of clusters for clustering algorithms like K-means and hierarchical clustering. The Gap statistic compares the total within-cluster variation for different numbers of clusters with their expected values under the null reference distribution of the data. The optimal number of clusters is the value k that maximizes the Gap statistic. The Gap statistic in a general form is the following:

$$\text{Gap}_n(k) = \mathbb{E}_n^*(\log(W_k)) - \log(W_k).$$

The studies done in the paper show that the Gap statistic usually performs better than other methods. The Gap statistic is a method that estimates the number of clusters in a dataset. It has the same goal as the elbow method, which is to find a distinctive difference in the measure of within-cluster variation. The resulting estimated number of clusters is applicable to any clustering method, such as K-means clustering, and any distance measure.

Purpose and Rationale

In the equation:

$$\text{Gap}_n(k) = \mathbb{E}_n^*(\log(W_k)) - \log(W_k),$$

k is the estimated optimal number of clusters, $\mathbb{E}_n^*$ is the expectation under sample size n of a reference distribution (which we will derive), and $\log(W_k)$ is the log of the pooled within-cluster sum of squares around the cluster means. The Gap statistic method compares the within-cluster dispersion of the observed data with a reference null distribution to identify the clusters that maximize the difference. The underlying rationale of the Gap statistic is that by comparing the data's clustering structure to the null hypothesis, which is a dataset with no obvious clustering structures, it measures how much the data's clustering structure strays from randomness. Doing this means that a significant gap indicates a more meaningful clustering structure.

Algorithm Versions

1. Uniform Feature Selection from Observed Values

One version of the algorithm uses a uniform reference distribution over the range of the observed data for the null distribution. It generates each reference feature uniformly. One way to do this is my identifying the range of each feature and generate values Uniformly. Example: Suppose observations $X_1, X_2, ..., X_n \in \mathbb{R}^p$. We find the min and max of each feature.

$$\text{For } j = 1, 2, .., p \quad \min = \min_{i=1}^{n} X_{ij} \quad \max = \max_{i=1}^{n} X_{ij}$$

We then generate new values for feature j from the uniform distribution $U(\min_j, \max_j)$ This version is simpler.

2. Principal Component Projections

Another version of the algorithm uses a reference distribution based on principal component projections. This is computed by assuming $\mu = 0$ and decomposing X into its singular value form, where $X = UDV^T$. Then by using $X' = XV$, we draw features uniformly distributed from the columns of X' called Z'. By transforming back Z' to $Z = Z'V^T$, which produces our reference data Z. These methods generate the null distributions (reference data) which will be utilized to create a more accurate comparison for the Gap statistic. The reference data is used in tandem with Monte Carlo to estimate.

$$\mathbb{E}_n^*(\log(W_k))$$

This is done by averaging from the B copies of $\log(W_k^*)$. We then have:

$$(1/B)\sum_b(\log(W_k) \approx \mathbb{E}_n^*(\log(W_k))$$

Rationale of Subtracting s_k

$$s_k = sd_k\sqrt{(1 + 1/B)}$$

$$Gap(k) = \frac{1}{B}\sum_{i=1}^{B}\log(W_{ki}^*) - \log(W_k^*)$$

We select $\hat{k}$ such that

$$\hat{k} = \text{smallest k} : Gap(k) \ \geq Gap(k+1) - s_{k+1}$$

Subtracting s_k is to correct the variability from the Monte Carlo simulations used to estimate the reference distribution. The term itself comes from accounting for the simulation error in $\mathbb{E}_n^*(\log(W_k))$. The factor $\sqrt{(1 + \frac{1}{B})}$ increases the standard deviation for a more conservative estimate. This factor adjusts for bias in the estimated expectation of $\log(W_k)$, which is from the finite number of simulations, B.

Theorem 1 Interpretation

Among all unimodal distributions, the uniform distribution is the most challenging for the Gap statistic because it is the most likely to suggest non-existent clusters. Therefore, this theorem states that the Gap statistic is more robust for structured data distributions compared to the uniform distribution.

Simulation Results

From the simulation results, we see that the Gap statistic performs well in different scenarios. It effectively identifies the correct number of clusters. Gap/pc has an advantage over Gap/unif because Gap/pc better adapts to the data structure and has a higher sensitivity to the true number of clusters. For example, when the authors simulated two elongated clusters in three dimensions, the Gap/unif is adversely affected by the oblong shapes, while Gap/pc performs better. Both used a uniform reference distribution.