Evaluating Clustering Techniques on Non-elliptical Data

Safiya Sirota, Yijin Wang, Bin Yang

Contents

1 Background and Objective **1**

2 Statistical Method **2**

2.1 K Means 2

2.2 Latent Class Analysis 2

3 Performance Metric **2**

4 Simulations Settings and Results **2**

4.1 Skewed Data - Data Generation 2

4.2 Skewed Data - Results 3

5 Conclusion and Discussion **4**

6 References **4**

1 Background and Objective

Latent Class Analysis and K-means are two successful clustering algorithms that partition the given data into distinct subgroups where observations in each group are very similar. They are popular choice of method in solving healthcare research problems such as identifying clusters on microscopic images (Amin et al. 2015), categorizing suicides by risk factors (Logan, Hall, and Karch 2011), and grouping elderly patients with their access to health care (Thorpe et al. 2011). One important assumption that both of these algorithms require is the normality of the given data.

In our project, we focus on the unsupervised learning scenario and we investigate how well these two algorithms perform when the normality assumption is violated. We designed 4 simulation settings as violation of the normality assumption: 1) skewed data, 2) data with heavy tail, 3) data with outliers, 4) multimodal data. In the end, we compare their performance with Rand Index in each settings with varying sample sizes.

2 Statistical Method

2.1 K Means

2.2 Latent Class Analysis

3 Performance Metric

4 Simulations Settings and Results

In simulations, we consider bivariate data with 2 true clusters. We generated data from different non-elliptical distributions in each setting.

For each simulation setting, we create 100 runs using randomly generated data sets in size of 500, 1000, 5000 and evaluate the Rand index for each method.

4.1 Skewed Data - Data Generation

In this setting, we want to test the performance of K-means and LCA when all features follow a skewed distribution. We assume the features X_1, X_2 are independent.

All data in the first feature $X_{11}, X_{12}, ..X_{1N}$ are random draws from a mixture model of Weibull distributions with density $0.5f_1(x|1,5) + 0.5f_2(x|12,14)$ where $f_1 \sim Weibull(1,5)$ and $f_2 \sim Weibull(12,14)$.

Similarly, all data in the second feature $X_{21}, X_{22}, ..X_{2N}$ are random draws from another mixture model of Weilbull distributions with density $0.5f_3(x|1,3) + 0.5f_4(x|1,4)$ where $f_3 \sim Weibull(1,3)$ and $f_4 \sim Weibull(1,4)$.

We generate 2 equal-sized clusters with total size being 500, 1000 and 5000. An example of a sample with size 5000 is shown below.

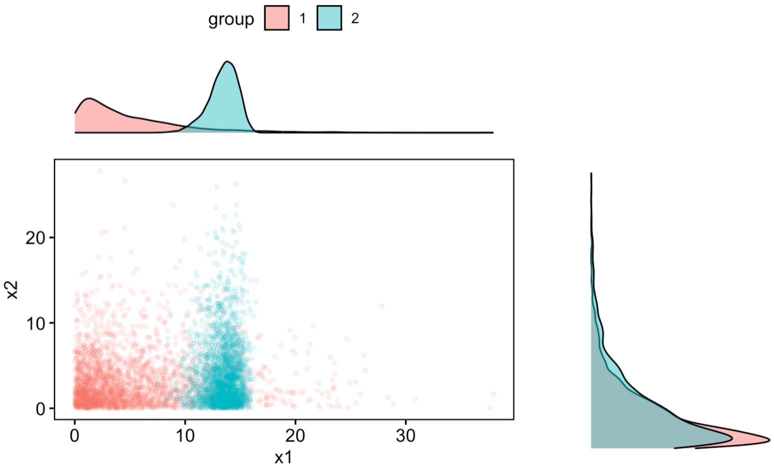

Figure 1: Skewed Data Sample

4.2 Skewed Data - Results

We are interested in how well K-means and LCA perform when the input data already has two true clusters. We calculate Rand Index for sample data in each simulation run and conclude that after 100 simulations, K-means performs better than LCA across all sample sizes. As sample size increases, the variance of Rand Index decreases.

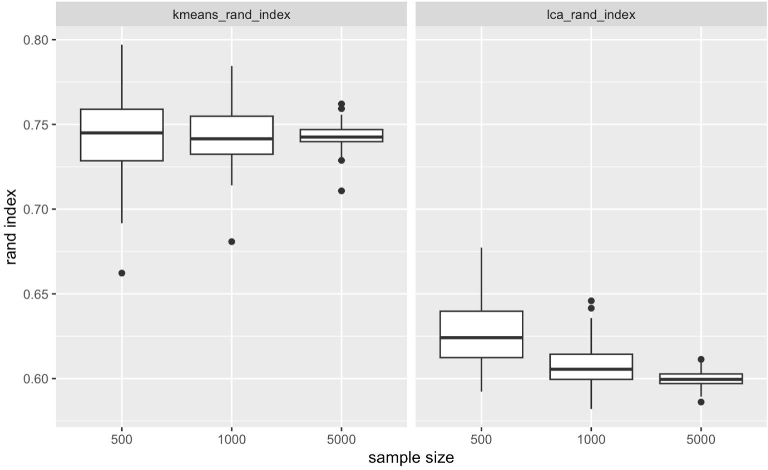

Figure 2: Skewed Data Comparison - Rand Index

To investigate some potential reasons behind this, we compare the optimal number of cluster number for each method in all simulations. Choosing the optimal cluster number is the first step for both methods. This number is crucial to our comparison because we predetermined the number of true clusters to be 2 and Rand Index uses the number of agreements in its calculation. Therefore, having a cluster number that is larger than 2 could hurt the performance.

In this figure, we compare the optimal number of clusters. Overall, K-means choose its optimal number of cluster around 3 which is much lower than that of LCA across all sample sizes. As sample size increases, the optimal number of cluster increases for LCA as well.

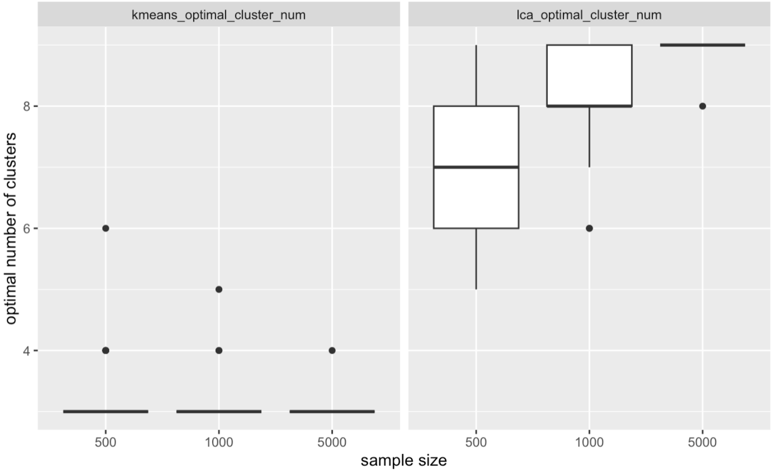

Figure 3: Skewed Data Comparison - Number of Optimal Clusters

5 Conclusion and Discussion

6 References

Amin, Morteza Moradi, Saeed Kermani, Ardeshir Talebi, and Mostafa Ghelich Oghli. 2015. "Recognition of Acute Lymphoblastic Leukemia Cells in Microscopic Images Using K-Means Clustering and Support Vector Machine Classifier." *Journal of Medical Signals and Sensors* 5 (1): 49–58. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4335145/.

Logan, Joseph, Jeffrey Hall, and Debra Karch. 2011. "Suicide Categories by Patterns of Known Risk Factors: A Latent Class Analysis." *Archives of General Psychiatry* 68 (9): 935–41. https://doi.org/10.1001/archgenpsychiatry.2011.85.

Thorpe, Joshua M., Carolyn T. Thorpe, Korey A. Kennelty, and Nancy Pandhi. 2011. "Patterns of Perceived Barriers to Medical Care in Older Adults: A Latent Class Analysis." *BMC Health Services Research* 11 (1): 181. https://doi.org/10.1186/1472-6963-11-181.