NC STATE UNIVERSITY

Department of Computer Science

G-means Clustering

Uddhav Bhosle(1), Deepak Patil(2), Satish Gurav(3), Rasika Pande(4), Atefeh Morasali(5), Kumar Nielarshi(6)
North Carolina State University

1. Introduction

One of the primary difficulties in cluster analysis is determining whether the obtained solution actually represents the underlying structure in the data or is merely an artifact of the procedure used to obtain that solution.

Here, we explore an improved algorithm, called **G-means**, for learning the value of k while clustering

It is based on a statistical test for the hypothesis that a subset of data follows Gaussian Distribution.

2. Data description

- The data used in the implementation was in numeric form.
- String input must be converted to numeric form.
- We have made use of two datasets,
  1. Synthetic dataset
  - 300 data points with 2 attributes
  2. ecoli dataset
  - 336 data points with 8 attributes obtained from UCI machine learning repository

3. Implementation

- Within a cluster X, find initial two centroids that are farthest from each other on major axis
- Run k-means on cluster X with k=2 which will give final centers c1 and c2
- Let v = c1-c2, a d-dimensional vector that connects to centers
- Project X onto v: $\boldsymbol{x_i' = (x_i, v)/}\ ||v||^2$ where X' is a 1-dimensional representation of the data.
- Transform X' so that it has mean 0 and variance 1
- Perform **Anderson-Darling** test:
  - Test Statistics:

$$\mathrm{A}^2(\mathrm{Z}) = \frac{-1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}(2\mathrm{i}-1)[\ln(\mathrm{Z}[\mathrm{i}]) + \ln(1-\mathrm{Z}[\mathrm{n}+1-\mathrm{i}])] - \mathrm{n}$$

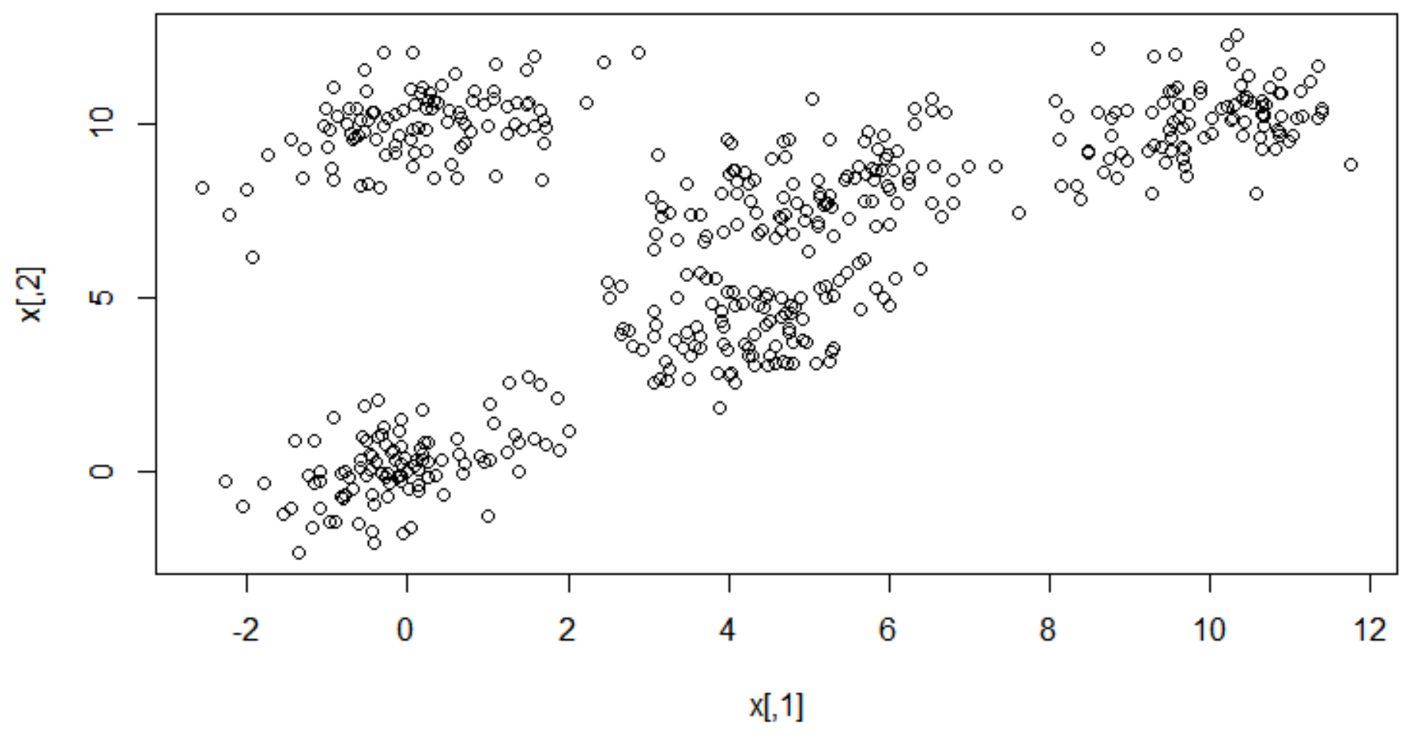
  - We must correct the statistic, $\mathrm{A}_*^2 = \mathrm{A}^2(\mathrm{Z})(1 + 4/\mathrm{n} - 25/\mathrm{n}^2)$
- If $\boldsymbol{A_*^2}$(Z) is in the range of non-critical values at **confidence level α**, then accept H0, keep the original center, and discard {c1, c2}. Otherwise, reject H0 and keep {c1, c2} in place of the original center
- Repeat above process till all the clusters are **Gaussian**

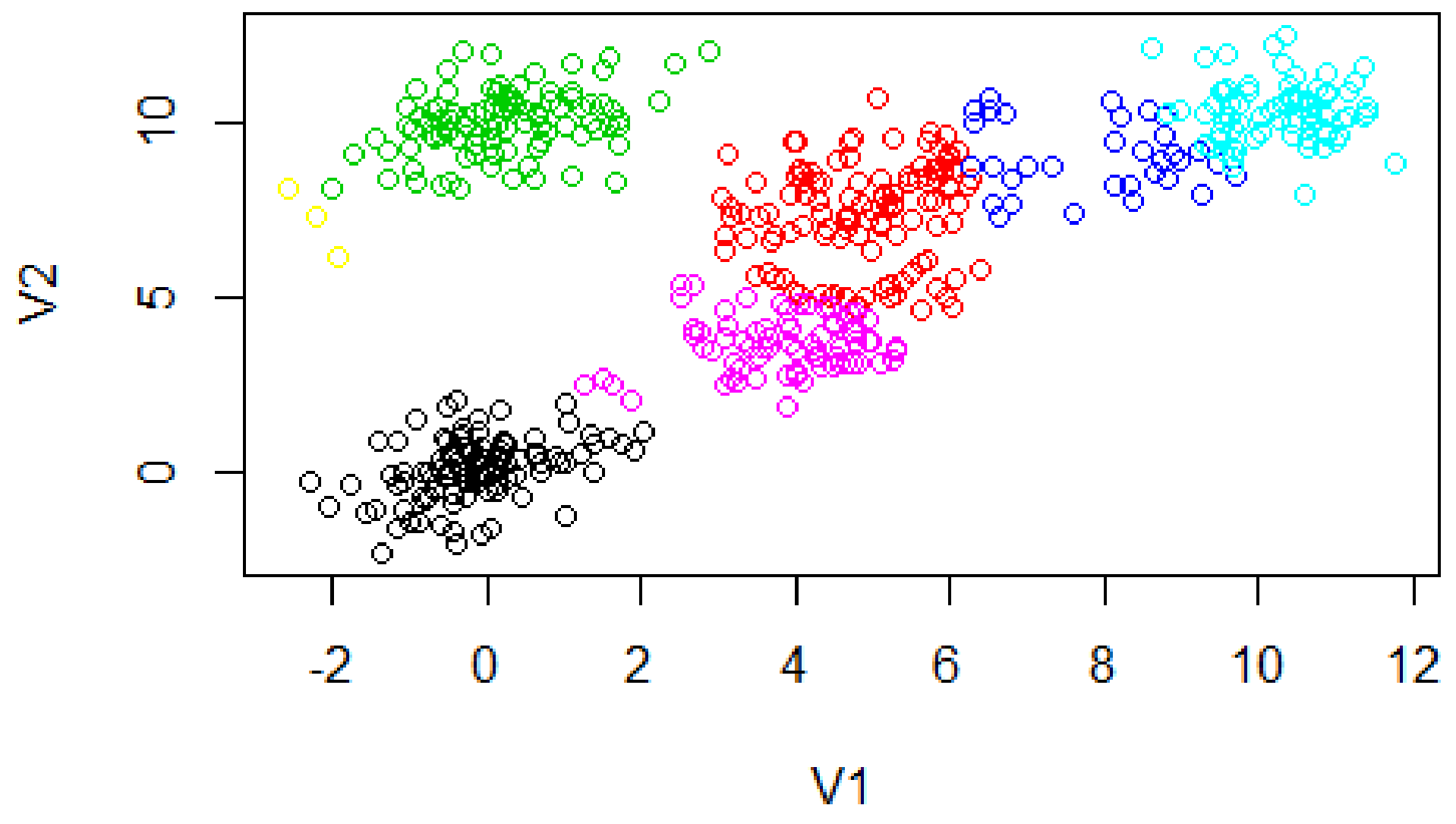
4. Results

1. Synthetic dataset

Predicted value of k = 6

Actual Cluster: 5

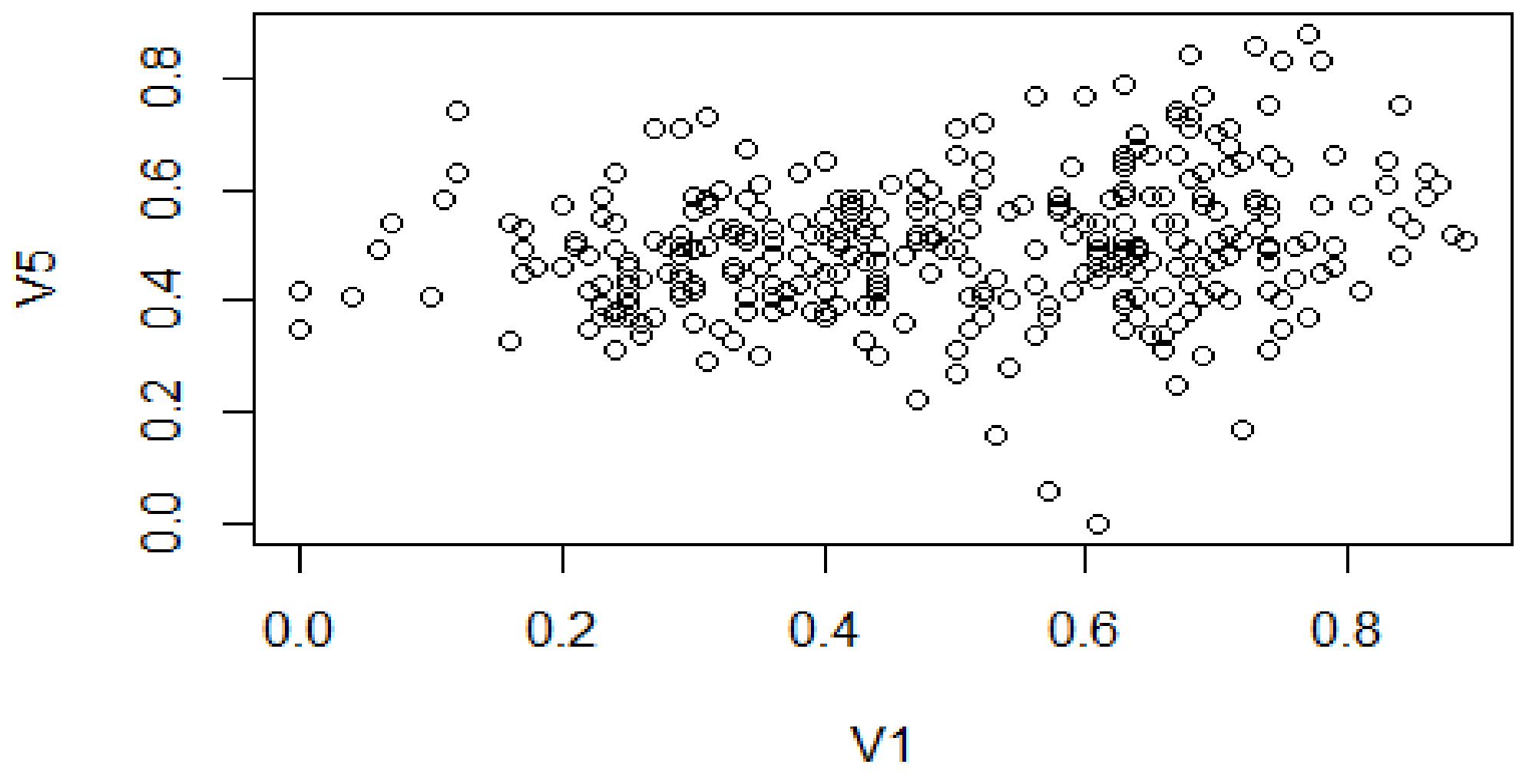


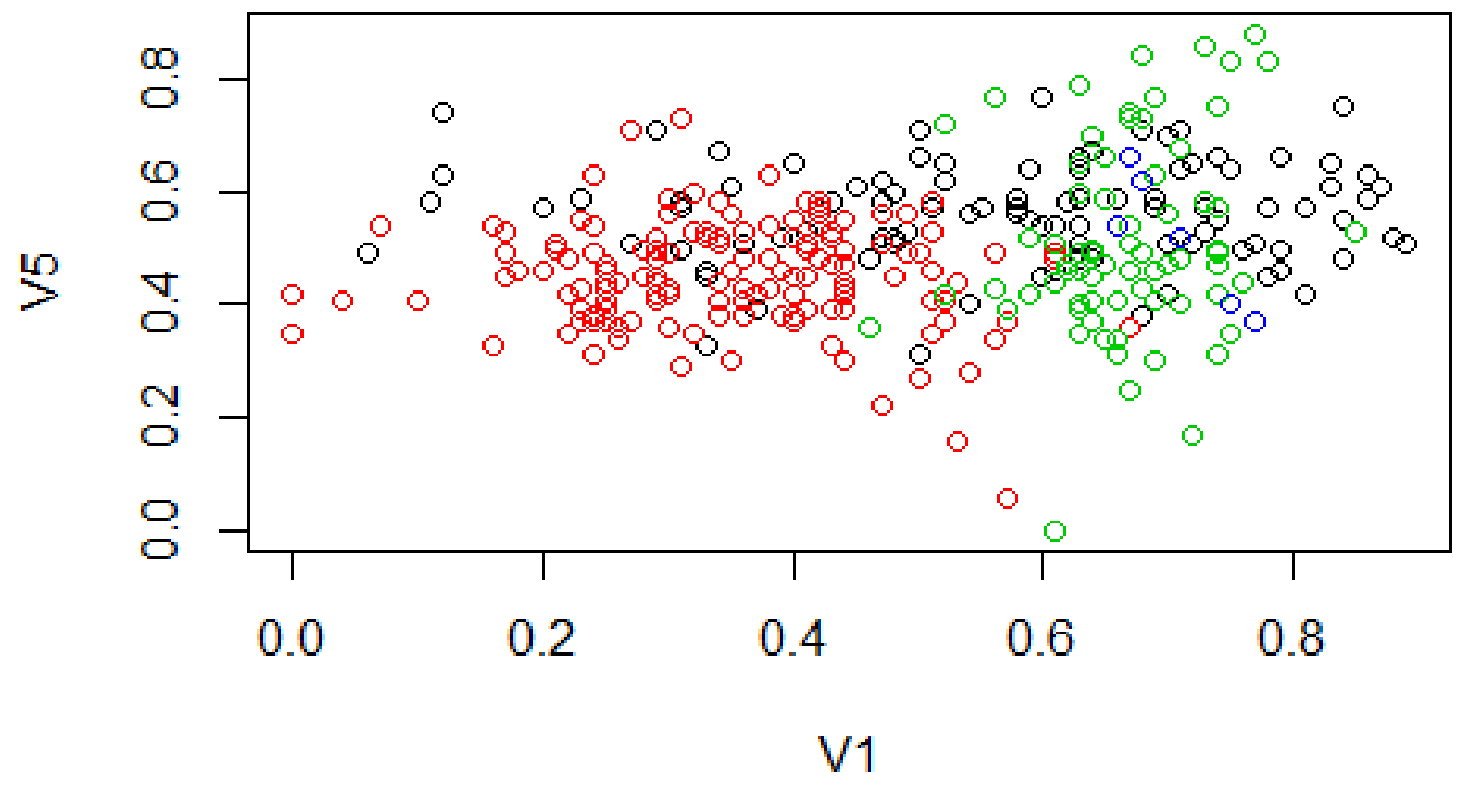


2. Ecoli dataset

Actual Clusters = 8 with 2 clusters of 2 instance

Predicted k = 4





5. Parameter choices

- Confidence $(\alpha) = 0.0001$
- Initial centers: farthest points on axes with maximum variance. Chosen to give two clusters with evenly allocated data points.
- Critical Value for Anderson-Darling test : 1.869
- Cut-off for clustering: 7 data points in a sub set of data

6. Conclusions

G-means clustering performs well in finding the optimal value of k, in k-means clustering algorithm and divides data into clusters such that they each cluster has a normal distribution.

7. References

Greg Hamerly, Charles Elkan, Learning the k in k-means, p.281-288

Dataset:
https://archive.ics.uci.edu/ml/datasets/ecoli

Github:
https://github.ncsu.edu/sjgurav/G-Means