K Means Genetic Algorithm for Wine Dataset Clustering

Bidhan Bashyal, Aarush Mathur

OverView

- K Means Clustering
- Genetic K Means Clustering
- Fitness Function
- Selection
- Crossover
- Mutation
- Dataset
- Result
- Conclusion

K Means Clustering

- Unsupervised data analysis in clustering data.
- It is most widely used clustering algorithm.
- ‘n’ number of centroid i.e. no.of clusters are selected and objects are assigned to the cluster based on different index.

GA K Means Algorithm

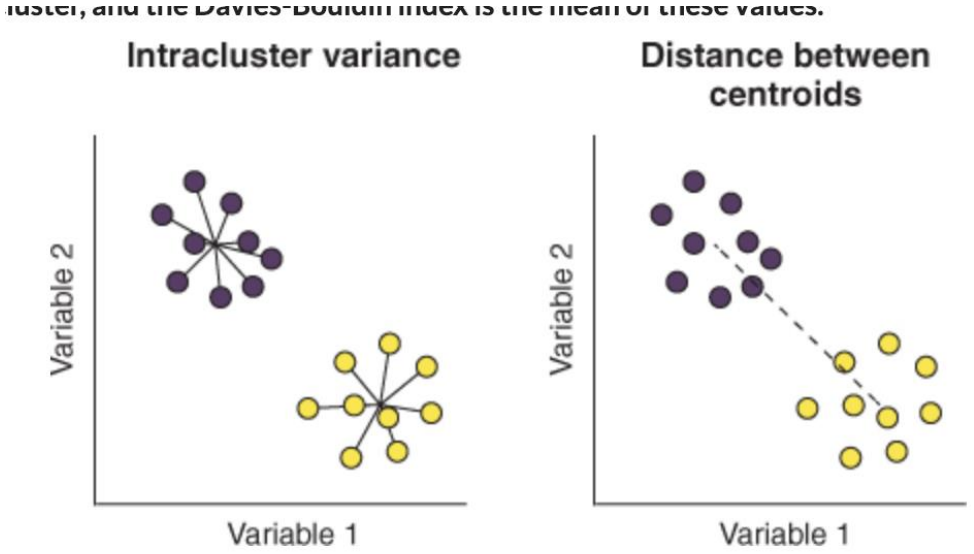
- There have been various GA based K Means algorithm with different variations.
- Our method also uses similar approaches.
- Tried on 2 different datasets.

Normalization of Data

- Data is normalized before used to calculate fitness.
- Normalized using standard minmax scaler.

Fitness Function

- Davis Bouldin Index as an index for evaluation of each cluster.
- DB index performed better compared to other criterion (Silhouette).
- This is based on a ratio between variance within cluster and between cluster variance.



Crossover

- N number of individuals (i.e. population size).
- Chromosome is randomly generated (length of chromosome=no.of features*no. of max clusters)
- One point crossover
- Children chromosomes were generated by cross over of two parent chromosome based on their fitness score (initial indexes of parent chromosome have better fitness score than later).
- Used Crossover rate as a parameter.

Mutation

- Similar to crossover, the mutation rate is also parameter which we provide.
- Performs a search that reassigns objects to cluster at lower distance.
- Remove from one cluster and add to another.

Selection

- Used Rank based selection
- Fitness of the chromosome is calculated
- Every chromosome is allocated selection probability with respect of its rank.
- Chromosome with less Ps is replaced with chromosome with high Ps.

Dataset

- Two datasets
- Wine Dataset for clustering (13 features such as Alchool, Malic Acid, Ash, Alcalinity,etc)
- Iris dataset (4 features)

Results

- Our GA based Kmeans method performed better in Iris dataset – 85.3% accuracy compared to 83.3%.
- However, didn’t performed better on the Wine Dataset with only 62% compared to 82%.

Discussion

- Results show that GA based K means does better in Iris Dataset compared to traditional K means

- High number of features?
- Selection of parameters such as Number of Individuals, Crossover rate, Selection rate and Mutation rate,

References

- 1. Pizzuti C., Procopio N. (2017) A K-means Based Genetic Algorithm for Data Clustering. In: Graña M., LópezGuede J.,Etxaniz O., Herrero Á., Quintián H., Corchado E. (eds) International Joint Conference SOCO'16- CISIS'16-ICEUTE'16. SOCO 2016, CISIS 2016, ICEUTE 2016. Advances in Intelligent Systems and Computing, vol 527. Springer, Cham. https://doi.org/10.1007/978-3-319-47364-2_21
- 2. Krishna, K., Murty, M.N.: Genetic k-means algorithm. IEEE Trans. Syst. Man Cybern. Part B 29(3), 433–439 (1999)
- 3. Sanghamitra Bandyopadhyay, Ujjwal Maulik, An evolutionary technique based on K-Means algorithm for optimal clustering in RN,Information Sciences,Volume 146, Issues 1–4,2002

THANK YOU