Finding The Optimal Number ‘K’ In The K-Means Algorithm

Author:
Ramesh Chandra Bagadi
Data Scientist
INSOFE (International School Of Engineering), Hyderabad, India.
rameshcbagadi@uwalumni.com
+91 9440032711

Technical Note

Abstract

In this research Technical Note the author has presented a novel method to find the Optimal Number ‘K’ in the K-Means Algorithm.

Theory

Definition of a Cluster based on Connectivity

We define a Cluster as follows:

A Cluster is a collection of Points (or objects) wherein they are scattered (their property is distributed) in such a fashion that, for a specified distance (measured in appropriate Metric of concern using appropriate Norm of concern) every point of this cluster has at least one neighbouring point also belonging to this cluster located within

(i) this specified distance* [1]

(ii) a certain small neighbourhood of this this specified distance, measured from the aforementioned point of concern.

Proximity Matrix

Given M number of points $\bar{x}_i \in R^N$, $i = 1\ to\ M$, each belonging to R^N, we find the Proximity Matrix P for each (M number of) point with each of all other (M Number of points) points, inclusive of itself. The Proximity can be found using Euclidean distance or using the concept stated in [1].

$$P = \begin{bmatrix} d(1,1) & d(1,2) & d(1,3) & .\ .\ .\ . & d(1,(m-1)) & d(1,m) & \\ d(2,1) & d(2,2) & d(2,3) & .\ .\ .\ . & d(2,(m-1)) & d(2,m) & \\ d(3,1) & d(3,2) & d(3,3) & .\ .\ .\ . & d(3,(m-1)) & d(3,1) & \\ . & . & . & .\ .\ .\ . & . & . & \\ . & . & . & .\ .\ .\ . & . & . & .. \\ . & . & . & .\ .\ .\ . & . & . & \\ . & . & . & .\ .\ .\ . & . & . & \\ d((m-1),1) & d((m-1),2) & d((m-1),3) & .\ .\ .\ . & d((m-1),(m-1)) & d((m-1),m) & \\ d(m,1) & d(m,2) & d(m,3) & .\ .\ .\ . & d(m,(m-1)) & d(m,m) & \end{bmatrix}$$

We now arrange the elements of the Proximity Matrix in Descending Order as a Set S_1 of at most $M_1 = \left(\frac{M^2 - M}{2} \right)$ elements, as the Proximity Matrix is Symmetric and all its diagonal elements are equal to zero.

We now plot this Set S_1 w.r.t to the x-axis of whole numbers. In this plot, there would be at most $M_1 = \left(\frac{M^2 - M}{2} \right)$ number of Levels. At this level, we can have at most $M_1 = \left(\frac{M^2 - M}{2} \right)$ number of Clusters, wherein we can find the points belonging to this Cluster by noting the indices of the operands responsible for the Proximity Differences in the Proximity Matrix. At this juncture, we can segregate almost similar Levels as one Level thus reducing the Number of Levels. The Final Number of levels gotten can be called as the Number ‘K’. Finally, we can find the points belonging to these K Clusters by noting the indices of the operands responsible for the Proximity Differences in the Proximity Matrix.

Higher Order Clusters

For this Set S_1, we again find the Proximity Matrix and similarly repeat the procedure again and find another Set S_2 which has $M_2 = \left(\frac{{M_1}^2 - M_1}{2} \right)$ number of Levels. At this level, we can have at most $M_2 = \left(\frac{{M_1}^2 - M_1}{2} \right)$ number of Clusters, wherein we can find the points belonging to this Cluster by noting the indices of the operands responsible for the Proximity Differences in the Proximity Matrix. At this juncture, we can segregate almost similar Levels as one Level thus reducing the Number of Levels. The Final Number of levels gotten can be called as the Number ‘ K_1’. Finally, we can find the points belonging to these K_1 Clusters by noting the indices of the operands responsible for the Proximity Differences in the Proximity Matrix.

We keep repeating this procedure again and again until $M_L = \left(\frac{{M_{L-1}}^2 - M_{L-1}}{2} \right)$ where L is such that $S_L \subset S_{L-1}$ within an error of a small neighbourhood. At this level, we can have at most $M_L = \left(\frac{{M_{L-1}}^2 - M_{L-1}}{2} \right)$ number of Clusters. At this juncture, we can segregate almost similar Levels as one Level thus reducing the Number of Levels. The Final Number of levels gotten can be called as the Number ‘K_L’. Finally, we can find the points belonging to these K_L Clusters by noting the indices of the operands responsible for the Proximity Differences in the Proximity Matrix.

References

1. Bagadi, R. (2017). Using the Appropriate Norm In The K-Nearest Neighbours Analysis. ISSN 1751-3030. *PHILICA.COM Observation number 180.* http://www.philica.com/display_observation.php?observation_id=173
2. http://www.philica.com/advancedsearch.php?author=12897
3. http://www.vixra.org/author/ramesh_chandra_bagadi