IJNRD.ORG ISSN : 2456-4184

INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG

An International Open Access, Peer-reviewed, Refereed Journal

Review on Deciding Optimal Number of Clusters in k-means Clustering

[1]Kushal Jain, [2]Mohit Kumar

[1]Student, [2]Student
[1]Department of Software Engineering,
[1]Delhi Technological University, Delhi, India

Abstract : In machine learning, the clustering approach is used to classify related data points or objects of a sizable dataset and is frequently applied in various fields, including image analysis, bioinformatics, and research fields like economics and biology. It is crucial to understand the initial parameters before clustering since patterns from any clustering algorithm depend on those parameters. Although there are various clustering techniques, we will concentrate on k-means grouping in this study which serves as the quickest and easiest. We focus on approaches for figuring out how many clusters are needed as clustering inputs. We can request in advance from end users that they submit the number of groups of the data point, but this is not practical because the end user needs to have knowledge regarding each dataset. There are numerous tackles that may be utilized for this, but we only concentrate on a few recent proposed and some widely used techniques.

***Keywords* - Number of clusters, k means, elbow method, silhouette, gap statistics, non parametric, curvature based method.**

1.INTRODUCTION

Clustering is a machine learning technique used to identify and group similar data points together from a large datasets or we can say grouping the unlabelled example based on similarity or proximity to each other is called clustering. Most clustering algorithms are built. That's the point they split the dataset onto a certain no of categories. However,the most difficult part of the process involves performing know the no of groups prior to clustering.Now, In order solve that one possible way is that we can ask the user to enter the no of groups prior to grouping, but that is challenging as consumer needs regarding have expert domain knowledge of the datasets. A no of accuracy indices & techniques has also been proposed in an effort to decide the appropriate cluster size automatically. There are multiple possibilities for clustering. K-means clustering is the method that is employed the most prevalently[1] among them due to its speed and simplicity. The only goal of this study is to count the no of groups in a k-means clustering. Finding total no of clusters has been approached in the variety of ways in the literature. In this article we concentrate on modern methods and some widely used old methods and see in what way new techniques are performing superior than old techniques and on which type of datasets. In this paper we focus on mainly ten approaches which includes elbow method, silhouette method, calinski-harabasz method, gap statistic method, slope method, Curvature based method, cross-validation method, minimum inter distance method, quantitative discriminant method of elbow point, cluster validity evaluation method. And also we see what are limitations and advantages of these approaches.

The article is structured as follows: way. Under Section 2 K-means is defined.Section 3 has detailed review of some methods published in some literature that are used for finding the right K in K-Means. Section 4 concludes the paper.

2. K-Means Clustering

Clustering of data is done using the uncontrolled machine learning algorithm K-means clustering. So in basic definitions, it is grouping of objects according to how similar and unlike they are to one another. The approach is used dividing the information towards the K clusters, where ‘K’ is a predetermined number chosen by the users.

Here is how the K means optimizer operates:

1. Selecting the value that belongs to K (the no of groups) and initializing ‘K’ cluster centroids at random.
2. Create a 'K' cluster by associating every point of data with the closest central.

3. Re calculate each grouping the centroids depending on the standard deviation number of the information points that make up that cluster.
4. Till convergence has been achieved, repeat steps 2 and 3. When centroids are not moving apart, convergence moves and moves very little.

K-means clustering has several advantages:

1. It is a simple and easy-to-understand algorithm that is widely used.
2. It is computationally efficient and can handle large datasets.
3. It works well when the data is well-separated and the clusters are roughly spherical.

Regards Clustering using K-means, however, also has several drawbacks:

1. The algorithm assumes that clusters are roughly spherical and have the same variance, they may not be true in all cases.
2. The algorithm is delicate till the beginning choice among the centers, which may produce various clustering results for different initializations.
3. It may not work well with datasets that have outlier's or noise.
4. The quantity of clusters 'K' should be known beforehand, which might not always be true in practice scenarios. In this paper we try to see the methods available to solve this issue.

Overall, 'K'-means clustering is the useful and commonly employed clustering method data, though it might not be suitable for all types of datasets and clustering tasks.

3. Different Approaches for finding a number of clusters

The methodology section outline the plan and method that how the study is conducted. This includes Universe of the study, sample of the study,Data and Sources of Data, study's variables and analytical framework. The detailsare as follows;

3.1 Elbow Method

The elbow method[2]is an approach used in data science as a tool to ascertain the ideal quantity of groups in a dataset for a clustering algorithm. This elbow approach is based on the notion indicating that when group numbers rise, the within-cluster distances multiplied by 2 (WSS) tends to decrease, but the rate of decrease typically reduces down the quantity of groups. The optimum quantity of groups is often approved as the "elbow point" on a plot of WSS versus how many groups there are.

The formula for WSS is:

$$WSS(K) = \Sigma(pi-ci)^2$$

Where pi is The area separating a piece of information from its assigned group center's, and ci is centroid of the cluster.

The limitation of the elbow method is that sometimes it can be subjective and difficult to use the graph to calculate the precise K number where the graph is pretty smooth to identify elbow point. Additionally, it is based on the supposition that the groups are circular & equal in dimensions, this might not be true in all datasets.

3.2 Silhouette method

Rousseeuw, P. J.[3], in 1987 proposed a method named as silhouette to identify the ideal number of groups for an information set or to assess the efficacy of grouping methods. The Silhouette method measures the effectiveness of grouping by assessing how closely every statistic resembles a particular group in relation to different groups. The method provides An evaluation assigned to each data item, which indicates how well The form of the silhouette wins. The number of clusters is often identified as The individual in question who maximizes the average rating, silhouettes.

The silhouette given the information the point, coefficients I's definition translates as the subsequent:

$$[s(i) = (b(i) - a(i)) / \max(a(i), b(i))]$$

where mean separation among position 'i' along with every other point within its group is given by a(i) and the mean separation amongst position 'i' and any additional locations of the cluster closest to it that are not occupied by it is given by b(i).

One limitation of the silhouette method is that it may not work well with datasets that contain many outliers or noise points, as these can distort the cluster shapes and lead to poor silhouette scores. Additionally, the method is computationally intensive for large datasets and high-dimensional data.

3.3 Calinski-Harabasz method

A dendrite approach for the cluster analysis which is also used to find the ideal number of clusters is proposed by Calinski, T., and Harabasz, J.[4] in (1974) . The optimum number of clusters is that which maximizes the difference in deviation across clusters as well as between clusters which can be measured by this method.

Calinski-Harabasz index for each value of k, which is given by the formula:

$$[CH(k) = (B(k)/(k-1)) / (W(k)/(n-k))]$$

where B(k) is for the between-grouping variance, W(k) is within-cluster variance, n is total amount of data points, and 'k' is the no of the groups present.

The k where Calinski-Harabasz index maximizes in a plot between Calinski-Harabasz index and number of clusters is taken as the optimal number of clusters.

Limitations of CH method include:

- It assumes that the clusters are roughly spherical and equally sized,that may be not available in real world data.
- It can be sensitive to outliers and noise in the data.
- It may not perform well on high-dimensional data, where the distance between points becomes less meaningful.

3.4 Gap Statistics method

Another method for figuring out the number of clusters for clustering algorithms was hypothesized in 2001 by Tibshirani et al[5]. With this technique, the within-cluster dispersion is compare to a reference distribution produced by a random dataset for various values of k. The "k" at which gap statistic maximizes, either the difference between the reference distribution and data's intra-cluster dispersion, is the ideal no of groups.

It is formula for gap analysis is given by:

$$gap(k) = [log(W\_r)] - [log(W\_k) + (1/n)] * [sum(log(W\_r) - log(W\_k))]$$

Where an allusion's inside-of-cluster distribution dataset is given by W_r, the within-cluster dispersion of the original dataset is given by W_k, The amount the details is n.

Authors conclude that the gap analysis method was found to be more efficient than some of the formerly stated methods like the Calinski method etc. So we can say that the gap statistics method as discussed above outperformed the several other methods available during that time. The limitations of the gap statistics method include its sensitivity to the choice of reference distribution and the assumption of spherical clusters with equal variances. It also requires a relatively large dataset to produce reliable results.

3.5 A Non-Parametric Method(slope statistic method)

The slope statistic method[6] works using the output of a clustering algorithm. This finds the maximum group number so that a dataset can be broken down . Also it has been shown that this method performs better(for some datasets) than some popular method that has been proposed in various literature.

In this approach ideal number of clusters k^* is a characterized through two factors: a high silhouette value when k^* is the number of clusters and a substantial decline in intrinsic worth of the silhouette when the number of groups is larger than k^*. This leads the following estimator $k^{\wedge}$for k^*.

$$k = arg\ max\ -\ [s(k+1) - s(k)s(k)^p]$$

When slope method was compared to BIC[7] which is parametric, it is pointed out that the slope method was equivalent or better than BIC even in situations where the data points were generated by multi-Gaussian distributions

3.6 Curvature-based method

Another approach for calculating the number of clusters in k means clustering that utilizes squares to represent inside the group (WSS) graph's curvature in relation to the number of clusters is proposed in 2017 by Zhang et al[8] . When the WSS curve is plotted it can be seen that the elbow joint of the curve is corresponding to the maximum curvature point of the second derivative curve and this method works on this intuition. The point where the WSS curve starts to flatten out consists of the ideal number of groups as this indicates that further The total amount of has increased clusters would not significantly reduce the within-cluster sum of the squares.

This limitation of this method is that it may not work well on datasets whose WSS curve formed have no clear elbow formed or WSS curve is irregular

The Iris dataset, the Wine dataset, the Seeds dataset, and the Glass dataset were among the synthetic and real-world datasets that the authors used to test their methodology.The experimental findings demonstrated that the suggest technique performed better than the several other methods, including the elbow method,gap statistic approach[5], the silhouette approach[3], and BIC[7] in terms of accuracy and noise resilience.

3.7 Minimum inter-center distance Method

When the number of clusters is increased in k-Means, the minimum pairwise distance between their centers decreases. It has been found that the final significant decline of this strategy occurs just prior to the user-defined number exceeding the number of naturally occurring clusters. Last Major Leap (LML) and the Last Leap (LL) are two strategies[9] that were created to find the number of clusters for k-Means.

3.7.1. The Last Leap (LL): This is for identifying the last noteworthy difference occurs in the values of d_k

$$LL(k) = \frac{d_k - d_{k+1}}{d_k}$$

The estimated $\hat{k}$(optimal K) is the k at which maximum of LL is there as follows:

$$\hat{k}_{LL} = arg\ arg\ LL(k)$$
$$k = 2, \ldots., k_{max}$$

A big value of the LL numerator, $d_k - d_{k+1}$, helps to identify areas where a significant decrease has taken place. Determining the location of the 'last' significant decline in LL is made easier because to the denominator's presence of dk, which typically decreases as k increases.

3.7.2. The Last Major Leap (LML): LML is specifically defined for checking the last noteworthy reduction in d_k:

$$I_{LML}(k) = \{1 \ \ , if \ \frac{1}{2}dk > \ \ max \ \ d1, 0 \ \ , otherwise,$$

$$K = \{|I_{LML}(k) == 1|\}, \qquad and$$

$$\hat{k}_{LML} = \{max\ K \ \ , if\ K \neq \emptyset. 1, \qquad otherwise$$

For larger l(l = k + 1, ...,k_{max}) values , $I_{LML}(k)$ function identifies a substantial leap when half of the present value of d_k is larger than value of d_l.. K represents the collection of cluster numbers (k) where there is significant leap in d_k occurs. The greatest number is chosen as the predicted k and is regarded as the most recent major jump from the set K.

Regarding the detection of single clusters: I_{LML} may fail to detect any notable leaps, resulting in an empty set K . In this case, we may assume that the data lacks any discernible cluster structures, hence , $\hat{k}_{LML}$ is set to 1. LL also has similar rule incorporated:

$$\hat{k}_{LL} = 1, if \frac{1}{2}d_k \ < max\ d_l$$

Authors observed that the maximum accuracy is attained by LL, LML,CE, FHV, BIC, MPC, CH, I, and PBMF, closely followed by Xu, Slope, PC, andDunn. The fact that LML and LL have extremely low O(k^2) computation complexity is a further benefit. When compared to the other good-performing cluster number estimation methods, the LML and LL execution time is consistently very low as data set size increases.

3.8 A cluster validity evaluation method

Liang et al[10] in 2019 The Relation of Departure of Sum-of-Squares with Euclidean separation ‘RDSED’ as an original cluster integrity the index, including an approach is developed a sequence to dynamically dChoose an appropriate proportion of groupings on basis of RDSED.

This method computes the RDSED value being big down tiny throughout the clustering number & area used for automatically completing the grouping validation procedure, using this indices number, producing the results of the clustering partition and a near to best amount of groups. A suggested strategy can efficiently evaluate clustering partition results and can typically obtain a cluster number which is close to the actual cluster number, which is demonstrated by experiments using both synthetic and Information from the real world. This is in juxtapose to some classical methods for evaluating the validity of the clustering.

RDSED: 'RDSED' stands for Ratio of Deviation of Sum-of-Squares and Euclid Distance. If the dataset that needs to be clustering comprises d-dimensional information, given its m n information framework, the dispersion from the total-of-squares, with its Euler relationship the disorder is defined by:

$$DSED(m) = |(SSW/SSB)\, SST - SID/ADB - (n - m)|$$

wherein m represents the total amount of cluster and n is the count of information elements in the information set. SSW[14] is Sum-of-Squares of Within-cluster, SSB[14] is sum-of-Squares of Between-cluster, SST[14] is Total Sum-of- Squares, SID Sum of Intra-cluster Distance and ADB is Average Distance Between clusters. The ratio of the difference between the DSED of adjacent two cluster numbers & the larger DSED value is defined as RDSED.

$$RDSED(m) = \frac{\Delta ADSED(m)}{[max\big(DSED(m)],[DSED(m+1)\big)]} = \frac{[DSED(m) - dsed(m+1)]}{[max(DSED(m)],[dsed(m+1))]}$$

A close-optimal amount of groupings must be identified. This defines the spectrum of the close-optimal amount of groups for picking. The smallest number of clusters is determined herein using the experimental results suggested through Dunn (1973)[11], Bouldin & Davies (1979)[12], and Krzanowski & Lai (1985)[13]. m_{min}= 2 and the maximum number of clustering is set as m_{max} =$\sqrt{n}$.

The 'RDSED' Within the group's integer assortment, values is determined as big to small. That 'DSED' or value for the largest quantity of groups comes first ($\sqrt{n}$) is determined, followed for the pre order Quantity of 'DSED' of $\sqrt{n}$ is determined according to ($\sqrt{n}+1$) therefore its corresponding The largest amount of groups' RDSED is found. The index's values for every cluster code are then determined individually. While RDSED falls within the range that is appropriate, the computation of the numerical value of subsequent group code begins; when RDSED fails to fall inside the appropriate range, The sheer amount of groups as it is is incorrect.

RDSED typically approaches 1 if [DSED](m) & [DSED (m+1)] are quite distinguished from one another.

RDSED performs relatively worse while handling information involving a tiny-granularity. However, RDSED can typically identify the nearly ideal quantity of groupings while assessing the grouping outcomes. division. Additionally, RDSED has a bias towards algorithms with low stability. This bias makes it more likely for RDSED To determine the actual amount of groups when the algorithm is rather unstable, but it also limits how stable RDSED may be.

3.9 A quantitative discriminant method of elbow point

As we know, the elbow method is one of the oldest and most popular techniques for finding the number of clusters but it is sort of a manual method as we need to identify the elbow points in a graph by visualizing it. Thus, when the curve is smooth it is difficult for experienced analysts to find the elbow point. In order to get a solution to this, Congming et al[15] in 2021 proposed an innovative elbow point discriminant approach that produces a mathematical measure which calculates the best group size for a spreadsheet's grouping.

This Forearm method, K-means++, a tool called Min as normalization, and cosine of forearm engagement degree as criterion are the foundation of the approach.

This starts by finding The collection of data X containing N locations' total deformation (MD), provided by

$$MD_i = \frac{SSE}{N} = \frac{\sum_{K=1}^{K} \quad \sum_{x_i=C_k} \quad x_i - \mu^2{}_{k_2}}{N}$$

SSE is sum of the squared Euclidean distances and MDi denotes the average distortion of the N data points in the dataset X, with i being the assigned cluster number for X.
After this change each MD_i using the Min Max Scaler and scale each changed value $N_{(md)}$ to a specific range [0-10], which is an empirical value.The absolute normalized SSE data space, N = $[n_1, n_{2,\dots\dots} n_k]$, where n_i has the definition:

$$ni = \frac{M_{Di} - MD_{(min)}}{MD_{(max)} - MD_{(min)}} * 10$$

Then using this finds the the angle $\angle\alpha_j$

$$\alpha_j = arccos\frac{E_{ij}^2 + E_{jk}^2 - E_{ik}^2}{2E_{ij}E_{jk}}$$

Where Eij represents Euclidean distance between data points i and j calculated using ni values

Over the course of α ∈ $[\alpha_1, \alpha_2, ..., \alpha_{k-2}]$, this uses the indicator for least as, K_{opt} that is treated as that projected maximum groups size for that analyzed information set, and that tiny, showing the perfect bend point relating the projected prospective ideal clusters amount without low certainty.

IV. CONCLUSION

Among most used clustering algorithms K-means is one of them and this is why it is important to study its properties. It is used not only in this field of information mining, categorization, & machine learning communities but also its use increased by practitioners of bioinformatics, engineering, marketing research, and other application domains.

The proper number of groups is one of many contentious grouping concerns, which some may consider as being unfounded since, in many instances, "clusters are not in data but in the viewing eye." This study discusses this topic. In this study, we investigate the situation in which data contain clusters, even if they are not perfectly traditional.

We have seen some not commonly used and newly proposed techniques which perform better than techniques used like silhouette method, calinski-harabasz method , gap statistics method etc. And in future we could compare these new methods and see what is best method among all these.

References

[1] Jain, A. K., & Dubes, R. C. (1988). *Algorithms for clustering data*. Prentice-Hall, Inc..
[2] Yuan, C., & Yang, H. (2019). Research on the K-value selection method of K-means clustering algorithm. *J*, *2*(2), 226-235.
[3] Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, 53-65.
[4] T. Caliński & J Harabasz (1974) A dendrite method for cluster analysis, Communications in Statistics, 3:1, 1-27, DOI: 10.1080/03610927408827101
[5] Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63(2), 411-423.
[6] Fujita, A., Takahashi, D. Y., & Patriota, A. G. (2014). A non-parametric method to estimate the number of clusters. Computational Statistics & Data Analysis, 73, 27-39.
[7] Schwarz, G. (1978). Estimating the dimension of a model. The annals of statistics, 461-464.
[8] Zhang, Y., Mańdziuk, J., Quek, C. H., & Goh, B. W. (2017). Curvature-based method for determining the number of clusters. Information Sciences, 415, 414-428.
[9] Gupta, A., Datta, S., & Das, S. (2018). Fast automatic estimation of the number of clusters from the minimum inter-center distance for k-means clustering. Pattern Recognition Letters, 116, 72-79.
[10] Li, X., Liang, W., Zhang, X., Qing, S., & Chang, P. C. (2020). A cluster validity evaluation method for dynamically determining the near-optimal number of clusters. Soft Computing, 24, 9227-9241.
[11] Dunn, J. C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters.
[12] Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. IEEE transactions on pattern analysis and machine intelligence, (2), 224-227.
[13] Krzanowski, W. J., & Lai, Y. T. (1988). A criterion for determining the number of groups in a data set using sum-of-squares clustering. Biometrics, 23-34.
[14] Zhao, Q., Xu, M., & Fränti, P. (2009). Sum-of-squares based cluster validity index and significance analysis. In Adaptive and Natural Computing Algorithms: 9th International Conference, ICANNGA 2009, Kuopio, Finland, April 23-25, 2009, Revised Selected Papers 9 (pp. 313-322). Springer Berlin Heidelberg.

[15] Shi, C., Wei, B., Wei, S., Wang, W., Liu, H., & Liu, J. (2021). A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. EURASIP Journal on Wireless Communications and Networking, 2021(1), 1-16.