# Are the clusters good?

Dr. Saptarsi Goswami Jul 27, 2020 · 7 min read

Understanding how to evaluate clusters

**Clustering** is defined as finding natural groups in the data. But this definition is inherently subjective.

**What are natural groups?**

If we see the below picture, can we figure out the natural group of the flowers? Is it by the shape or is it by the color? It may even be by the size or species of the flower. Hence, *the notion of a natural group changes based on what characteristics we are focussing on.*

Fig 1: Flowers (Source: Unsplash)

Let's take another example, where we have some points or observations in a 2D plane, i.e. we have two attributes only

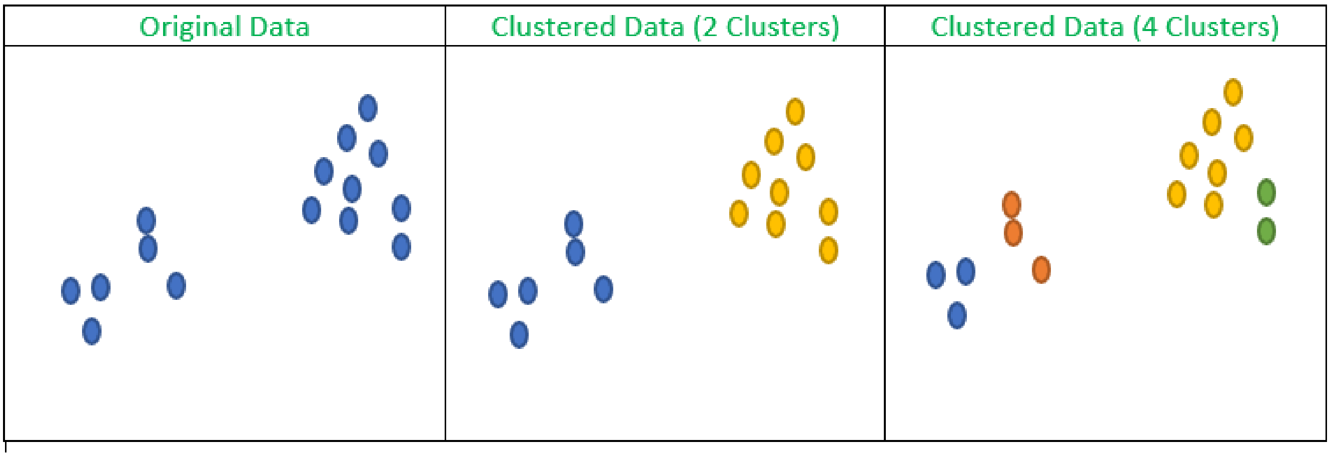


Fig 2: Original Data and clustering with different number of clusters (Image Source: Author)

If we look at the above figure which has three subfigures. The first subfigure has the original data, the second and third subfigure shows clustering with the number of clusters as two and four respectively (Observations belonging to the same cluster are marked with the same color).

Fortunately, we can still visualize and try to gauge the quality of the clusters, however, if we go for more numbers features, we can't visualize and see. Hence, there needs to be a mechanism, some measure which can make us compare two or more sets of clusters, or maybe two or more clustering algorithms on the same set of data. *Unfortunately, like the way we can compare classification algorithms using accuracy or in case of regression using mean squared error, it's not so clear cut for clustering.*

### Clustering Tendency:

What if the data do not have any clustering tendency, even if the data is random and we apply k-means, the algorithm will generate k-clusters. Hence, how do we measure, if the data has a clustering tendency or not? To measure the same we take the help of Hopkins Statistic.

### Hopkins Statistic (H)

In this scheme, as many artificially generated random points are added as there are original data points in the dataset. For each of the original points, the distance with it's nearest neighbor is calculated, denoted by **w** and the same exercise is repeated for the artificially generated points. Here, distance with the nearest neighbor is calculated as **u.**

$$H = \frac{\sum_{i=1}^{p} w_i}{\sum_{i=1}^{p} u_i + \sum_{i=1}^{p} w_i}$$

A value near 0.5 indicates the data do not have clustering tendencies as both of w and p are equal.

### Cluster Evaluation Measures:

### Sum of Squared Error (SSE):-

The most used clustering evaluation tool is the sum of squared error which is given by the below equations.

$$\mathbf{c}_i = \frac{1}{m_i} \sum_{\mathbf{x} \in C_i} \mathbf{x}$$

$$K$$

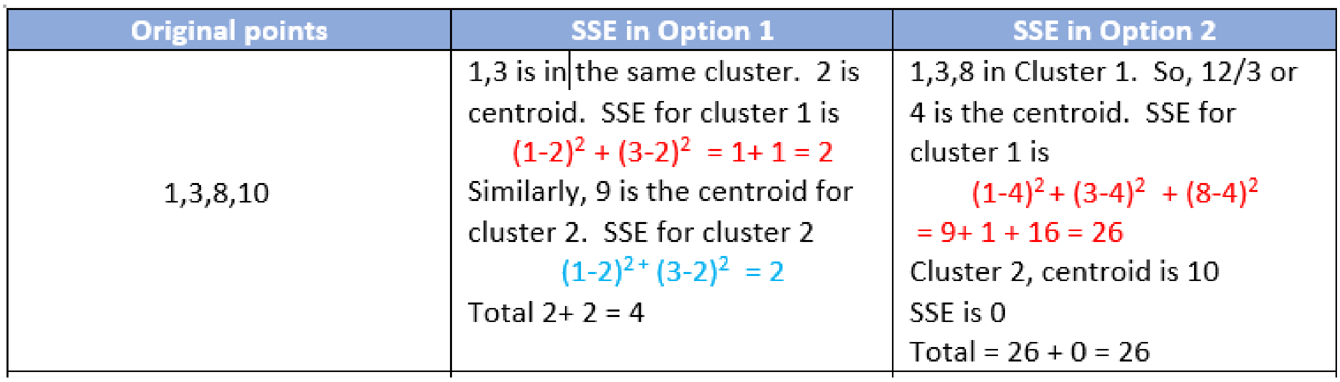
$$\mathrm{SSE} = \sum_{i=1} \sum_{x \in C_i} dist(c_i, x)^2$$

SSE Equations (Image Source: Authors)

Basically, at the first step, we find the centroid of each cluster by taking an average of all the observations in that cluster.

- Then we find how much the points in that clusters deviate from the center and sum it.
- Then we sum this deviation or error of individual clusters.
- SSE should be as low as possible.

I always understand the intuitions better with an example, let’s just do that

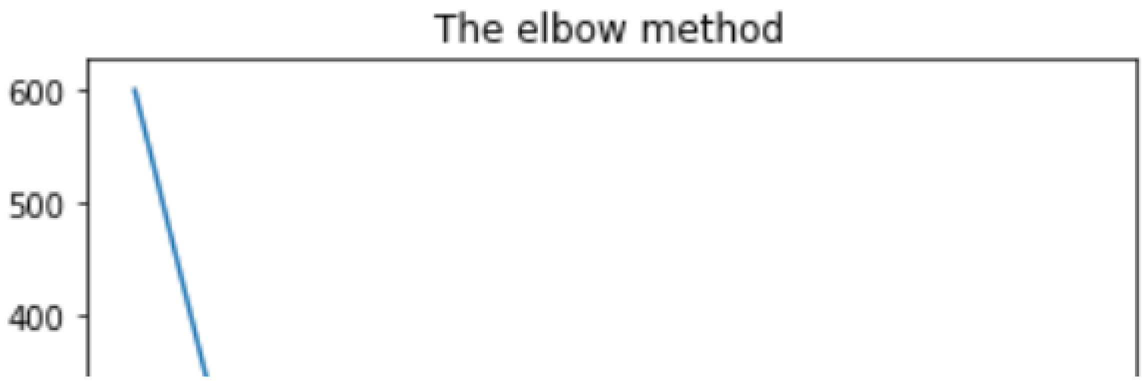


| Original points | SSE in Option 1 | SSE in Option 2 |
|---|---|---|
| 1,3,8,10 | 1,3 is in the same cluster. 2 is centroid. SSE for cluster 1 is<br>$(1-2)^2 + (3-2)^2 = 1+ 1 = 2$<br>Similarly, 9 is the centroid for cluster 2. SSE for cluster 2<br>$(1-2)^{2+} (3-2)^2 = 2$<br>Total 2+ 2 = 4 | 1,3,8 in Cluster 1. So, 12/3 or 4 is the centroid. SSE for cluster 1 is<br>$(1-4)^2 + (3-4)^2 + (8-4)^2$<br>$= 9+ 1 + 16 = 26$<br>Cluster 2, centroid is 10<br>SSE is 0<br>Total = 26 + 0 = 26 |

Fig 3: Illustration of SSE. (Image Source: Author)

I think the above example is self-explanatory. (1,3) and (8,10) are natural cluster organization with the number of clusters as 2. SSE is 4. On the other setting where we have kept 1,3,8 in the same cluster and 10 in the other SSE has shot up to 26.

SSE is widely used to find, the number of clusters (k) especially for K-means.



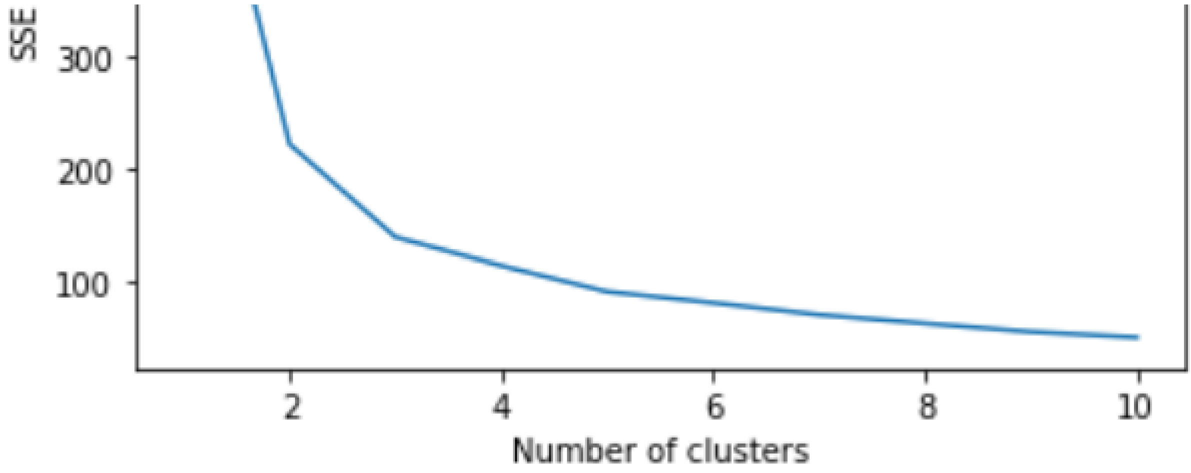


Fig 4: SSE for different number of clusters on the iris dataset (Image Source: Author)

Here also, we are comparing cluster qualities for different options of cluster numbers. This is a monotonically decreasing function, if we continue increasing the number of clusters, it will keep on reducing. Hence, the optimal number of clusters is taken when drop-in SSE stabilizes or forms an elbow. SSE is a good measure if we are trying to find spherically shaped clusters.

Let's take a more generalized view, what we want in a well-formed set of clusters are as follows

- the observations within a cluster to be as close as possible. This is referred to as **Cohesion**
- the observations from two clusters should be far from each other. This is referred to as **Separation**

illustrated using the below Figure

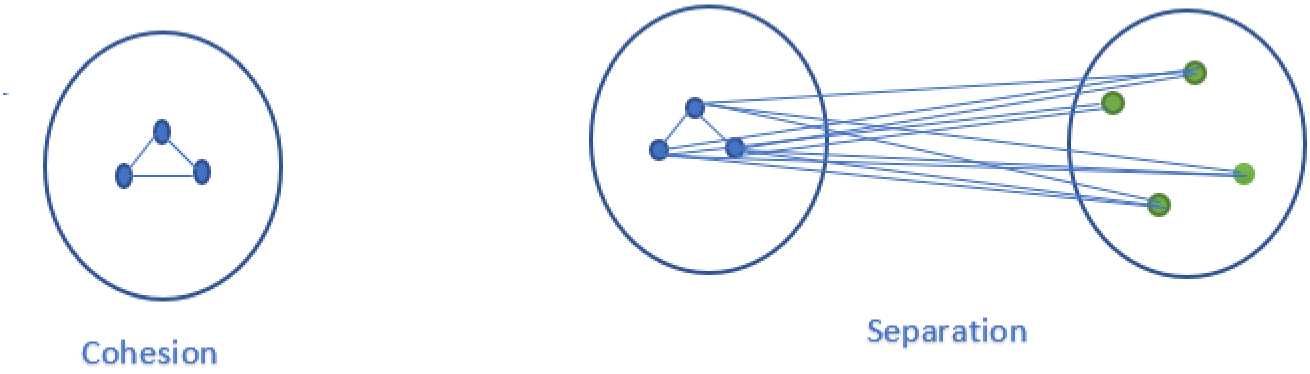


Fig 5: Cohesion and Separation ( Image Source: Author)

The evaluation measures can be categorized into two ways:-

**Internal Measure:**

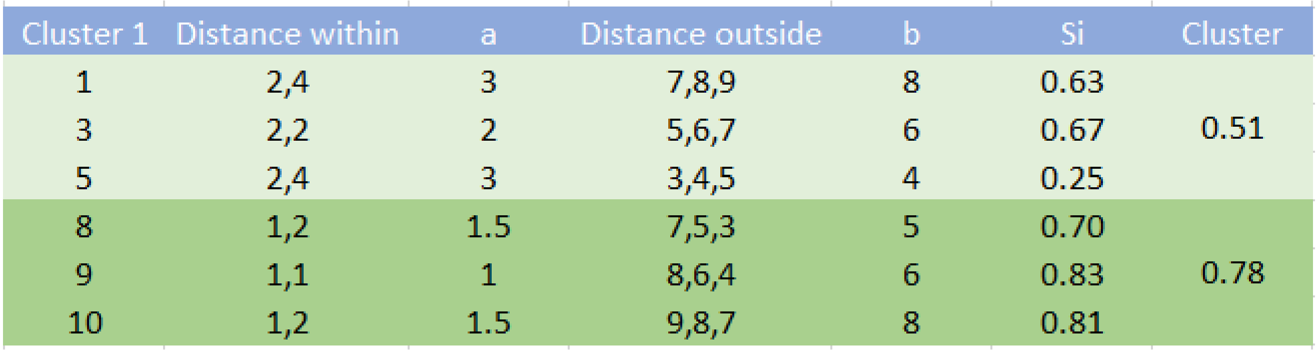
This is the more general one when the class label is not available. The **silhouette coefficient** is one such popular measure. This uses the concept of both cohesion and separation. Let's take a point i from cluster j, first, we calculate the distance of the point i from all the points in j and find the average let's call this ai (Cohesion). Now, let's take another cluster k, similarly, we find the average distance of the point i, from all the points in the cluster k, let's call this as b(Separation). Ther can be many such clusters, the minimum of them is taken as bi.

Then Silhouette coefficient for the point i is

$$s_i = (b_i - a_i) / \max(a_i, b_i).$$

Silhouette Coefficient ( Image Source: Author)

For a cluster, the Silhouette coefficient needs to be calculated for all the points in a cluster and an average is taken. The value lies between -1 and 1, higher the value better is the cluster. The below figure encloses, a worked-out example



| Cluster 1 | Distance within | a | Distance outside | b | Si | Cluster |
|---|---|---|---|---|---|---|
| 1 | 2,4 | 3 | 7,8,9 | 8 | 0.63 | 0.51 |
| 3 | 2,2 | 2 | 5,6,7 | 6 | 0.67 | |
| 5 | 2,4 | 3 | 3,4,5 | 4 | 0.25 | |
| 8 | 1,2 | 1.5 | 7,5,3 | 5 | 0.70 | 0.78 |
| 9 | 1,1 | 1 | 8,6,4 | 6 | 0.83 | |
| 10 | 1,2 | 1.5 | 9,8,7 | 8 | 0.81 | |

Fig 6: Silhouette coefficient ( Image Source: Author)

1,3,5,8,9,10 are taken as the initial set of points. 1,3,5 is assumed to be the first cluster and 8,9,10 in the second cluster.

So for point 1, a1 is ( 2+ 4)/2 = 3 and b is (7+8+9)/3 =7 and then we use the formula. We can observe

- The second cluster has a better Silhouette coefficient as it is more compact

- The center points have a better Silhouette coefficient compared to others
- The boundary points have the lowest Silhouette coefficient

Some other well used measures are **Dunn Index and DB Index**.

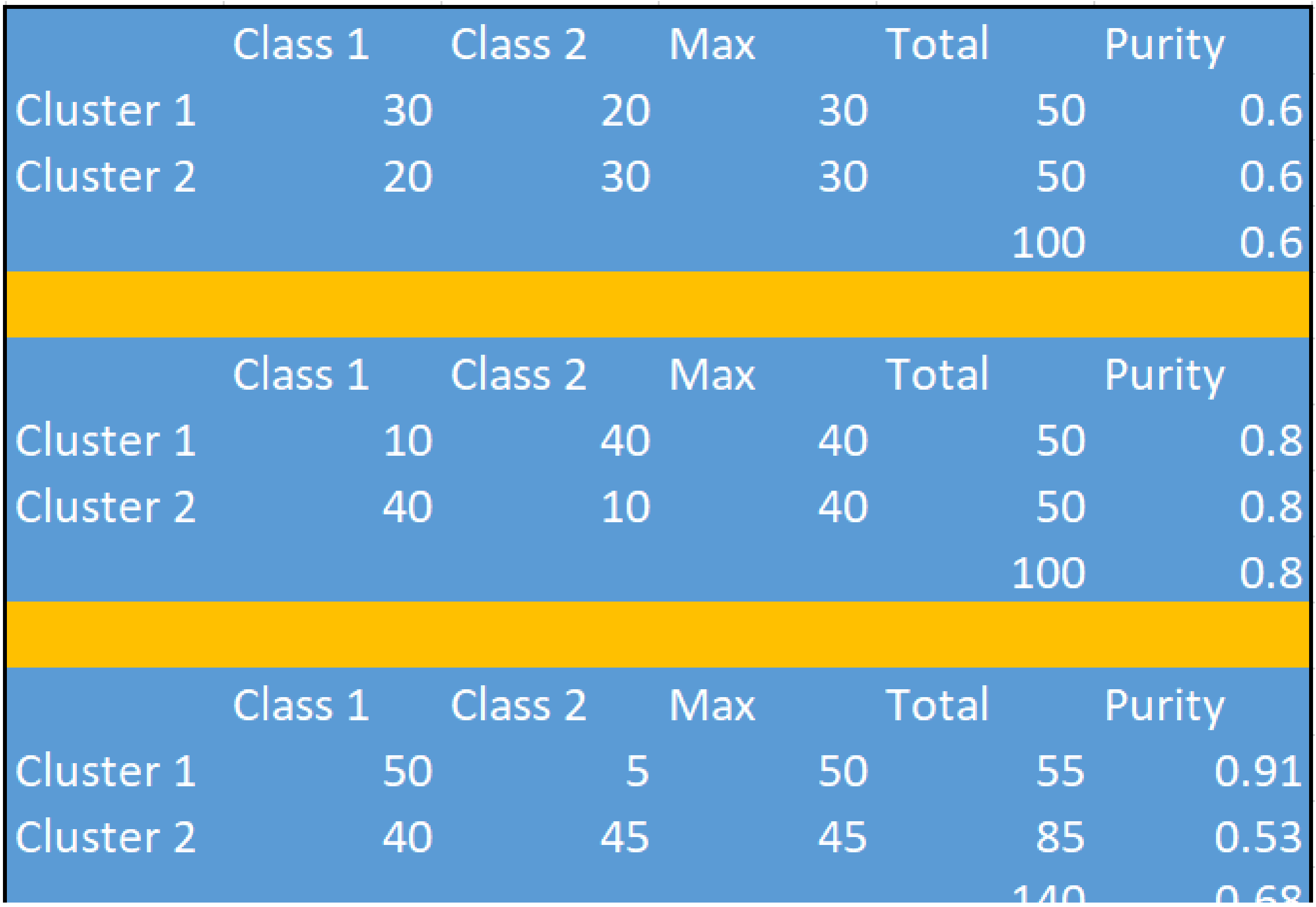
**External Measure:**

We have a class label available, and we use this class label to evaluate the clustering results. Ideally, each one of the classes should form a cluster. One such measure is called purity, as given by the following formula.

$$Purity = \sum_{i=1}^{K} \frac{m_i}{m} * p_i$$

Equation of Purity (Image Source: Author)

K is the number of clusters, mi is the total number of observations in the cluster and m is the total number of observations. Pi is the proportion of the majority class in that cluster. As an example, if cluster i has 5 observations from class 1 and 20 from class 2. Then class 2 is the majority class and the purity is 20/25 or 0.8. This is further illustrated with the below example



| | Class 1 | Class 2 | Max | Total | Purity |
|---|---|---|---|---|---|
| Cluster 1 | 30 | 20 | 30 | 50 | 0.6 |
| Cluster 2 | 20 | 30 | 30 | 50 | 0.6 |
| | | | | 100 | 0.6 |
| | | | | | |
| | Class 1 | Class 2 | Max | Total | Purity |
| Cluster 1 | 10 | 40 | 40 | 50 | 0.8 |
| Cluster 2 | 40 | 10 | 40 | 50 | 0.8 |
| | | | | 100 | 0.8 |
| | | | | | |
| | Class 1 | Class 2 | Max | Total | Purity |
| Cluster 1 | 50 | 5 | 50 | 55 | 0.91 |
| Cluster 2 | 40 | 45 | 45 | 85 | 0.53 |
| | | | | 140 | 0.68 |

Fig 7: Illustration of Purity ( Source: Author)

In the above diagram, three variants of clustering results are shown The calculations are self-explanatory. For option 1 and option 2, both the clusters are equal-sized. The second option is more homogeneous hence better purity. For option 3, cluster 2 dominates cluster 1.

**Purity has a value 0 and 1, the closer it is to 1, the better is the purity.** There are more measures like Entropy, F Score, etc. for details, I highly recommend you to read the book by Pang, Steinbach, Kumar. Why should we use clustering, when we have labels available, to maybe to come up with a new clustering algorithm, test different configurations, and assumptions. It is to be noted, certainly, the clustering is done on the data after removing the class label and then the label is used to validate the cluster quality.

We do not discuss evaluations for Hierarchical clustering, which is a different ball game.

**Conclusion:**

- Clustering is an inherently complex task and hence the quality of the clustering needs to be evaluated.
- This is useful to compare multiple clustering algorithms, as well as a different result of the same clustering algorithm with different parameter values
- At first, we may test, whether there is a clustering tendency or not
- A cluster quality measure should consider cohesion and separation
- It can be internal and external based on the availability of class labels

**References**:

[1] Tan PN, Steinbach M, Kumar V. Introduction to data mining. Pearson Education India; 2016.

[2] Jain AK. Data clustering: 50 years beyond K-means. Pattern recognition letters. 2010 Jun 1;31(8):651–66.

[3] https://youtu.be/_KKT55JohcU

[4] https://towardsdatascience.com/clustering-evaluation-strategies-98a4006fcfc

Clustering Data Science Machine Learning Cluster Analysis Cluster