VERİ BİLİMİ DERGİSİ
www.dergipark.gov.tr/veri

Comparison of the Methods to Determine Optimal Number of Cluster

Fatih Emre OZTURK[1], Neslihan DEMIREL[2]*

[1]*Dokuz Eylul University, The Graduate School of Natural and Applied Science, Data Science, İzmir*
[2]*Dokuz Eylul University, Faculty of Science, Department of Statistics, İzmir*

Abstract

Clustering is an unsupervised learning that divides observations into groups based on their similarity. The most widely used clustering algorithm is k-means. However, in this clustering algorithm, the number of clusters needs to be determined in advance. In this study, the most widely used methods for determining the number of clusters, namely Average Silhouette, Caliński-Harabasz, Davies-Bouldin and Dunn Index were used. The performances of these methods were compared by Rand Index and Meila's Variation of Information (MVI) criteria on nine real data sets where the number of clusters was known in advance. According to these criterias, Average Silhouette was given more successful results.

Keywords: *Cluster analysis, Average Silhouette, Dunn Index, Davies-Bouldin, Calinski-Harabasz*

Makale Bilgisi

Başvuru:
05/03/2023
Kabul:
15/05/2023

Optimal Küme Sayısının Belirlenmesinde Kullanılan Yöntemlerin Karşılaştırılması

Özet

Kümeleme, gözlemleri benzerliklerine göre gruplarına ayıran bir denetimsiz öğrenme şeklidir. En yaygın olarak kullanılan kümeleme algoritması k-ortalamadır. Ancak bu kümeleme algoritmasında küme sayısının önceden belirlenmesi gerekmektedir. Bu çalışmada en çok kullanılan küme sayısı belirleme yöntemlerinden Ortalama Silüet (Average Silhouette), Caliński-Harabasz, Davies-Bouldin ve Dunn Endeksi kullanılmıştır. Bu yöntemlerin performansları küme sayısı önceden belli olan dokuz gerçek veri seti üzerinde Rand Endeksi ve Meila bilgi kriteri (Meila's Variation of Information-MVI) kriterleri ile karşılaştırılmıştır. Bu kriterlere göre değerlendirildiğinde Ortalama Silüet ile daha başarılı sonuçlar elde edilmiştir.

Anahtar Kelimeler: *Kümeleme analizi, Ortalama Silüet, Dunn Endeksi, Davies-Bouldin, Calinski-Harabasz*

* İletişim e-posta: neslihan.ortabas@deu.edu.tr

1 Introduction

Machine learning is an evolving branch of computer algorithms that are designed to imitate human intelligence. Techniques supported by machine learning are applied with success in numerous fields starting from pattern recognition, computer vision, finance, entertainment, and computational biology to medical specialty and medical applications [1]. It is possible to divide machine learning into three categories. The first one is supervised learning. Supervised learning is just a rationalization of the concept of learning from experiences. In supervised learning, the learner program is given two sets of data, a training and a test. The concept is for the learner to "learn" from a group of labeled examples within the training data so it can establish unlabeled examples in the test data with the best attainable accuracy [2]. The most widely used algorithms of supervised learning are tree-based algorithms for classification and regression. The second category is semi-supervised learning (SSL). It is the conclusion of the idea that labeling the training data in the real-world project is difficult, computationally expensive, or taking too much time, because it needs human expertise with specific domain expertise and training [3]. SSL addresses this issue by permitting the model to integrate half or all of the offered unlabeled data in its supervised learning. The goal is to maximize the training performance of the model through such newly-labeled examples while minimizing the work needed for human expertise [3]. The third category of machine learning is unsupervised learning. In contrast to supervised learning, the data is not labeled in unsupervised learning. Instead, it captures patterns as probability densities [4]. Unsupervised learning is used in many fields such as psychology [5], biology [6], computer security [7] pattern recognition [8], and image processing [9]. The most widely used algorithms in unsupervised learning are dimension reduction and clustering. Clustering is unsupervised learning that separates observations in a dataset into clusters based on their similarity [10]. Similarity is based on distance measurement. Observations that are close to each other are considered similar and observations that are far away are considered dissimilar. Observations in the same cluster are expected to be similar, while observations in different clusters should be dissimilar. The Euclidean distance metric is generally used to measure similarity/distance [11]. In addition, Manhattan and Cosine distance metrics are also used. There are two main criterions for clustering. The former is that the within-cluster variances of the clusters should be as low as possible. This ensures for observations in the same cluster to be more similar. The letter is to separate the clusters from each other as much as possible. This ensures for observations in the different clusters to be dissimilar.

There are many clustering algorithms (e.g. k-means, k-medois, clara, hierarchical, fuzzy, model-based, and density based) [12]. However, it would not be wrong to say that the most widely used method is k-means [13-17]. In k-means, each cluster center is located according to the arithmetic mean of the observations in that cluster. k-means selects a point and considers it as the center of the cluster, then divides the other observations into clusters according to their distance from the mean [18]. However, the number of clusters must be determined in advance [19]. Determining the number of clusters is too complex to be achieved by simply observing the overall structure of the data set.

There are more than thirty methods for determining the optimal number of clusters [20]. While some of these methods are valid only for data sets that meet certain conditions, there are also methods that are valid for all types of data sets. In this paper, we compare four mostly used methods [21-24] (Average Silhouette, Calinski-Harabasz, Davies - Bouldin and Dunn Index) that determine the optimal number of clusters for nine, different types of real datasets. The success of the methods in estimating the number of clusters of the datasets is compared and the Rand Index and Melia VI scores based on whether each observation is placed in the correct clusters is analyzed.

The remainder of this paper consists of three sections. The second section explains the k-means algorithm in detail and then shows how the methods used to determine the optimal number of clusters which are Average Silhouette, Calinski-Harabasz, Davies - Bouldin and Dunn Index work. At the end of the section Rand Index and Melia VI Indexes, which is used to compare the success of the methods, is explained. In the third section, each dataset is explained and the data cleaning performed before the clustering algorithm is explained. Then, the number of clusters suggested by each method is compared with the real number of clusters and the Rand Index and MVI value are

compared. In the discussion section, the results of the studies are presented.

2 Methods

This section provides a theoretical explanation of the k-means clustering algorithm and methods to determine the optimal number of clusters that are used in this study. First, the k-means clustering algorithm for clustering data sets will be explained, followed by the methods used to determine the number of clusters (Average Silhouette, Davies-Bouldin, Calinski-Harabasz, and Dunn Index). Concrete illustrations are included in these explanations. Finally, Rand Index and MVI methods will be explained to test the success of clustering results.

2.1 k-means algorithm

There are various k-means algorithms available. However, the common approach is the Hartigan-Wong algorithm [25], which sums the squared distances between observations and the matching centroid to determine the total within-cluster variation.

Since the data to be clustered is generally not tidy, the data should be prepared before it is clustered. There are some steps to be followed for this preparation process. The first of these steps is to check for missing values. The presence of missing values should be detected and necessary actions should be taken regarding this issue. Secondly, the unit of measurement of each variable will undoubtedly be different. These differences in measurement units will negatively affect distance (similarity) calculations. In order to avoid this negative impact, standardization, which means transforming all observations in the data set with a mean of 0 and a standard deviation of 1, should be performed [26]. Another preparatory step is to look at the correlation matrix of the variables in the data set. If there is a high correlation between combinations of pairs of variables, this can both negatively affect the analysis and prevent the clear distinction of observation clusters from being recognized. To avoid these situations, principal component analysis should be applied to the data set. Since principal component analysis (PCA) retains only the variables with highest variance, it is likely for clusters to be more visible [27]. Other advantages of the PCA are both to avoid the curse of dimensionality, and to avoid computing the k-means becomes computationally expensive [28].

For a clearer understanding of the k-means clustering algorithm, it is useful to explain it step by step. After the preparation phase, the first step of the k-means algorithm is to determine the number of clusters. As mentioned before, this step requires detailed analysis. In the following sections, the methods used to determine the number of clusters are explained in detail. In the second step of the k-means clustering algorithm, cluster centers (centroids) are randomly selected as many as the number of clusters selected in the previous step. The average distance (usually Euclidean Distance is used) of the other observations in the dataset to the centroid is calculated. According to these calculations, observations are assigned to clusters according to their distance to each centroid. The new centroids of the new clusters formed after these assignments are calculated by taking the mean of each cluster. The other parts after the second stage continue until convergence and the number of iterations is reached [29]. Convergence is the condition where the total within sum of squares within has the minimum value [30]. *Equation 1* shows how total within sum of squares which intended to be minimized is calculated:

$$\text{tot.withinss} = \sum_{k=1}^{K} \sum_{x_i \in c_k} (x_i - \mu_k)^2 \qquad (1)$$

where,

k is the cluster number,
x_i is the observation,
c_k is the cluster that observation x_i is assigned to,
μ_k is the mean of c_k.

2.2 Average silhouette method

Average silhouette is a cluster interpretation and validation method based on the comparison of cluster tightness and separation [31]. Cluster tightness assesses how accurate the cluster assignment of the observation is, while cluster separation assesses how well the observation separates from the cluster to which it was not assigned. In order to construct silhouettes, a clustered data set and calculation of the distance between observations are required.

2.2.1 Cluster tightness

For each observation in the data set, cluster tightness is calculated separately. For each observation, the average Euclidean distance

between x_i and $x_{i'}$ nd is calculated. Assume dark colored dot is x_i and for a concrete illustration please see *Figure 1*. *Equation 2* explains how cluster tightness is calculated.

$$a_i = \frac{1}{n_k - 1} \sum_{x_i, x_{i'} \in c_k} d(x_i, x_{i'}) \quad (2)$$

where,
c_k is the cluster that observation x_i assigned to,
n_k is the total observation number of c_k ,
$x_{i'}$ are all observations other than x_i in c_k ,
$d(x_i, x_{i'})$ is Euclidean distance between x_i and $x_{i'}$

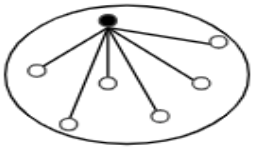

Figure 1. illustration of a_i calculation

2.2.2 Cluster separation

For each observation in the data set, cluster separation is calculated separately. For the observation x_i , cluster separation is calculated as the minimum Euclidean distance between x_i and all the observations in the clusters that are not in the cluster that $x_i$is assigned to[31]. Assume dark colored dot is x_i and for a concrete illustration please see *Figure 2*. *Equation 3* explains how cluster separation is calculated.

$$b_i = min(\, d(x_i, x_{i'}) \,) \quad (3)$$

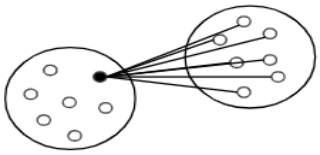

Figure 2. illustration of b_i calculation for two clustered data

After both cluster tightness and separation are calculated, average silhouette value is calculated by using the Equation 4:

$$S_i = \frac{b_i - a_i}{max(a_i, b_i)} \quad (4)$$

As mentioned before, the silhouette value is calculated separately for each observation. This means that the validity of clustering is measured for each observation. The sum of the silhouette values of each observation divided by the number of observations is sufficient to measure the validity of the clustering result. The value obtained as a result of this process is called the average silhouette value. The average silhouette takes values between -1 and 1, which means poor clustered and well clustered, respectively.

2.3 Caliński - Harabasz Method

Essentially, Caliński-Harabasz method is a method to minimize the variation of each cluster [32]. It is calculated separately for each cluster. The higher value of the Caliński-Harabasz is considered as the dataset is clustered well. There are two metrics to calculate in order to get the Caliński-Harabasz value. One is the within-cluster sum of squares, and the second one is the between-cluster sum of squares.

2.3.1 Within cluster sum of squares

In order to calculate within cluster sum of squares (WCSS), for each cluster, one needs to calculate the average distance between each observation and cluster centroid [32]. Assume the dark colored dot is a cluster centroid and for a concrete illustration please see *Figure 3*. *Equation 5* and *6* explains how WCSS is calculated.

$$WCSS_k = \sum_{i=1}^{n_k} (\, x_i - c_{ck} \,)^2 \quad (5)$$

$$WCSS = \sum_{k=1}^{K} WCSS_k \quad (6)$$

$WCSS_k$ is the within cluster sum of squares of cluster i,
n_k is number of the observations in cluster that x_i is assigned to,
x_i is the ith observation of c_k ,
c_{ck} is the centroid of cluster c_k .

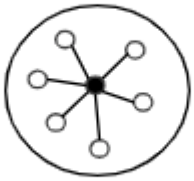

Figure 3. illustration of the WCSS calculation for a cluster

2.3.2 Between cluster sum of squares

For each cluster, between cluster sum of squares (BCSS) is the Euclidean distance between the centroid of that cluster and the centroid of the dataset(barycenter) [32]. Assume the dark colored dot in the middle is barycenter, the other dark colored ones are the centroids of the clusters and for a concrete illustration please see *Figure 4*. Equation 7 explains how BCSS is calculated.

$$BCSS = \sum_{k=1}^{K} n_k \times (c_{ck} - C)^2 \qquad (7)$$

where C is the barycenter of the dataset.

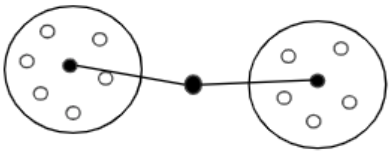

Figure 4. illustration of BCSS calculation for two clustered data

After WCSS and BCSS are calculated, Caliński-Harabasz value can be calculated by using equation 8.

$$CH = \frac{BCSS}{WCSS} - \frac{n-k}{K-1} \qquad (8)$$

2.4 Davies - Bouldin Method

Essentially, Davies - Bouldin is an index to calculate similarity between clusters [33]. In a four step calculation, Davies - Bouldin method finds the most similar cluster for each cluster. It is calculated separately for each cluster. Davies-Bouldin takes values between 0 and 1. Lower value of the Davies - Bouldin means the dataset is clustered well.

2.4.1 Intra-cluster dispersion

Intra-cluster dispersion measures distribution of the cluster. It is, for each cluster, the average distance between cluster centroid and each observation. Assume the dark colored dot is a cluster centroid and for a concrete illustration please see *Figure 5.* Equation 9 explains how intra-cluster dispersion is calculated.

$$S_k = \frac{1}{n_k} \sum_{i=1}^{n_k} (x_i - c_{ck})^2 \qquad (9)$$

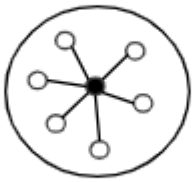

Figure 5. illustration of S_k calculation

2.4.2 Separation criteria

In this step, the sum of squared distances between each cluster centroid is calculated. In this way, how well the clusters separation is determined. It should be noted that for each cluster centroid, there are $k-1$ distances. Let the cluster number be three. To calculate separation criteria for cluster 1, distance between both centroid of cluster 1 and centroid of cluster 2, and centroid of cluster 1 and centroid of cluster 3 is calculated. Assume the dark colored dots are the centroid of the clusters and for a concrete illustration for two clustered data, please see *Figure 6.* Equation 10 explains how separation criteria is calculated.

$$M_{kk'} = (c_{ck} - c_{ck'})^2 \qquad (10)$$

where,
c_{ck} is the centroid of cluster k,
$c_{ck'}$ is the centroid of cluster k'.

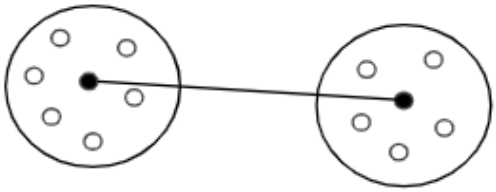

Figure 6. illustration of the calculation of Separation Criteria for two clustered data

2.4.3 Cluster similarity & most similar clusters

Let K be 3 again. In order to calculate cluster similarity for the cluster 1 and 2, values for cluster 1 and 2 calculated in the first step are summed. Then, it is divided to the value for cluster 1 and 2 that is calculated in the second step. Just like the second step, for each cluster there are *k-1* similarity measurements $R_{kk'}$. Higher value of the $R_{kk'}$ means that for cluster k, the most similar cluster is k'. Equations 11 and 12 explains how similarity and most similar clusters are calculated:

$$R_{kk'} \equiv \frac{S_k + S_{k'}}{M_{kk'}} \tag{11}$$

$$R_k \equiv max(R_{kk'}) \tag{12}$$

where,
S_k is the intra cluster dispersion of cluster k,
$S_{k'}$ is the intra cluster dispersion of cluster not k,
$M_{kk'}$ is the separation criteria of cluster k and cluster not k.

After the first three steps are calculated, Davies - Bouldin index is calculated by applying Equation 13.

$$\bar{R} \equiv \frac{1}{K}\sum_{k=1}^{K} R_k \tag{13}$$

where K is the total cluster number.

2.5 Dunn index

Dunn Index is calculated by dividing minimum separation to maximum diameter [34] It is calculated separately for each cluster. Higher value of the Dunn Index means data is clustered well. The minimum distance between observations that belong to the separate clusters is called minimum separation. Assume dark colored dots represent the closest observations between two separate clusters and for a concrete illustration please see *Figure 7*. On the other hand, the maximum distance between observations that are assigned to the same cluster is called maximum diameter. Assume dark colored dots represent the furthest observations in the same cluster and for a concrete illustration please see *Figure 8*. *Equation 14* and *15* explains how minimum separation and maximum diameter are calculated:

$$min_{separation} = min(d(x_i, x_{i'})_{c_k \neq c_{k'}}) \tag{14}$$

where,

x_i are the observations in c_k
$x_{i'}$ are the observations in $c_{k'}$.

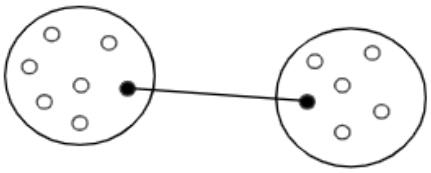

Figure 7. illustration of the minimum separation calculation for two clustered data

$$max_{diameter} = max(d(x_i, x_{i'})_{x_i, x_{i'} \in c_k}) \tag{15}$$

where,

x_i are the observations in c_k
$x_{i'}$ are the observations other than x_i in c_k .

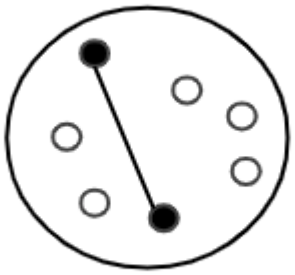

Figure 8. illustration of the maximum diameter calculation

2.6 Rand index

The Rand Index is a measure of the similarity between clustering result and label of the dataset. It was introduced by William M. Rand in 1971 and is defined as the ratio of the total number of observations that clustered correctly to the total number of observations [35]. The Rand Index takes values from 0 to 1, with a value of 1 indicating that the dataset is clustered perfectly, and a value of 0 indicating that dataset clustered imperfectly. It is often used as a measure of the quality of a clustering algorithm or as a measure of the similarity between different clustering techniques applied to the same dataset.

To compute the Rand Index, we first need to define a confusion matrix, which is a matrix that compares the elements in the clustering result and the actual values. *Table 1* explains the confusion matrix in detail. The confusion matrix has four entries: a, b, c, and d. The entry a represents the number of observations that are in the same cluster in both clustering result and label. The entry b represents the number of observations that are in the same cluster in the label but in different clusters in the clustering result. The entry c represents the number of observations that are in different clusters in the label but in the same cluster in the clustering result. The entry d represents the number of observations that are in different clusters in both clustering result and label. *Equation 16* explains how similarity and most similar clusters are calculated.

Table 1. Confusion matrix

		Label	
		1	0
Clustering result	1	a	b
	0	c	d

$$RI = \frac{a+d}{a+b+c+d} \tag{16}$$

2.7 Meila's variation of information

MVI is a measure of the dissimilarity between clustering result and label of the dataset. It was introduced by Marina Meila in 2005 as an alternative to the Rand Index and other measures of cluster similarity that are based on the confusion matrix. The VI is defined as the sum of the mutual information between the two clustering and the entropy of each clustering [36]. Mutual information is a measure of the amount of information shared between two random variables, and entropy is a measure of the amount of uncertainty or randomness in a random variable. MVI takes values between 0 to infinity, with a value of 0 indicating that clustering result and the label are identical and a larger value indicating greater dissimilarity. It is often used as a measure of the quality of a clustering algorithm or as a measure of the similarity between different clustering techniques applied to the same dataset. MVI is calculated by computing the mutual information between the clustering result and label and the entropy of clustering result and label. *Equation 17* explains how the mutual information is calculated:

$$MI = \sum_{k=1}^{n_k} \sum_{k'=1}^{n_{k'}} p(k,k') log \frac{p(k,k')}{(p(k)p'(k'))} \tag{17}$$

where,

$p(k)$ and $p'(k')$ are random variables associated with the clusterings c_k and $c_{k'}$,

$p(k,k')$ is the probability that a pair of elements belong to c_k in the clustering result and $c_{k'}$ in the label

Second step is the calculation of the entropy of the clustering result and the label. *Equation 18* explains how the entropy is calculated for the clustering result. It is also calculated for the label.

$$H = -\sum_{k=1}^{n_k} p(k)\log(p(k)) \quad (18)$$

Finally, the Variation of Information is calculated by using *Equation 19* follows:

$$VI = H(k) + H(k') - 2(MI) \quad (19)$$

where $H(k)$ and $H(k')$ are the entropies of the clustering result and the label, MI is the mutual information between clustering result and the label.

3 Real data analysis

R programming language was used in all of the analyses performed in this study. The *factoextra* [37] package was used for the clustering algorithms, the *NbClust* [38] package was used for the functions to determine the optimal number of clusters, and the *fpc* [39] package was used to calculate the Rand Index and MVI scores.

In this section, analyses that are done in nine real datasets from different study areas are used. Except for User Knowledge [40] and Appendicitis [41] datasets, all other datasets, which are wine, e.coli, breast cancer wisconsin, column3c, iris, haberman, and breast tissue, were taken from the UCI Machine Learning Depository [42]. In all of the data sets used in the study, there is one variable containing class information; however, this dependent variable was removed from the data set before clustering. Class information is used to calculate Rand Index and MVI scores at the end of clustering and to measure whether the number of clusters is correctly decided. Information on how many clusters(class) the data sets consist of and the process applied in the data preparation process can be found in *Table 2.* The codes of the study can be found at https://github.com/ozturkfemre/optimal_k .

Table 2. Information about datasets

Dataset	Abbreviation	# of class	Preparation
Wine	d1	3	Standardization
E.coli	d2	8	Standardization
Breast Cancer Wisconsin	d3	2	PCA
Column3c	d4	3	PCA
Iris	d5	3	PCA
Haberman	d6	2	Standardization
Breast Tissue	d7	6	PCA
Appendicitis	d8	2	PCA
User Knowledge	d9	4	Standardization

As can be seen from *Table 2*, PCA was applied to some datasets during the data preparation, while some datasets were standardized. When the correlation matrix of breast cancer wisconsin, column3c, iris, breast tissue, and appendicitis datasets are examined, high correlation is noticed in the combinations of variable pairs. PCA was applied to these datasets to prevent this correlation from negatively affecting the analysis.

Table 3 shows the number of clusters suggested by the four methods for each data set. Among the four methods, Caliński-Harabasz and Average Silhouette are found to be the most successful method by correctly determining the cluster of 2 out of 9 datasets, while the others seem to be correctly determining in 1 dataset.

Table 3. Suggested cluster numbers

Data	# of class	Average Silhouette	Caliński - Harabasz	Davies - Bouldin	Dunn Index
d1	3	3	3	3	3
d2	8	4	4	5	6
d3	2	2	2	7	6
d4	3	2	2	9	10
d5	3	2	9	2	2
d6	2	4	5	5	6
d7	6	2	4	4	2
d8	2	3	4	10	10
d9	4	9	2	9	8

At this point, it may be necessary to draw attention especially to the d2 dataset. In *Figure 9* which is the scatter plot of the d2 dataset, it was found that the number of observations included in some of the 8 classes of the dataset was less than five (see purple, green, and pink observations. Considering the number of clusters proposed by all methods, it was realized that these methods had difficulty in detecting small clusters.

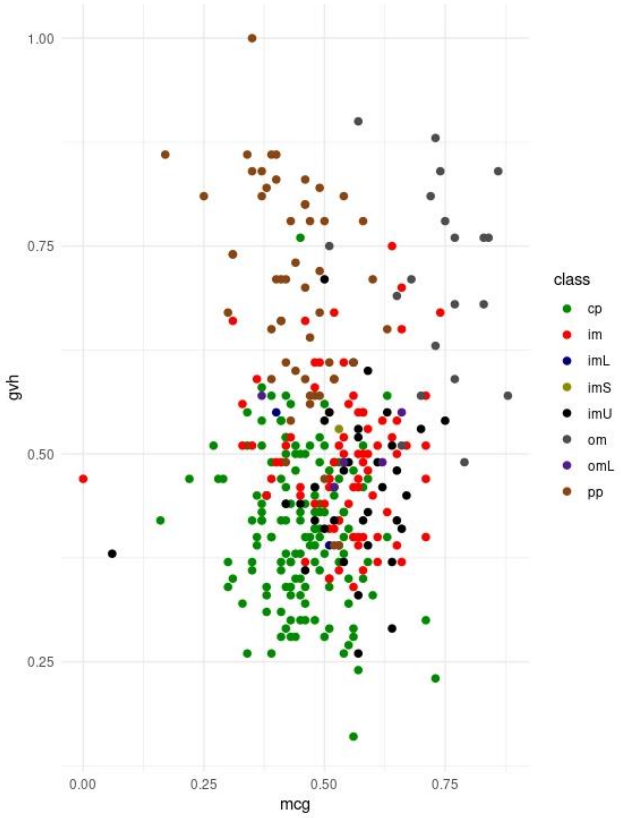

Figure 9. Scatterplot of d2
(each color represents a class)

Another striking feature is in the d5. Three of the four methods suggested two cluster numbers for the d5. As can be seen in *Figure 10*, the scatterplot of the PCA applied d5 data set shows a two-cluster decomposition, but in any case, it needs to be noted that the methods failed because the actual number of classes is three. It is also noteworthy that the Average Silhouette method suggests two clusters in four out of nine datasets, no matter how diverse the datasets and how different the number of classes.

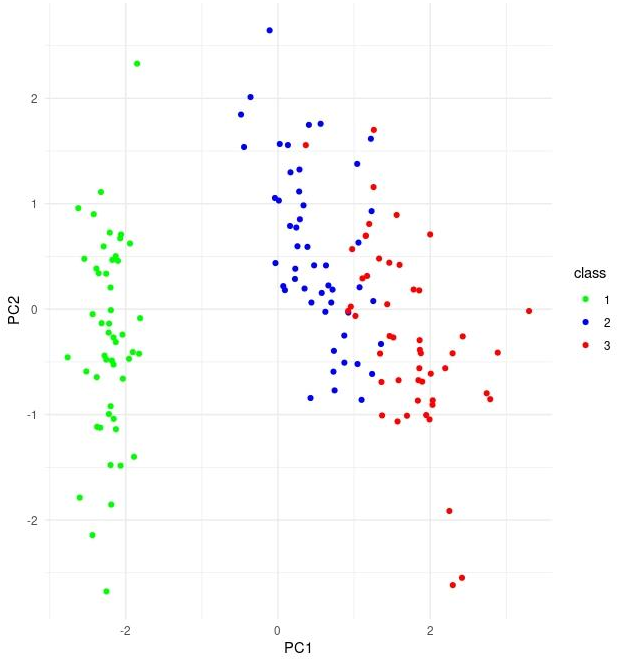

Figure 10. Scatterplot of d5
(each color represents a class)

4 Discussion & Conclusion

This study set out to compare four different methods (Average Silhouette, Caliński-Harabasz, Davies-Bouldin, and Dunn Index) to determine the optimal number of clusters in nine real datasets [43]. Each dataset is clustered by the k-means algorithm. The most striking result to emerge from this study is that methods failed to determine the optimal number of clusters correctly on the majority of the data sets. *Table 4* shows the averages of Rand Index and MVI values calculated for each method for nine real data sets and success rate of each method. When the Rand Index and MVI metrics, which measure the accuracy of clustering observations into clusters, are analyzed, it is possible to have information about how the methods approach the correct number of clusters, even if they cluster incorrectly. Although Average Silhouette and Caliński-Harabasz correctly determined an equal number of datasets, the average Rand Index and average MVI values suggest that Average silhouette is closer to the correct number of clusters in the datasets that it incorrectly determined.

Table 4. Average metrics and success rates of methods

Methods	Rand Index	MVI	Success Rate
Average Silhouette	**0.40**	**1.14**	**0.22**
Caliński - Harabasz	0.38	1.29	**0.22**
Davies - Bouldin	0.33	1.54	0.11
Dunn Index	0.30	1.56	0.11

Although there are no other studies that use validation metrics (Rand Index and Meila's Variation of Information) for evaluation, there are some comparisons in several studies that the methods used in this study is compared The same results were obtained on the Iris data set where NbClust package is presented [44]. Nanjundan et al. [45] also compared Average Silhouette method with the proposed method on six data sets, including Iris and Breast Cancer Wisconsin, and obtained the similar results. On 12 data sets, 7 of which are the same, the I-nice algorithm [46] is compared with the Elbow and Average Silhouette approaches. The same results were obtained.

The evidence from this study suggests that methods tend to fail in some specific cases. Future research should therefore concentrate on developing new methods and algorithms. For the former, since Average Silhouette and Caliński-Harabasz are the methods which are closer to the better success, they can be starting points for following studies.

References

[1] El Naqa, I., & Murphy, M. J. What is machine learning?. In *machine learning in radiation oncology* (pp. 3-11). Springer, Cham. 2015

[2] Learned-Miller, E. G. Introduction to supervised learning. *I: Department of Computer Science, University of Massachusetts*, 3. 2014

[3] Hady, M. F. A., & Schwenker, F. Semi-supervised learning. *Handbook on Neural Information Processing*, 215-239. 2013

[4] Hinton, G., & Sejnowski, T. J. (Eds.). Unsupervised learning: foundations of neural computation. *MIT press*. 1999

[5] Holzinger, K. J., & Harman, H. H. *Factor analysis; a synthesis of factorial methods*. 1941
[6] Sokal, R. R. Numerical taxonomy. *Scientific American*, 215(6), 106-117. 1966
[7] Barbará, D., & Jajodia, S. (Eds.). Applications of data mining in computer security (Vol. 6). *Springer Science & Business Media.* 2002
[8] Mirkin, B. Clustering for data mining: a data recovery approach. *Chapman and Hall/CRC*. 2005
[9] Chou, C. H., Su, M. C., & Lai, E. A new cluster validity measure and its application to image compression. *Pattern Analysis and Applications,* 7(2), 205-220. 2004
[10] Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., & Perona, I. An extensive comparative study of cluster validity indices. *Pattern recognition*, 46(1), 243-256. 2013
[11] Berthold, M. R., & Höppner, F. On clustering time series using euclidean distance and pearson correlation. *arXiv preprint arXiv*:1601.02213. 2016
[12] Kassambara, A. Practical guide to cluster analysis in R: Unsupervised machine learning (Vol. 1). *Sthda*. 2017
[13] Marutho, D., Handaka, S. H., & Wijaya, E. The determination of cluster number at k-mean using elbow method and purity evaluation on headline news. *In 2018 international seminar on application for technology of information and communication (pp. 533-538).* IEEE. 2018, September
[14] Govender, P., & Sivakumar, V. Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019). *Atmospheric pollution research*, 11(1), 40-56,2020
[15] Sinaga, K. P., & Yang, M. S. Unsupervised K-means clustering algorithm. *IEEE* access, 8, 80716-80727. 2020
[16] Yuan, C., & Yang, H. Research on K-value selection method of K-means clustering algorithm. *J*, 2(2), 226-235. 2019
[17] Celebi, M. E., Kingravi, H. A., & Vela, P. A. A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert systems with applications*, 40(1), 200-210. 2013
[18] Hartigan, J. A., & Wong, M. A. Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society.* Series C (Applied Statistics), 28(1), 100–108. https://doi.org/10.2307/2346830. 1979
[19] Löster, T. Determining the optimal number of clusters in cluster analysis. *Proceedings of the 10th International Days of Statistics and Economics*, Prague, Czech Republic, 8-10. 2016.
[20] Desgraupes, B. Clustering indices. *University of Paris Ouest-Lab Modal'X*, 1, 34. 2013
[21] Hasanpour, H., Asadi, S., Meibodi, R. G., Daraeian, A., Ahmadiani, A., Shams, J., & Navi, K. A critical appraisal of heterogeneity in obsessive-compulsive disorder using symptom-based clustering analysis. *Asian journal of psychiatry*, 28, 89-96. 2017.
[22] Xu, R., Xu, J., & Wunsch, D. C. A comparison study of validity indices on swarm-intelligence-based clustering. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(4), 1243-1256. 2012.
[23] Anitha, S., & Metilda, M. A. R. Y. (2019). An extensive investigation of outlier detection by cluster validation indices. *Ciencia e Tecnica Vitivinicola-A Science and Technology Journal*, 34(2), 22-32. 2019.
[24] Baarsch, J., & Celebi, M. E. Investigation of internal validity measures for K-means clustering. *In Proceedings of the international multiconference of engineers and computer scientists (Vol. 1, pp. 14-16)*. sn. 2012.
[25] Hartigan, John A., Manchek A. Wong. Algorithm AS 136: A k-means clustering algorithm. J*ournal of the royal statistical society. series c (applied statistics)* 28., 100-108, 1979
[26] Kassambara, Alboukadel. *Practical guide to cluster analysis in R: Unsupervised machine learning.* Vol. 1. Sthda, 2017.
[27] Ben-Hur, Asa, and Isabelle Guyon. Detecting stable clusters using principal component analysis. *Functional genomics*. Humana press, 159-182, 2003.
[28] Ding, Chris, and Xiaofeng He. K-means clustering via principal component analysis. *Proceedings of the twenty-first international conference on Machine learning*. 2004.
[29] Kassambara, Alboukadel. Practical guide to cluster analysis in R: Unsupervised machine learning. Vol. 1. *Sthda*, 2017.
[30] Hartigan, John A., Manchek A. Wong. Algorithm AS 136: A k-means clustering algorithm. J*ournal of the royal statistical society. series c (applied statistics)* 28., 100-108, 1979
[31] Rousseeuw, Peter J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics,* 1987, 20: 53-65.
[32] Caliński T, Harabasz J. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods* 3.1, 1-27, 1974.
[33] Davies DL, Bouldin, DW. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2), 224-227, 1979
[34] Dunn JC. *A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters.* 32-57, 1973
[35] Warrens, M. J., & van der Hoef, H. Understanding the rand index. In Advanced Studies in Classification and Data Science (pp. 301-313). *Springer,* Singapore. 2020
[36] Meilă M. "Comparing clusterings -an information based distance." *J.Multivariate Analysis*, 98(5), 873-895, 2007.

[37] Kassambara, Alboukadel, and Fabian Mundt. "Factoextra: Extract and Visualize the Results of Multivariate Data Analyses." *https://CRAN.R-project.org/package=factoextra*. 2020

[38] Charrad, Malika, Nadia Ghazzali, Véronique Boiteau, and Azam Niknafs. "NbClust: An r Package for Determining the Relevant Number of Clusters in a Data Set" 61. *https://www.jstatsoft.org/v61/i06/.* 2014.

[39] Hennig, Christian. "Fpc: Flexible Procedures for Clustering." *https://CRAN.R-project.org/package=fpc.* 2020.

[40] H. T. Kahraman, Sagiroglu, S., Colak, I., Developing intuitive knowledge classifier and modeling of users' domain dependent data in web, *Knowledge Based Systems*, vol. 37, pp. 283-295, 2013.

[41] WEISS, Sholom M.; KULIKOWSKI, Casimir A. Computer systems that learn: classification and prediction methods from statistics, neural nets, machine learning, and expert systems. M*organ Kaufmann Publishers* Inc. 1991.

[42] Dua, D. and Graff, C. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: *University of California, School of Information and Computer Science*. 2019.

[43] Öztürk, F.E., and Demirel, N. Comparison of the Methods to Determine Optimal Number of Clusters. International *Symposium in Graduate Researches on Data Sciences*, 2-3 December 2022.

[44] Charrad, Malika, Nadia Ghazzali, Véronique Boiteau, and Azam Niknafs. "NbClust: An r Package for Determining the Relevant Number of Clusters in a Data Set" 61. *https://www.jstatsoft.org/v61/i06/.* 2014.

[45] Nanjundan, Sukavanan, et al. Identifying the number of clusters for K-Means: A hypersphere density based approach. *arXiv preprint arXiv:1912.00643*. 2019.

[46] Masud, Md Abdul, et al. "I-nice: A new approach for identifying the number of clusters and initial cluster centres." *Information Sciences* 466.129-151. 2018.