www.revistageintec.net
ISSN: 2237-0722
revista GEINTEC
Gestão, Inovação e Tecnologias

A Comprehensive Study on the Importance of the Elbow and the Silhouette Metrics in Cluster Count Prediction for Partition Cluster Models

A.A. Abdulnassar[1]; Latha R. Nair[2]
[1]School of Engineering, Cochin University of Science and Technology, India.
[1]nasrishabaa@gmail.com
[2]School of Engineering Cochin, University of Science and Technology, India.
[2]latha5074@gmail.com

Abstract

Proper selection of cluster count gives better clustering results in partition models. Partition clustering methods are very simple as well as efficient. Kmeans and its modified versions are very efficient cluster models and the results are very sensitive to the chosen K value. The partition clustering algorithms are more suitable in applications where the data are arranged in a uniform manner. This work aims to evaluate the importance of assigning cluster count value in order to improve the efficiency of partition clustering algorithms using two well known statistical methods, the Elbow method and the Silhouette method. The performance of the Silhouette method and Elbow method are compared with different data sets from the UCI data repository. The values obtained using these methods are compared with the results of cluster performance obtained using the statistical analysis tool Weka on the selected data sets. Performance was evaluated on cluster efficiency for small and large data sets by varying the cluster count values. Similar results obtained from the three methods, the Elbow method, the Silhouette method and the clustering by Weka. It was also observed that the fast reduction in clustering efficiency for small changes in cluster count when the cluster count is small.

Key-words: Cluster, Data Mining, Kmeans Partition Algorithm, Cluster Count Prediction, Elbow and Silhouette Method.

1. Introduction

Clustering and classification are the major machine learning methods used in data mining. Partition algorithms are simple and best suited for many cluster applications [1]. Cluster count selection and the initial seed selection affect the performance of clustering applications. There are

Received: 17.06.2021 – Accepted: 18.07.2021

many conventional methods suggested for the proper selection of cluster count to improve the quality of cluster [2]. The performances of k value selection methods vary with data sets.

The major drawbacks of the partition algorithm are found to be its cluster count selection and initial seed selection. The performance of different initialization algorithms and cluster count selection methods are compared in this study. The Silhouette and Elbow methods are two popular statistical methods used for finding the optimal cluster count. Small variation in the selection of cluster count as well as initial seed selection can affect the cluster performance [3].

The Elbow method gives the results based on the general behavior of the samples in the data sets. The Silhouette metric systematically measures the compactness of the samples in the cluster. The Silhouette is a measure of how close are the samples within a cluster and how far from the nearest cluster. In many applications the Silhouette metric predicts the cluster count more accurately. Many tools and algorithms are available to evaluate the performance of the partition models.

2. Cluster Count Selection Methods

In machine learning problems cluster algorithms are used as a pre-processing tool [4]. This section reviews existing methods for selecting the most appropriate cluster count value k for the partition algorithms. Both conventional methods and statistical methods are used to find the cluster count.

2.1. The Elbow Method

It is an ambiguous and heuristic method to find cluster count value in a data set [9]. In this method, a graph is plotted with the data prepared between the percentage variance as a function cluster count value. The plotted graph is analysed and the elbow position is identified in it and chosen as cluster count value. The variance function is calculated as the ratio of the between group variance to the total variance of samples in a cluster. Objective function reflects the intra cluster distance relative to inter cluster distance in Kmeans cluster models.

2.2. The Silhouette Coefficient Method

Silhouette coefficient is calculated using the mean inter cluster distance (i) and the mean nearest cluster distance (n) for each sample. The Coefficient for a sample is calculated using the formula, sc= (n-i) / max(i,n), where n is the distance between a sample and the nearest cluster that the sample is not a part of. We can compute the mean Silhouette Coefficient over all samples and use this as a metric to decide the cluster count. This method is more accurate to determine the cluster count value. Cluster performance depends on the low inter cluster distance and the high intra cluster metric [16].

2.3. Other Conventional Methods

The user can specify the cluster count value by performing the cluster algorithm after varying the values of k. By comparing the cluster performance with different data sets and variations of models, user can predict almost correct value for the cluster. The small changes in the value selected in cluster count can cause deficient cluster formation. The cluster accuracy and the cluster formation performances for a range of values using different data sets can decide the most suitable cluster count value. Depending on the types of data sets used, the count can also vary. In many applications the clustering and classification are done based on the class field. In such cases, the class value can be chosen as cluster count to form more accurate clusters. In the statistical approach, the Gaussian distribution and Bayesian classifiers are used to find the suitable cluster count value [5]. Monte Carlo techniques, which are associated with the null hypothesis is also used for assessing the clustering performance and for finding cluster count. Probabilistic clustering methods do not take into account the distortion inside a cluster and so the clusters created by applying such methods may not correspond to the clusters in partitioning clustering.

3. Previous Works

Cluster analysis is one of the promising areas used as a preliminary step in grouping the data for various data mining applications. Many modifications have been suggested for improving the performance of conventional Kmeans partition clustering algorithms [6]. The major emphasis of the research works are in the area of selection of cluster count value. Selecting proper initial centroids is another method used in cluster models. Some of the important works in this area are discussed in this section.

Gregory Wilkin, Xiuzhen and Huang suggest two types of K means algorithms and compared the algorithms using data from biological domain. They discussed many conventional cluster count selection procedures in this work. The running time and the compactness of cluster are taken as the performance measures [5]. The study claims that the progressive greedy Kmeans clustering algorithm gives better results.

Shi Na, Liu Xumin and Guan Yong suggest an improved Kmeans cluster method, avoiding the distance calculations in each iteration. They propose a data structure to store the intermediate distances to avoid the unnecessary repeated distance calculations. In this method, the results of previous iterations can be stored and used for next iterations. Efficiency of the algorithm can be improved by avoiding unnecessary calculations. The method improves the cluster speed by reducing the repeated distance calculation in iteration [1]. Experimental results show the improvement in cluster accuracy and speed by eliminating unnecessary complex calculations.

T Baolin Yi and Haiquan Qiao have proposed a new method for selecting initial cluster centroid for Kmeans algorithm and tried to minimize the sensitivity issue in selecting proper initial cluster centre and correct cluster count[7]. The experiment is done using the data set from UCI, fixing the initial centroids in a previously identified dense area. The proposed method promises better initial seed selection in Kmeans clustering.

Wang Yintong and Li Wanlong suggests a new method for selecting the initial cluster centre[8]. In this method density based method is used to choose initial seeds. The experiment have generated more compact clusters. Data from the UCI repository is used for the comparison. This method can reduce the noises in initialising the cluster centres and can improve the quality of clusters.

Purnima Bholowalia and Arvind Kumar suggests K-means based quick clustering algorithms to produce a new cluster scheme for WSNs with dynamic selection of the number of the clusters [9]. Since the correct K value is very important in this work, they use Elbow method to find the cluster count k.

A. Rahman and B. Verma studied the nature of different cluster algorithms and proposed a novel ensemble classifier[10]. They suggested a new ensemble model with two stages known as Base classifier and fusion classifier. The base classifier produces the cluster confidences by learning the cluster boundaries. The fusion classifier takes the decisions by combining the cluster confidences. The work is done using the data set from UCI data repository and the performance of the models is shown using two tailed sign test.

Mengxing Huang and Hongjing Lin analyzes the advantages and disadvantages of the traditional K-means algorithm, and proposes the K-means clustering algorithm of events based on variable time granularity[11]. The experiments show that the improved algorithm is more suitable for clustering analysis of Weibo event, improves the efficiency of clustering algorithm, and solves the initial cluster centers sensitive issue, compared with the traditional K-means algorithm.

Caiquan Xiong and Zhen Hua discuss the dependency of the cluster performance with initial seed selection. They suggest a method to find out suitable initial centres by removing the isolated points from the data sets[12]. The experimental results claim that the performance of cluster is improved by selection of initial cluster centre from denser area in the data sets.

Channamma Patil and Ishwar Baidari suggests a new method called depth difference to calculate the optimal number of clusters (*k*) in a dataset based on data depth[13]. This method initially calculates the cluster count and use the value for the cluster algorithm. In this method depth within clusters and between clusters are calculated and the depth difference is used to determine the k value.

Chunhui Yuan and Haitao Yang analysed the importance of k value selection for fast converging of the Kmeans algorithm[14]. They analysed the Elbow method, Silhouette method, Gap statics and the Canopy method for the k value selection. They used IRIS data sets from the UCI for their study. The analysis helps the researchers to select proper cluster count finding methods for their application.

Congming Shi, Bingtao Wei and Shoulin Wei used Elbow method to find the cluster count value in partition cluster. It is not easy to identify the elbow point from the plotted curve if the plotted curve is fairly smooth. They proposed a new elbow point discriminant method to get a metric which is used to find the cluster count value of the dataset[15]. They claim that the experimental results show that the method gives better result than the silhouette method.

Tai Dinh, Tsutomu Fujinami and Van-Nam Huynh proposes categorical clustering algorithm to predict the k value. The algorithm uses the kernel density estimation approach to define cluster centers and information theoretic based dissimilarity to measure the distance between centers and objects in each cluster[16]. They apply the silhouette analysis based approach to assess the quality of cluster to select the best cluster count value.

The above review listed some of the important works undertaken to improve the problems faced by the partitioning algorithms. They suggest suitable methods to improve the selection of cluster count value for better and more accurate clustering results. Many initial seed selection algorithms have been proposed by them to improve the quality of clusters in Kmeans algorithm.

4. Methodology

The main objective of the work is to analyse the performance of the methods used to find the initial cluster count value. Two popular methods are used to find the K value namely, the Elbow method and the Silhouette method. The results obtained using these methods are compared with that of other methods. The algorithms are compared using three standard data sets from the UCI data repository.

The study discussed the importance of rectifying the deficiencies of partition algorithms. The work also analyses the cluster efficiency variations by changing the value obtained by Elbow and Silhouette methods. The experiments are conducted by implementing the Elbow and Silhouette methods in Python language using the standard data sets from the UCI data repository. The performances of the algorithms are compared using different k values. The three different sets of data are used for the performance analysis are given in Table 1. The experimental sample points in Data1 data set are also used for the performance evaluation of the Silhouette metric.

Table 1 - Data Sets Used

Dataset Name	No. of Attributes	No. of Instances
Abalone	9	4177, 1253
Wine	14	178
Iris	5	150

There are many tools available to analyse the results of the clustering. The field based cluster checking is very easy in cluster analysis. The fields which are having specific classes can easily be used to calculate the accuracy of the cluster. Hence we mainly select the class field for easy and convincing analysis [17]. The efficient partition algorithms, Kmeans and Farthest First are used for the analysis. Two famous metrics, Elbow and Silhouette are mainly used to find the cluster count value. The performance of the cluster accuracy for small changes in K values has also been conducted for small and large data sets.

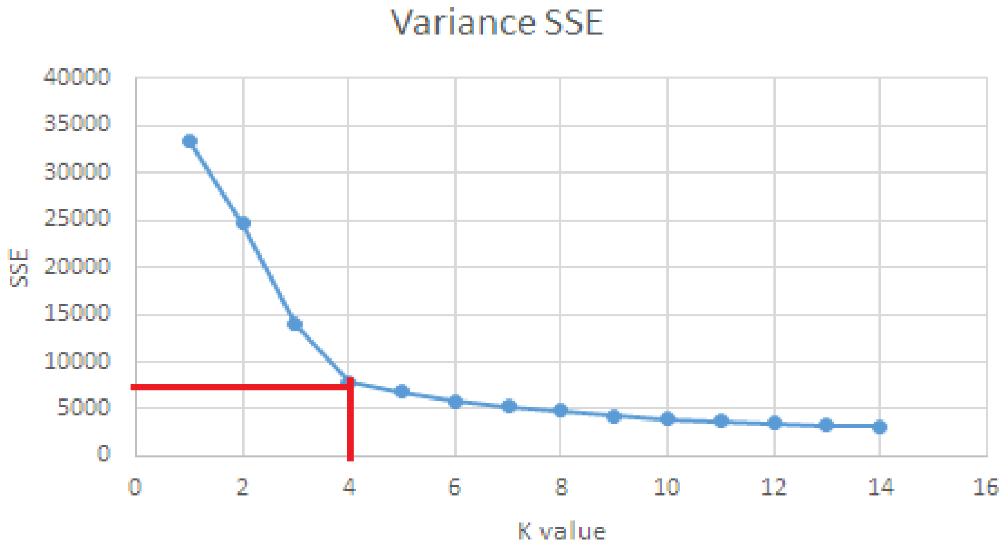
5. Results and Discussions

The main aim of this study is to find the suitable cluster count value for the Kmeans clustering. The variations of the Elbow method and the Silhouette method are used to find the k value for the Kmeans clustering. The performances of the methods are analysed using standard data sets from the UCI repository and some general data samples are used for the study.

5.1. Elbow Method

The Sum of Squared Error of the clusters corresponding to different k values are found for the Abalone data set. The variance values corresponding to each k values are taken to plot the graph. The line graph between the different K value and Sum of Squared Error are drawn. The Elbow chart of Abalone data set shows elbow position corresponding to the k value 4. In the case of Abalone data set, the cluster results are drawn for the data sets to find the accuracy of cluster formed. The weka tool is used to check accuracy of clusters formed with the cluster count value 4.

Fig. 1 - Elbow Chart for Abalone Data Set

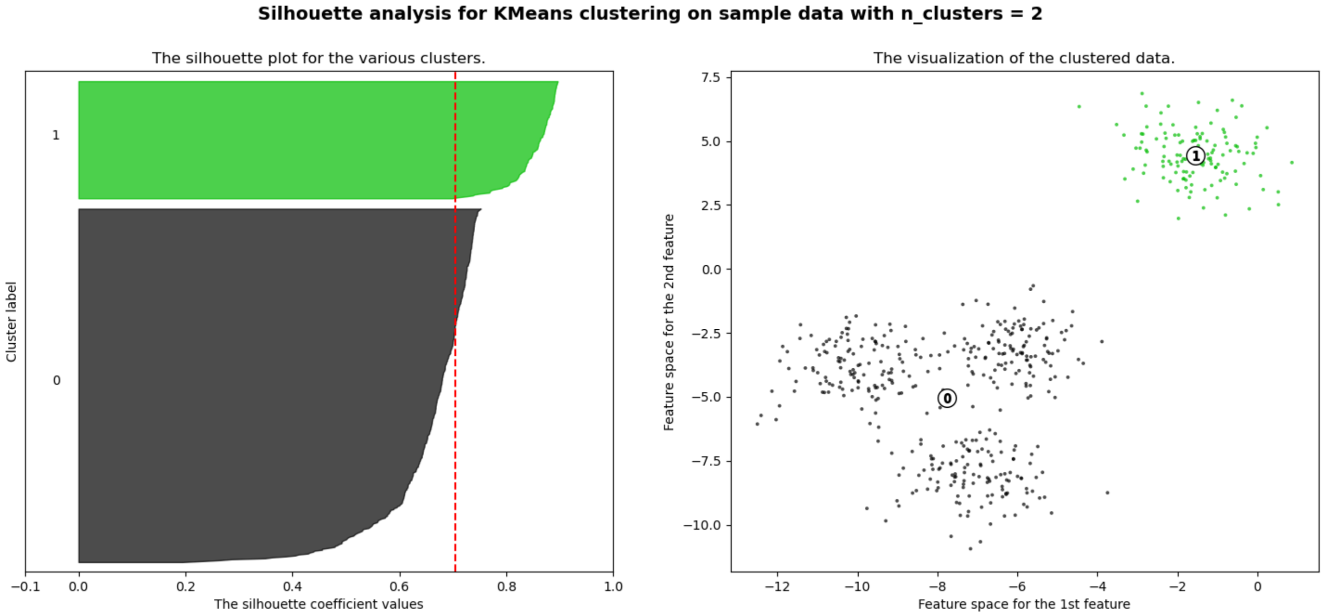


5.2. Silhouette Method

This is an efficient method to find the cluster count suitable for a clustering application. The silhouette value is a measure of compactness of samples within a cluster and the separation between different clusters. The performance of cluster is improved if the inter cluster distance among samples is low and the intra cluster distance among samples in one cluster to the neighboring cluster is high. The silhouette plot shows the measure of how close each point in one cluster is to points in the neighboring clusters. The method can be used to find the optimum cluster count. Silhouette value ranges between -1 and 1. The value +1 indicates that the sample is far away from the neighboring clusters. The 0 value indicates that the sample is very near to the nearby clusters and the negative value indicates the presence of outlier in the cluster [19].

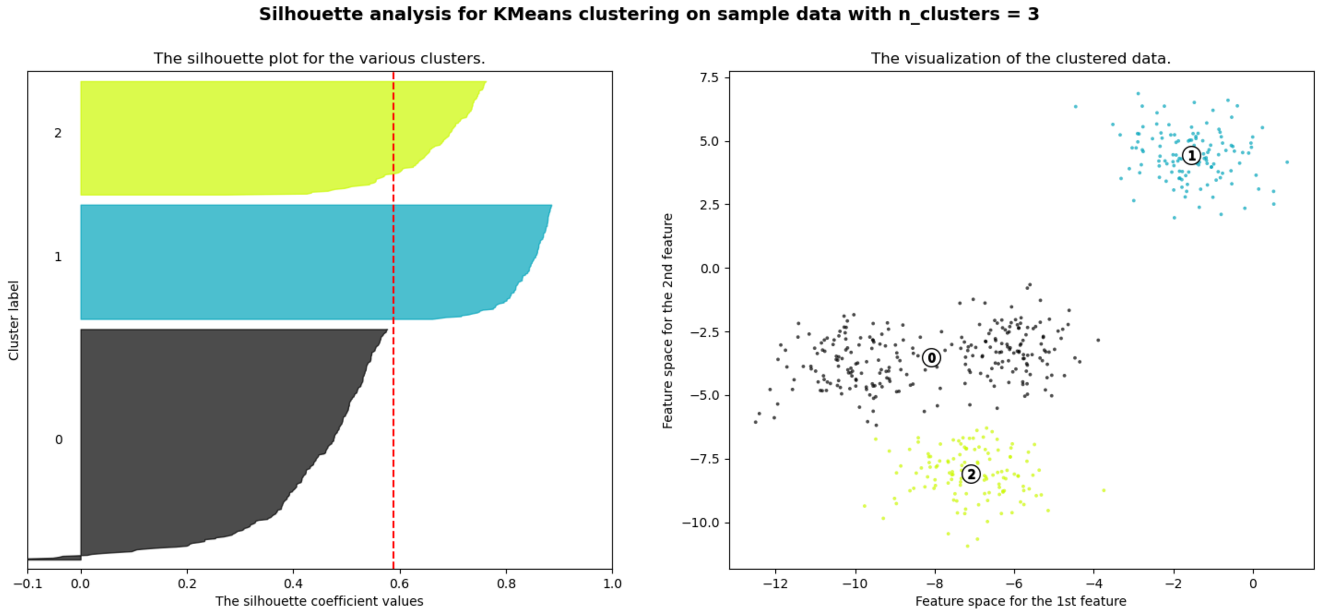
The silhouette values for 3 data sets are used for the analysis. The values and the graph of Abalone data sets are shown in Figure 2. The silhouette value of each sample and the average silhouette value for each cluster for different cluster count values are taken for the study. The higher average cluster silhouettes and the average overall silhouettes are considered for the cluster count selection.

Fig. 2 - Silhouette Plot with K Value 2



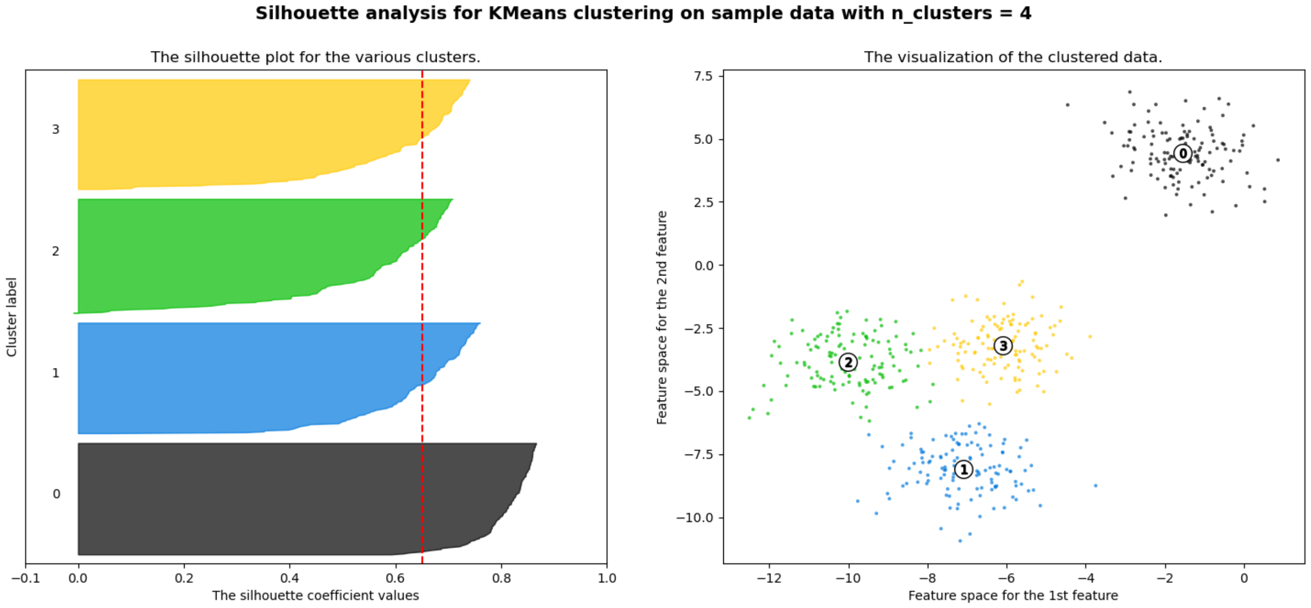
Here the result of silhouette plot with cluster count value 2 is shown. Silhouette value of first cluster is very high and the overall average silhouette is also high. But the average silhouette of second cluster is below the threshold value. Hence cannot consider this as k value.

Fig. 3 - Silhouette Plot with k Value 3



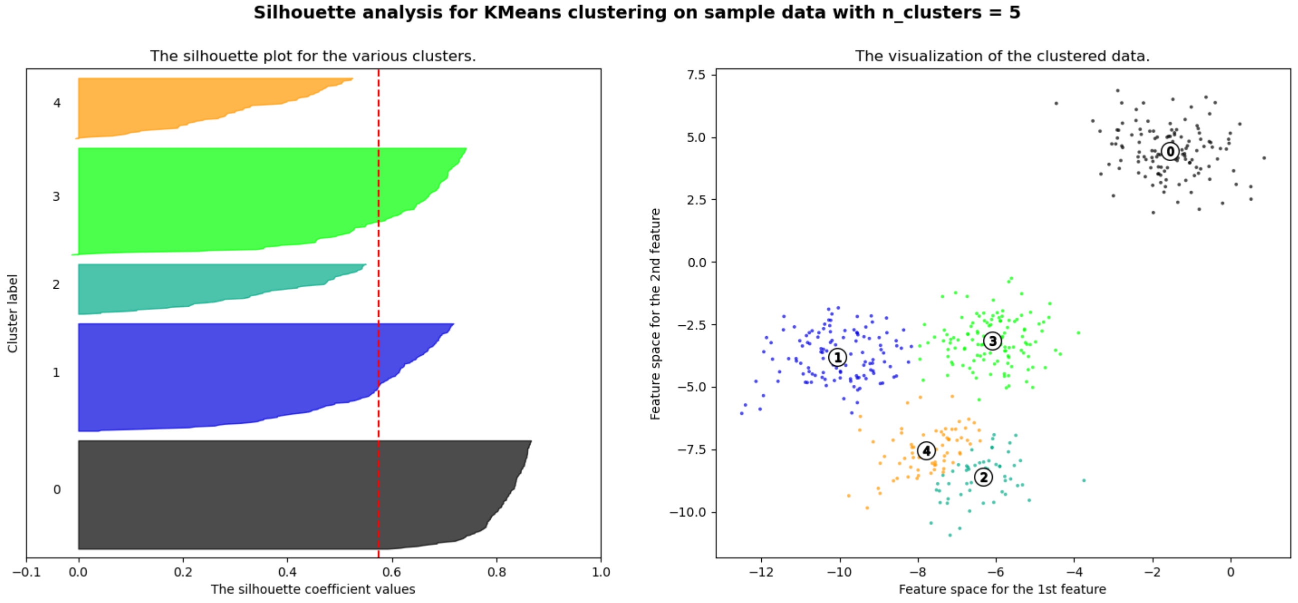
When k value is 3, one cluster has small silhouette value compared to other clusters and also has large cluster size. Size of cluster indicates the samples in that cluster. We expect clusters of almost similar sizes for good clustering. Small silhouette value indicates the lower compaction of the samples in that cluster. The Negative silhouette value shows the outlier samples in the data set.

Fig. 4 - Silhouette Plot with k value 4



When cluster count value is 4, all the cluster silhouettes were found to be greater than the threshold value 0.6. Almost all clusters are of similar size also. Hence the method suggests an optimal cluster count value of 4.

Fig. 5 - Silhouette Plot with k Value 5



Even though we got cluster count as 4, we can try for higher silhouettes. When cluster count value is 5, we got small silhouettes for two clusters and the cluster sizes are not similar. Two clusters have small silhouettes and their size is smaller than other two clusters. In case of cluster count 6, we got four clusters with small silhouette values. The sizes of four clusters are not similar to other two clusters. Similarly if we increase the cluster count 7, the results obtained are not appreciable. The results show that for the Abalone data set the optimal cluster count value 4.

5.3. Analysis using Combination of Average Silhouette and Individual Cluster Silhouette

The Silhouette analysis for optimal cluster count can be done in two ways. In multivariate Silhouettes the overall average of cluster Silhouettes is used to find the optimal cluster count. In single varied Silhouette index individual cluster Silhouettes is used to find the optimal cluster count. Three data sets have been chosen for the experiment. For a K value the silhouettes of each sample, average silhouettes of each cluster and overall average of total clusters are calculated. Table 3 gives the low, average and high silhouette values of different data sets for different cluster counts K.

In the case of IRIS data set, overall average cluster count value is high for k equals 2. But for one cluster average cluster silhouette is less than the threshold value. But in the case of k=3, both overall silhouettes and individual cluster silhouettes are greater than threshold. Hence optimum cluster count value for the IRIS data set is 3.

In the case of Abalone data set, the highest overall silhouette average is for cluster count 2. But here also average silhouette value for one cluster is less than threshold value. Both the overall average and individual cluster averages are greater than the threshold when cluster count is 4. Hence 4 is taken as the optimal cluster count. Similarly both the Silhouette measures are higher and greater than the threshold for sample data set Data1 for k=4.

The analysis of the clustering results shows that both average overall cluster Silhouettes and average individual cluster silhouettes can be considered for the selection of cluster count.

Table 2 - Average of Overall Silhouette and Individual Cluster Silhouette

Data set	Silhouette values											
	Clusters-2			Clusters-3			Clusters-4			Clusters-5		
	low	high	avg	low	high	avg	low	High	avg	low	high	avg
IRIS	.59	.75	.68	.6	.78	.65	.41	.63	.51	.39	.59	.49
Abalone	.50	.80	.68	.42	.80	.62	.61	.8	.65	.46	.8	.57
Data1	.58	.67	.64	.57	.61	.59	.61	.74	.69	.59	.69	.64

5.4. Evaluation Based on Accuracy of Clusters

Here for checking the correctness of the results obtained from the Elbow and the Silhouette methods, we use Weka the tool and the data sets from UCI. The partitioning algorithms used for the study are Kmeans and Farthest First. The clustering of data set can be performed based on different fields. The experimental result shows that more accurate and visible clustering can be observed when cluster using the class field. Ensemble the cluster algorithms using different k values and initial seed selection can improve the cluster accuracy [18].

5.4.1. Small k Value

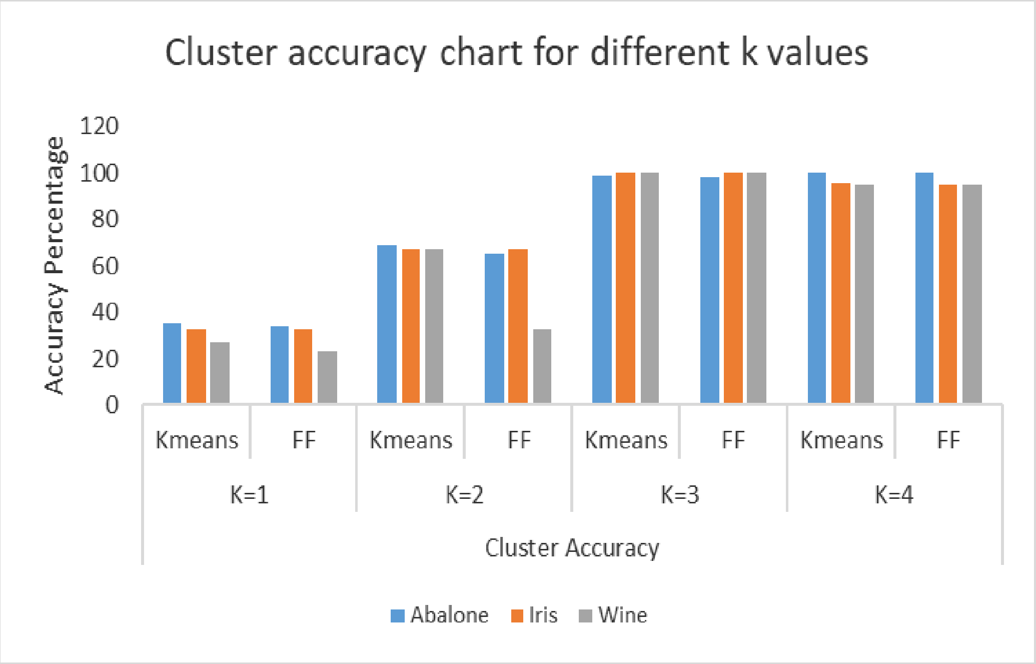
The clustering depends on many factors. Here we point out the importance of selecting the most appropriate value for k. To understand the importance of variation in the cluster formation, three data sets with small K value is selected. Cluster accuracy for k means and Farthest First is done separately. Data sets Abalone, Iris and Wine from UCI are used for the study.

Table 3 - Cluster Accuracy for K means and Farthest First

Data sets	Cluster Accuracy							
	K=1		K=2		K=3		K=4	
	Kmeans	FF	Kmeans	FF	Kmeans	FF	Kmeans	FF
Abalone	35	34	69	65	99	98	100	100
Iris	33	33	67	67	100	100	96	95
Wine	27	23	67	33	100	100	95	95

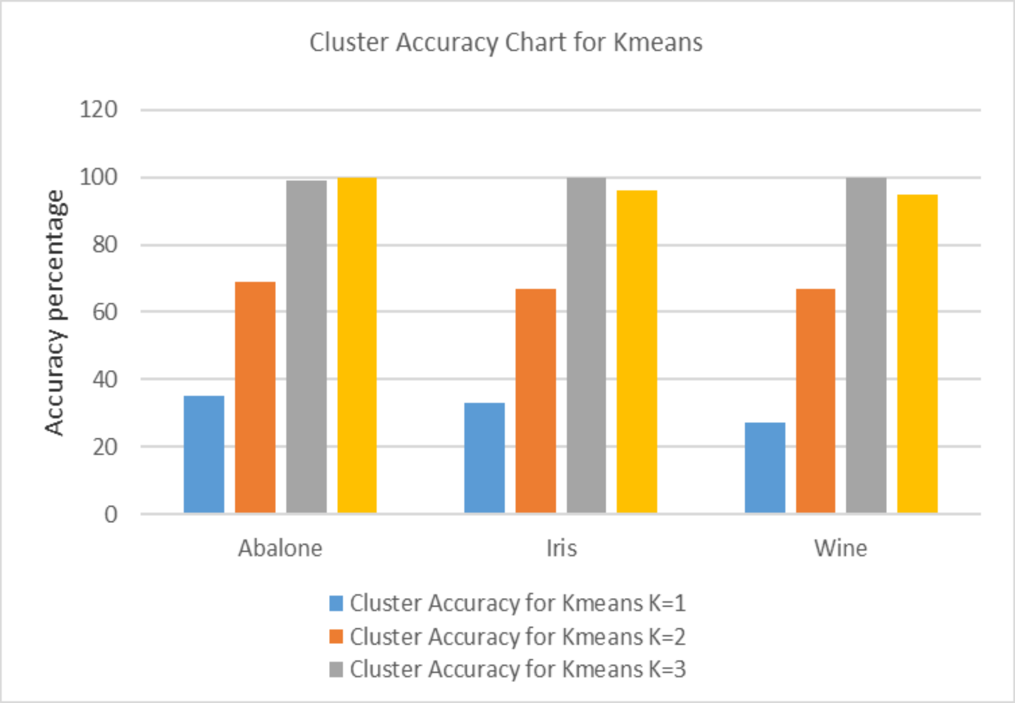
Table 3 shows almost similar results for K means and Farthest First algorithms. The value of cluster count in which the most accurate cluster is formed is taken as k value. The figure 6 gives the pictorial representation of cluster accuracy of the three data sets for different cluster count values.

Fig. 6 - Cluster Accuracy for different k values



The variations of accuracy of cluster using K means for the three data sets are shown in Figure 7. The cluster count value for around 100 percentage cluster accuracy is considered as the optimum value for K. Thus we obtain optimal cluster count for Abalone as 4, Iris as 3 and Wine as 3.

Fig. 7 - Cluster Accuracy Chart for Kmeans



5.4.2. Large k Value

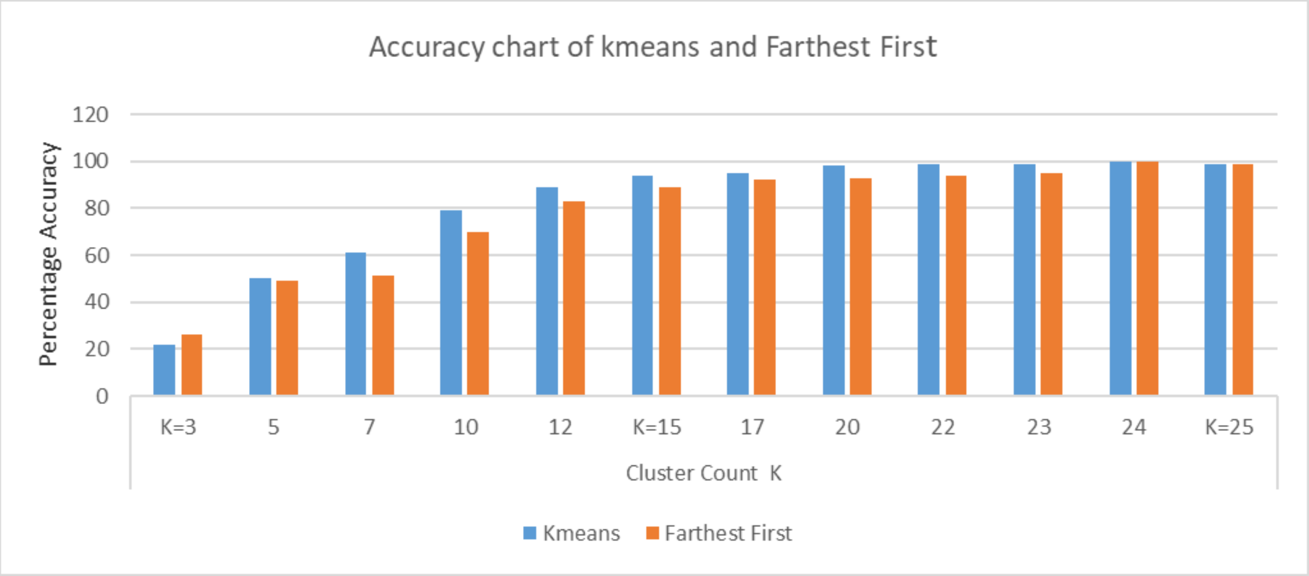
The accuracy variations of clustering based on the variations in k value was determined for a large cluster count value. Here also we used the class field for clustering. Table 4 shows the cluster accuracy of data sets when used K means and Farthest First algorithms. The subset of Abalone from the UCI repository is used for performance evaluation. The clustering was done on the data set based on the class field and obtained the optimum result for the cluster count as 24. There are 24 different classes in the selected subset of the data set. The class field with higher cluster count value was selected to understand the accuracy variations of cluster with different cluster count values. More accurate results can be observed using large data sets. The quality of cluster can be enhanced by solving the issues of outliers in the sample space [19].

Table 4 - Cluster Accuracy for K means and Farthest First

Algorithms	Cluster Count K											
	K=3	5	7	10	12	K=15	17	20	22	23	24	K=25
Kmeans	22	50	61	79	89	94	95	98	99	99	100	99
Farthest First	26	49	51	70	83	89	92	93	94	95	100	99

The figure 8 shows the cluster accuracy chart of the abalone data set. The maximum clustering accuracy is obtained for cluster count value 24. The errors occurred due to small change in cluster value is small in the case of large cluster count applications.

Fig. 8 - Cluster Accuracy with Varying Cluster Count



6. Conclusion

The value of cluster count is a very important feature in Kmeans clustering. The cluster accuracy of models depends on the cluster count value chosen. The Silhouette method and the Elbow method can be used to find the cluster count k value of the Kmeans algorithm. Three data sets from the UCI repository have been used for the study. In all the data sets, both the methods give the same results. More accurate results are obtained by the combination of separate cluster Silhouette and overall average of Silhouette instead of considering them separately. The Elbow and the Silhouette coefficients are effective tools to find the cluster count value and also to understand the compactness of samples in the cluster. Clustering accuracy can be improved by ensemble cluster count methods and the best initial seed selection models. Our future work is focused in ensemble cluster models to improve the clustering accuracy and reduce clustering error.

References

Shi Na, Liu Xumin and Guan Yong, "Research on k-means Clustering Algorithm: An Improved k-means clustering algorithm", *IEEE Third International Symposium on Intelligent Information Technology and Security Informatics,* 10.1109/IITSI.2010.74, 2010.

A. Fahad, N. Alshatri, Z. Tari and A. Alamri "A survey of clustering algorithms for Big Data", *IEEE Transactions on Emerging Topics in Computing, 2*(3), 267–790, 2014.

Ashish Jethi, Manishi Kalra, and Iranjan Bhattacharyya, "Analysis of Breast Cancer Recurrence using Combination of Data Mining Techniques" *IJCSMC,* 4(6), 392 – 397, 2015.

Pranav Nerurkar and Archana Shirke, "Empirical Analysis of Data Clustering algorithms", *Science Direct Procardia Computer Science, 125*(2018) 770–779 1877-0509 © 2018 http://www.journals.elsevier.com/procedia-computer-science

Gregory Wilkin, Xiuzhen and Huang, "K-Means Clustering Algorithms: Implementation and Comparison", *IEEE* Explore.ieee.org/4392591, 2016.

Christian Lopez and Scott Tuckerb, "An unsupervised machine learning method for discovering patient clusters based on genetic signatures", 29 July 20181532-0464/ © 2018 Elsevier Inc.

Baolin Yi Haiquan Qiao Fan Yang and Chenwei Xu, "An Improved Initialization Center Algorithm for K-Means Clustering", *IEEE* Ex*plore*: 30 December 2010.

Wang Yintong, Li Wanlong and Gao Rujia, "An improved k-means clustering algorithm" IEEE, World Automation congress, ISBN,978-1-889334-47-9, 2012.

Purnima Bholowalia and Arvind Kumar, "A Clustering Technique based on Elbow Method and K-Means in WSNE", *International Journal of Computer Applications © 2014 by IJCA Journal,* 105(9), 2014.

A. Rahman and B. Verma, "A Novel Layered Clustering based Approach for Generating Ensemble of Classifiers", *IEEE Transaction on Neural Networks,* 22(5), 781– 792, 2011.

Mengxing Huang and Hongjing Lin, “A New Method of K-Means Clustering Algorithm with Events Based on Variable Time Granularity”, *IEEE Explore, 13th Web Information Systems and Applications Conference (WISA),* 2016.

Caiquan Xiong, Zhen Hua, Ke Lv and Xuan Li, “An Improved K-means Text Clustering Algorithm by Optimizing Initial Cluster Centers”, *IEEE,7th International Conference on Cloud Computing and Big Data (CCBD)*, 2016.

Channamma Patil & Ishwar Baidari, “Estimating the Optimal Number of Clusters *k* in a Dataset Using Data Depth”, *Data Science and Engineering* volume 4, pages132–140. 2019.

Chunhui Yuan and Haitao Yang, “Research on K-Value Selection Method of K-Means Clustering Algorithm”, *Multi Disciplinary Scientific Journal, Graduate institute, Space Engineering University,* Beijing 101400, China; yuanyuan19821988@163.com, June 2019.

Congming Shi, Bingtao Wei and Shoulin Wei, “A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm”, *EURASIP Journal on Wireless Communications and Networking,* 2021, Article number: 31 (2021).

Tai Dinh, Tsutomu Fujinami and Van-Nam Huynh, *“Estimating the Optimal Number of Clusters in Categorical Data Clustering by Silhouette Coefficient”,* Book: Knowledge and Systems Science**s**, 1-17, DOI:10.1007/978-981-15-1209-4_1, November 2019.

A. Rahman and B.Verma, “A Novel Layered Clustering based Approach for Generating Ensemble of Classifiers”, *IEEE Transaction on Neural Networks,* vol 22, no 5, pp781– 792, 2011.

Brijesh Varma and Ashfaqur Rahman, “Cluster-Oriented Ensemble Classifier: Impact of Multicluster Characterization on Ensemble Classifier Learning”, *IEEE Transaction on Data and Knowledge Engineering,* 24(4), April 2012.

Min Wang, Zachary B. Abrams, Steven M. Kornblau & Kevin R. Coombes, “Determining the number of clusters while removing outliers”, *BMC Bioinformatics, 19*(1), 1 – 15, 2018. https://doi.org/10.1186/s12859-017-1998-9