Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.777129

A NOVEL STUDY OF SILHOUETTE METHOD TO SOLVE THE ISSUES OF OUTLIER AND IMPROVE THE QUALITY OF CLUSTER

Abdulnassar. A. A
Research Scholar, Computer Science and Engineering, School of Engineering CUSAT Cochin, India, Email: nasrishabaa@gmail. com

Latha . R . Nair
Associate Professor, Computer Science and Engineering, School of Engineering CUSAT Cochin, India, Email:latha5074@gmail. com

ABSTRACT-The silhouette method is a famous statistical method to find the cluster count value as well as to solve the issues of outliers in the sample space. An outlier is a data object that deviates significantly from the rest of objects. Silhouette coefficient value of a sample is a clear indication of the outlier in the data set. This study aims to improve the cluster quality by detecting and removing the outlier using different cluster methods. The value can be used to determine the compactness of formed clusters. In partition methods, the cluster results are very sensitive to the cluster count value we select. The performance of the Silhouette method is analysed with different data sets from UCI data repository. We propose two methods to detect and remove outliers. One method uses the silhouette value of sample and the other method measures the distances of sample with all cluster centroids and decide the sample as outliers based on a threshold distance. We have implemented methods in Python and the results are checked using different data sets from UCI and large public data sets. The performances of the cluster quality are checked using the cluster evaluation indexes such as Silhouette, Dunn, DB and C indexes. The removal of outliers improves the quality and compactness of the newly formed cluster. Analysis is done to study performance as well as cluster efficiency by removing the outlier from the sample space.

KEYWORDS Cluster Compactness, Data Mining, Dunn index, Kmeans. Outlier, Partition Algorithm and Silhouette.

1. Introduction

Data mining has a great role in the management of people's daily activities. The size of the data to be processed grows dramatically in many applications. The technique of extracting useful information from vast amounts of data is known as data mining. Machine learning methods are used to classify, cluster and order the data for various data mining application. The complexity and cost of machine learning procedures rise as data volumes grow. The primary machine learning tools used in data mining are clustering and classification. Samples are sorted into several groups or clusters through the clustering process.

The major cluster methods used are partition methods, Hierarchical methods, Density based and Grid based method. For many cluster applications, partition algorithms are the most suitable. They are very simple and easy to cluster. Performance of applications is influenced by initial cluster count selection and initial centroid selection. There are numerous established techniques that have been recommended for the proper selection of cluster count to raise cluster

quality. Silhouette and Elbow methods are popular statistical methods used for finding the cluster count. Small variation in selection of cluster count as well as initial seed can affect the cluster performance. Elbow method give the results based on the general behavior of the samples in data sets. Silhouette metric systematically measures the compactness of the samples in the cluster. Silhouette coefficient is a measure of how close the samples within cluster and how far from nearest cluster. In many applications the silhouette metric gives more accurate cluster count results.

Outliers are the data points stand out from the rest of the dataset. They are typically unexpected observations that skew the data distribution as a result of poor data input or inaccurate observations. Outliers badly affect the results of machine learning operations. The challenge of identifying data patterns that differ from other data patterns in some way is known as outlier detection. Such outliers can occur for a variety of reasons, including malicious behavior, instrument error, environmental change, human mistake, sampling error, experimental error, data processing error, etc. It is used in economical purpose such as detecting changes in stock market, observe changes in demand supply curve of a market and in medical science also relies upon outlier detection in some cases. The key factors that make the outlier identification complex include the challenges associated with accurately modeling both normal objects and outliers. Modeling an outlier data set becomes quite challenging. Outlier detection is an important research area in machine learning. Many techniques have been developed in this domain. These techniques can be broadly classified into three groups – supervised, semi-supervised and unsupervised methods.

II. Outliers

In data mining and machine learning, Outlier detection and correction are very important task to improve the quality of machine learning operation. These types of abnormal data patterns are referred to by a variety of terms, including outliers, anomalies, discordant observations, exceptions, faults, defects, aberrations, noise, errors, damage, surprise, novelty, oddities, or contamination in various application domains. Unexpected cluster results are produced when data mining and machine learning methods are applied to datasets with outliers. Outliers are different from noisy data. Outliers are classified as Global outliers, Contextual outlier and Collective clusters. Outlier detection models can be classified as Supervised, Semi supervised and unsupervised based on the analysis of sample of labeled data. The outlier detection methods based on assumption can be classified as Statistical methods, Proximity based methods and Cluster based methods.

A. Types of Outliers

Outliers are divided into three different types

Global or point outliers

Collective outliers

Contextual or conditional outliers

Global Outliers

Global outliers are also called point outliers. The simplest kind of outliers are considered to be global outliers. The term global outlier refers to a data point that differs from every other data point in a given data set. Almost all outlier detection techniques aim to identify global outliers.

Collective Outliers

A group of data points that deviates from the majority of the data in a given set is referred to as a collective outlier. In this case, the specific group of data objects might not behave as outliers, but the data objects as a whole might. To identify thistypes of outliers, you need to go through background information about the relationship between the behaviour of outliers shown by different data objects.

Contextual Outliers

Contextual outlier is used in applications based on some context. These are also known as Conditional outliers. These forms of outliers arise when a data object deviates from the other data points due to any particular condition in a given data set. Contextual and behavioural attributes are the two different sorts of characteristics that make up data objects. Users can investigate outliers in various contexts and circumstances with the help of contextual outlier analysis.

B) Outliers Classification

The process of identification of the outliers in a dataset is called outlier analysis. Outliers are discarded at many places when data mining is applied. But it is still used in many applications like fraud detection, medical, etc. Any unusual response that occurs due to medical treatment can be analyzed through outlier analysis in data mining. The major outlier analysis area is, Fraud detection in the telecom industry, Market Analysis, Fraud detection in banking and finance. It is also known as outlier mining.

Supervised methods

Data with accurate labels are a requirement for supervised methods. Label is used here to indicate both typical and anomalous data. Building prediction models for both the outlier and normal classes using the training data is the conventional strategy in such a situation. The two models are compared to each instance of unlabeled (test) data to determine which class it belongs to. Since supervised outlier identification techniques explicitly distinguish between normal and outlier behavior, precise models can be created. This effectively turns the outlier identification problem into a classification challenge when using the supervised outlier detection method. In many cases, labeling is done manually which is error-prone.

Semi-supervised methods

Semi-supervised algorithms rely on the availability of labelled examples for just one class. Collecting all the other classes frequently becomes extremely challenging. These methods typically train a model using the labelled data of a single class. Any test sample is examined using the trained model to establish the sample's class. The absence of labelled data is the fundamental disadvantage of semi-supervised algorithms.

Unsupervised methods

Unsupervised outlier identification techniques make the implicit assumption that data are properly grouped. The samples may be part of more than one cluster or they may be located on one cluster. The samples inside the clusters exhibit strong cohesiveness with considerable spacing between them. The outliers may be a small number of isolated points far from the data

clusters. Unsupervised learning's primary strategy is to identify clusters by looking at the data's immediate characteristics. Unsupervised methods are more frequently used in this field because labelling the data is not necessary in this situation. But only a small part of those studies has performed clustering and outlier detection simultaneously.

Algorithms for supervised and semi-supervised learning have some limitations because they need labelled datasets. The labelling mistakes are common for large datasets. Unsupervised learning techniques use data that has not been tagged. The computational complexity and training time of some supervised learning techniques are relatively high. They occasionally require computers with GPU support. unsupervised methods are comparatively less complex and less time consuming.

III. Literature Review

Cluster analysis is one of the promising areas used as pre arranging the data for various data mining applications. Suitability of the models for applications are determined using various data mining strategies. When we review the literature, we can see different types of cluster algorithms are used to detect the outliers from the data sets. The accuracy and quality of cluster depends the way we determine the outliers. The different strategies can be used to detect outliers from the data sets. Many researchers propose methods based on the distances of the samples from the centroids. The merits and demerits of the previous works are reviewed in this section. Many works had done to improve the deficiency of the partition cluster algorithms. Many improvements suggested to increase the quality of cluster using kmeans cluster models. Many papers discuss the importance of selecting proper cluster counts and initial seeds in cluster models.

An enhanced Kmeans cluster approach is proposed by skipping the distance calculations in each iteration [1]. To prevent needless repeating distance estimates, they suggest a data structure to hold the intermediate distances. The outcomes of earlier iterations can be kept and applied to the current method's subsequent iterations. Avoiding pointless calculations will increase the algorithm's efficiency. By lowering the number of repeated distance calculations performed during each iteration, the approach increases cluster speed. According to experimental findings, cluster accuracy, speed, and needless complex calculations have been improved.

A new method for selecting initial centre in kmeans algorithm is proposed and tried to minimize the sensitivity issue in selecting proper initial cluster centre[2]. The experiment is done using the data set from UCI and fixing the initial centres in previously identified dense area. Proposed method promises best initial seed selection in K means clustering. A brand-new ensemble model comprising Base classifier and Fusion classifier as its two phases are proposed [3]. By learning the boundaries of the clusters, the base classifier generates the cluster confidences. Combining the cluster confidences results in the decisions made by the fusion classifier. The work is completed using data from the UCI data repository. The two tailed sign tests are used to demonstrate how well the models perform.

A density based initial cluster centre selection method is studied [4]. The experimental findings indicate a more compacted cluster. The comparison uses data from the UCI repository. This technique can lower initial cluster centre noise levels and enhance cluster quality. The study of

different cluster algorithms and the outlier detection methods are done [5]. They conducted extensive experiments and compared the most representative algorithm from each of the categories using a large number of real (big) data sets. The effectiveness of the candidate clustering algorithms is measured through a number of internal and external validity metrics and scalability tests.

A method is proposed to eliminate isolated points from the data sets in order to determine appropriate initial cetres [6]. The experimental findings support the claim that choosing the initial cluster centre from denser data sets improved cluster results. The study is done to find to examine the significance of k value selection for the quick convergence of the k means algorithm [7]. They analysed the methods known as the Elbow method, Gap statics, and Canopy for the k value selection utilizing the iris data sets from the UCI. The study compares the benefits and drawbacks of these methods and find the silhouette method suitable to determine the right value for k.

A new method is created to detect and remove the outliers from data set using Kmeans and Hierarchical methods. [8]. To discover outliers from each clustering, they first applied the clustering algorithms K-Means and Hierarchical clustering to a data set. Outliers are found in K-Means clustering using both cluster-based method and a distance-based approach. Dendrograms are used in the case of hierarchical clustering to identify outliers. The goal of the project is to detect the outlier and remove the outliers to make the clustering more reliable.

A Kmeans algorithm based on the characteristics of Weibo event is proposed [9]. They suggest the K-means clustering algorithm of events based on variable time granularity. 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

A novel method known as Thresher is implemented in R package to predict optimum cluster count value using genetic data set[10]. The concepts from principal component analysis and outlier filtering are used. The Monte Carlo simulation research methodology is used to compare Thresher with other techniques for identifying outliers and the cluster count. Thresher is the best approach for predicting the ideal number of clusters when the number of objects being grouped is less than the number of variables used for clustering. Thresher also exhibited good sensitivity and specificity for detecting and removing outliers.

The performance analysis of different silhouette index based methods is done [11]. There are two approaches to calculate the index, and the first method makes use of the average of the average silhouettes of all clusters. The second, method is, by averaging the silhouettes throughout the entire dataset. These different indexing strategies have a considerable impact on determining the appropriate number of clusters in a data collection. Two well-known hierarchical methods, the complete-linkage and the single-linkage algorithms, were used to examine the index's performance as the underlying clustering techniques.

The K-means technique was used to cluster data and identify outliers concurrently [12]. Two parameters were needed: K and l, which stand for the desired number of clusters and top outliers, respectively. All distance metrics that can be stated as a Bregman divergence have been included in the authors' method. They have experimented with both made-up and actual

datasets. In addition, the authors suggested a modified K-means algorithm and offered a polynomial-time solution to the issue.

The sliding window-based ensemble method is proposed to detect outliers in a streaming fashion [13]. The proposed method uses a combination of clustering algorithms to construct subgroups (clusters) representing different data structures. These structures are later used in a one-class classification algorithm to identify the outliers.

The Outlier Detection and Clustering (ODC) algorithm, a variant of the K-means algorithm, was created to identify outliers [14]. A data point was designated as an outlier in the ODC method, which was if its distance from its cluster centre was at least p times the average distance. The authors' modified K-means algorithm to suit it for identifying outliers. Several publicly accessible datasets are used to demonstrate the effectiveness of the technique. The method has only taken a few pointless outlier cases for the study.

The outlier identification method using feature extraction by stacked encoders are proposed [15]. They used a combination of probabilistic neural networks to implement a majority voting system. The approach can handle situations when there are extremely few outliers in comparison to total data points. Due to the rapid expansion of the Internet of Things, researchers have also applied outlier identification algorithms in the healthcare sector (IoT). A survey was conducted to examine the technological perspectives of changing healthcare IoT systems and the action for enhancing Quality of Service with the use of contemporary technology [16]. Researchers applied density-based clustering techniques to aggregate the data's high-density regions.

A 2-step algorithm, called Clustering Based Outlier Detection (CBOD) is developed [17]. In its first phase, a clustering algorithm was applied to divide a dataset into hyper spheres with almost the same radius. In the second phase, outlier factors for all clusters obtained from the first step were calculated and the clusters were sorted based on their outlier factors. Clusters with high outlier factors were considered as outliers. The time complexity of CBOD is nearly linear with the size of dataset and the number of attributes, which results in good scalability and adapts to large dataset. The method can mislead if the data clusters vary a lot in terms of number of points.

A weighted majority voting method for clustering based on normalized mutual information (NMI)is proposed [18]. NMI is a supervised method where the true labels for a training set are required to calculate NMI. In this study, they extend their previous work of aggregating the clustering results to develop an unsupervised weighting function where a training set is not available. The proposed weighting function is based on Silhouette index, as an unsupervised criterion. As a result, a training set is not required to calculate Silhouette index. This makes their new method more sensible in terms of clustering concept.

A cluster-based ensemble technique is proposed to detect outliers [19]. They have used unsupervised learning algorithms since they are aware of the limitations of supervised and semi-supervised learning. Three clustering algorithms, K-means, K-means++, and fuzzy C-means, were employed in the cluster-based ensemble approach. In order to determine if a data point belongs to a certain cluster, their model mixes the findings from many clustering techniques intelligently and assigns probability to each data point. In order to maintain the flexibility of combining hard and soft clustering methods, they have devised a methodology to

assign a membership value to a data point in the case of hard clustering techniques. They have used five different cluster validity indices in their work to measure the goodness of the clusters formed, considering the results of eight widely used datasets for evaluation of the proposed model amongst which three are large datasets. It is noticed that a significant improvement in the cluster validity indices after applying our outlier detection algorithm. The experimental results prove that the proposed method is empirically sound.

A unique SO-GAAL method for outlier detection was put out, based on the mini-max game between a generator and a discriminator [20]. In a high-dimensional space, authors have dealt with data sparsity. On synthetic and real-world datasets, the approach is validated. The technique includes user-defined parameters. The approach is based on generative adversarial active learning (GAAL). Since many samples to train using this method, the training time is more.

A 3-phase modified K-means algorithm to cluster data and detect outliers is proposed [21]. In the first phase, fuzzy C-means algorithm was applied on the data. In the second phase, local outliers were identified and removed from the dataset. Hence, the cluster centers needed to be recalculated. In the third phase, certain clusters with low inter-cluster separation were merged and global outliers were identified. The algorithm has the capability of finding the local and global outliers separately. The method also supports the discovery of clusters of different density, shape, sizes and non-spherical shapes. The algorithm takes into account the density of a cluster while identifying local outliers. This might sometimes fail in case of scattered data. The final outcome is also sensitive towards the cluster centroid initialization.

A quantitative metric called Local Distance-based Outlier Factor (LDOF) to measure and detect an outlier in scattered datasets [22]. The technique outperforms K-nearest neighbor (KNN) and local outlier factor (LOF) for neighborhood detection. The Non-exhaustive Overlapping Kmeans (NEO-K-means) algorithm has developed to detect outliers during the clustering process. The authors have suggested a modified K-means algorithm that also manages overlapping clusters. They have put forth an objective function that unifies the problems of overlap and non-exhaustiveness. The procedure uses manually adjusted parameters. The outcome is greatly dependent on the right selection of parameter values. Wang et al. used the cluster algorithm OPTICS is used to identify outliers [23]. Outlier identification techniques based on graph theory were also employed in this work.

A brand-new statistical outlier detection technique that automatically located cluster centroids from the decision graph is proposed [24]. Density-based clustering has a significant flaw. In the case of density-based clustering, the clusters are chosen based on distances from the allocated centre. The density-based clustering is capable of handling the task if there are any discrete outliers. But there are many instances where outliers are a cluster in and of themselves. When this occurs, because there are so many nearby points, the outlier cluster will be treated as data. As a result, the efficiency of density-based clustering may substantially decline.

IV **Methodology**

One of the aims of this work is to analyse the performance of methods used to find the initial cluster count value. The popular methods such as Elbow and Silhouette coefficient are used to find the K value. The results obtained using these methods are checked with the results

of other methods. The Study also verify the variations of cluster efficiency 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 standard data sets from UCI data repository.

The main aim of this work is to design and develop the algorithms for outlier detection in clusters. There are different methods used to find the outliers. Two methods are introduced in this work. One is based on the Silhouette values of a sample in data space. These silhouettes are calculated using two measures. One is intra-cluster distances of the samples in a cluster and the other is the distances of samples with samples in the nearest cluster.

The other method suggested is to find the distances of a sample with the centroids of all the clusters in the set. The sample is fixed in cluster which has minimum distance with a centroid. If the probability measure is less than that of threshold then the sample is considered as outlier. The threshold value is a small value which is added with the (one by k th probability value)of the sample in cluster.

V. Results and Discussions

There are different works done in the area of prediction of cluster count value using silhouette method, different outlier detection methods, cluster evaluation indexes and the performance comparison works using different thresholds in both the silhouette method and probability distance measure method. The data set used to establish the results are also mentioned.

A Cluster Count

The cluster performance in partition cluster depends the cluster count we select. Hence before outlier detection, we have to find the cluster count values of different data sets. We are using different methods to fix the k value. The variations of well-known Elbow methods and Silhouette methods are used to find the k value of kmeans cluster. The cluster count value of the datasets used for outlier detection is calculated first. The standard data sets from UCI and some general data samples are used for the study. The silhouette method is mainly used to fix the cluster count in a data set.

1) Silhouette Method

This is an efficient method to find the cluster count of an application. The silhouette value is a measure of compactness of samples within cluster and the separation between other clusters. The performance of cluster depends the low inter cluster distance among samples and the high intra cluster distance among samples in one cluster to the neighboring cluster. The silhouette plot shows the measure of how close each point in one cluster and how far is to points in the neighboring clusters. The method can be used to find the cluster count value. Silhouette value ranges between -1 and 1.

The Silhouette 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 that the assigned cluster is wrong.

The silhouette values are found using different data sets from the UCI and public applications. The results of the silhouette values for the Abalone data sets using different K values are analysed here. The method suggests to select the k values of higher cluster silhouettes and also the similar cluster sizes.

The silhouette analysis for cluster count can be done in two ways. One is finding the overall average of cluster silhouettes known as multivariate silhouettes and the other is considering the individual cluster silhouettes or single varied silhouette index. Here for the analysis, we use three data sets and find the silhouettes of each sample and average silhouettes of each cluster and also overall average of total clusters for a k value.

In case of IRIS data set, overall average cluster count value is high for k equals 2. but for one average cluster silhouette is less than threshold value, hence cannot consider this value as cluster count. But in case of k=3,both overall silhouettes value is high and also all the cluster silhouettes are greater than threshold. Hence optimum cluster count value for the IRIS data set is 3.

In case of Abalone data set, when the cluster count value is 4, both the overall average and individual cluster averages are greater than threshold. Hence, we select 4 as cluster count. Similarly, both the silhouette measures are higher and greater than threshold for data set Covid-19 for k=4. The result of the analysis is clear that both overall cluster average silhouettes and individual cluster average silhouettes are considered for the selection of cluster count value.

Table1. Silhouettevalues for different cluster count values

Data set	Silhouette value								
	Clusters-2			Clusters-3			Clusters-4		
	low	high	avg	low	high	avg	low	High	avg
IRIS	. 59	. 75	. 68	. 6	. 78	. 65	. 41	. 63	. 51
Abalone	. 50	. 80	. 68	. 42	. 80	. 62	. 61	. 8	. 65
Covid19	. 51	. 55	. 52	. 52	. 56	. 54	. 53	. 62	. 57

The detailed analysis of the Abalone data set is given below. When k value is 3, one cluster has small silhouette compares to other and also have large cluster size. Small silhouette value indicates the lower compaction of the samples in that cluster.

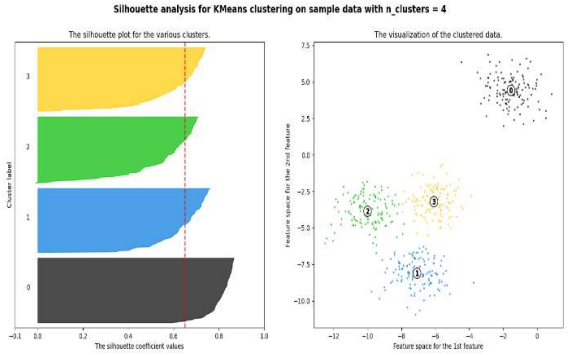


Fig. 1 Silhouette Plot with k value 4

When cluster count value is 4, Almost all the cluster silhouettes values are greater than 0. 6 and all the clusters are almost same size. Hence the method suggests the cluster count value 4. 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. Hence, we cannot recommend 5 as cluster count value.

B. Datasets

We used different types of data sets to detect the outliers from the data sets. The datasets used are IRIS, Abalone, Covid_19_clean_complete, Yeast, and Shuttle.

1) The IRIS data set consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters.
2) Abalone data set consists of 4177 instances of data with eight attributes.
3) Covid19_clean_complete is data set used to cluster the patient based on the different symptoms. There is total 49069 instances and 15 attributes.
4) Yeast dataset consists of a protein-protein interaction network. It is multivariate data set with eight attributes and 1484 instances
5) ETF data sets (large data sets)

B. Cluster Validation Techniques

Cluster validation can be categorized into two classes, external clustering validation and internal clustering validation. All these clustering assessment techniques fall under two categories, supervised evaluation that uses an external criterion and unsupervised evaluation that uses an internal criterion.

The unsupervised evaluation metrics generally considers intra-cluster and inter-cluster distance objectives of a cluster outcome. The sum of squared distance between each point and the centroid of the cluster it is assigned, is a measure to compute clustering quality. Dunn Index, Rand Index, Purity, Sum of Square Distance (SSD), and Average Silhouette Coefficient, Cindex, Calinski-Harabasz index, DB index, are widely used clustering evaluation metrics.

1) Dunn index

The Dunn index (DI) is a metric for evaluating clustering algorithms and it is introduced by J. C. Dunn in 1974. The higher Index value indicates better cluster performance. It is calculated as the lowest inter cluster distance (ie. the smallest distance between any two cluster centroids) divided by the highest intra cluster distance.

The Dunn index for c number of clusters is defined as :

$$\text{Dunn index} = \min_{1\le i\le c}\{\min_{1\le j\le c, j\ne i}\{\frac{\delta(s_i, s_j)}{\max_{1\le j\le c}\{\Delta(s_k)}\}\}$$

$\delta(S_i, S_j)$ is the inter cluster distance

$\Delta(S_k)$ is the intra cluster distance of cluster S_k.

2) Calinski-Harabasz index

The CH Index is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation) . This index also known as the Variance Ratio Criterion. It is calculated as the ratio between the sum of between-clusters dispersion and inter-cluster dispersion for all clusters. The high score indicates the better performance. For a set of data S of size nis clustered into k clusters, the Calinski-Harbaasz score s is defined as the ratio of the between cluster dispersion mean and the within cluster dispersion.

$$s = \frac{tr(B_k))}{tr(W_k)} * \frac{n-k}{k-1}$$

where tr (B_k) is the trace between group dispersion and tr(W_k) is the trace of the within cluster dispersion. The score is higher when clusters are dense and well separated. The Calinski-Harabasz index is generally higher for convex clusters than other type of clusters.

3. Silhouette Coefficient

The Silhouette Coefficient is defined for each sample and is composed of two scores a(i) and b(i). The silhouette Coefficient for a single sample is given by

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

Here a(i) is the mean distance between a sample and all other points in the same cluster. This score measures the closeness of points in the same cluster and b(i) is the mean distance between a sample and all other points in the next nearest cluster [18]. This score measures the distance of points in different clusters. The silhouette Coefficient for a set of samples is given by the mean of the Silhouette Coefficient for each sample. The score is bounded between -1 for incorrect clustering and +1 for highly dense clustering. Scores around zero indicate overlapping clusters. The Silhouette Coefficient is generally higher for convex clusters and the computational complexity is high, ie, $O(n^2)$

4) Davies-Bouldin Index

This index is a measure of the average similarity between clusters. A lower Davies-Bouldin index relates to a model with better separation between the clusters. The computation of Davies-Bouldin is simpler than that of Silhouette scores. The index is computed only quantities and features inherent to the dataset. The usage of centroid distance limits the distance metric to Euclidean space. The Davies-Boulding index is generally higher for convex clusters than other types of clusters.

$\text{DB index(U)} = {}^{1}/_{k} \sum_{i=1}^{k} \max_{i \neq j} \{ \frac{\Delta(s_i)+\Delta(s_j)}{\delta(s_i,s_j)} \}$, where

$\delta\ (s_{i,}\ s_j)$ is the inter cluster distance

$\Delta(s_k)$ is the intra cluster distance of cluster $S_{k.}$

5) C index

The C Index was formulated by Hubert & Levin in 1976. Its purpose is to compare the dispersion of clusters of data relative to the total dispersion in a dataset. Ideally, the value of the number of clusters that minimizes the C Index will also be the optimal number of clusters to partition a dataset. Its purpose is to compare the dispersion of clusters of data relative to the total dispersion in a dataset. Ideally, the value of the number of clusters that minimizes the C Index will also be the optimal number of clusters to partition a dataset.

The C Index is calculated as $\text{Cindex} = \frac{S_w - S_{min}}{S_{max} - S_{min}}$

Here, S_w is the sum of within-cluster distance measurements (only intra cluster combinations of data are summed within each cluster - not between clusters) S_{min} is the sum of the m smallest point-wise distances between points within the entire dataset Smax is the sum of the m largest point-wise distances between points within the entire dataset. m is the total number of pairs of observations belonging to the same cluster. It is the same as total combinations of points within clusters taken two at a time.

6). External Validation Indexes

Random Index, Bayesian information criterion, Fowlkes-Mallows Score, V index and Purity of cluster are some of the external cluster evaluation indexes used to assess the performance of generated cluster [20]. Random index is a way to compare the similarity between two different clustering methods. It is calculated as$R = \frac{(a+b)}{(\mathbf{nC_2})}$. The Adjusted Rand score is introduced to determine whether two cluster results are like each other. The Bayesian information criterion (BIC) is normally used to avoid overfitting of samples. It is defined as, $BIC = -\ln(L) + v\ln(n)$, Where n is the number of objects, L is the likelihood of the parameters to generate the data in the model, and v is the number of free parameters in the Gaussian model. A clustering result satisfies homogeneity if all its clusters contain only data points which are members of a single class. The Fowlkes-Mallows Score is an evaluation metric to evaluate the similarity among clusters obtained after applying different clustering algorithms. The V-Measure is defined as the harmonic mean of homogeneity and completeness. Purity in cluster is an external evaluation criterion of cluster quality. It is the percent of the total number of samples that were classified correctly. It's value range between 0 and 1.

D. Outlier detection Methods

In this work, we are using two different methods to detect the outliers from the data sets one is using the silhouette method and the other using the distance method, which is calculated using the probability of distance of sample with centroids of the clusters formed.

1) Silhouette method to detect outlier

In this method Silhouette of each sample is calculated using the formula

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

Here we can get the silhouettes of all the samples in the data sets. Silhouette value ranges between -1 and 1. Negative value indicates wrong inclusion of the sample in the cluster. The value 0 indicates that the sample are in the boarders of cluster. Very small values near to zero and border samples are considered as outliers. The detected the outliers can be removed from the cluster and can get better cluster results. Based on the importance and the efficiency required in applications we can fix thresholds to detect and remove the outliers from the clusters.

2) Distance method to detect outlier

The other ensemble method used to detect the outlier from the cluster is Distance Method. Here firstly, the probability of a sample to belong to all clusters is calculated using the distances of the sample with different centroids of the clusters. We are using a dataset with N samples. The samples are of d dimension. The data samples can be denoted as S_1, S_2, …. , S_n. Let us assume the cluster count as K. The cluster centres are denoted by C_1, C_2, …, C_k. After hard clustering, data object S_i belongs to j^{th} cluster with center C_j. We calculates probabilities of Si to belong to clusters 1, 2, . . , K by the given equation.

$$P[i][j] = \frac{\exp(-Eudi\ (S_i, c_j))}{\sum \exp(-Eudi\ (S_i, C_k))}\ , j = 1,2, \dots, k \qquad (1)$$

where Eudist (S_i, C_k) is the Euclidian distance between sample S_i and C_k .

The probabilities of the samples in each cluster can be used to determine the position of the samples. P[i][j] gives the probability of data object Xi to belong to J^{th}clusterwith centre C_j. The data object Xi is assigned to J^{th} cluster if P[i][j] is maximum among P[i][1], . . . P[i][k]

We decide the Xi as outlier if $\max_j(P[i][j]) < (1/k) + T$, where T is the threshold value

In soft clustering,the probability distance value is checked using the formula to decide the sample is outlier.

The performance of the resultant cluster formed after removing the outliers can be analysed using the different cluster performance analysis indexes. The improvement in the cluster evaluation indexes is the clear indication of the cluster performance.

E. Experimental Results

We conducted the experiment using 7 different data sets. The large data sets ETF and small dataset IRIS from UCI for comparing the cluster results. Cluster evaluation metrics Silhouette, CH index, DB index, C index and Dunn Index of the clustered results are given in different tables.

The higher value score of Dunn Index indicates better cluster results. In almost all the datasets Ensemble models gives better cluster results.

The silhouette score is the clear indication of the compactness of the samples in the cluster and differences of samples in different clusters. The high silhouette value indicates the better cluster quality. The value ranges between 0 and 1. The ensemble models gives better performance. The boosted ensemble model gives better silhouette results.

Calinski-Harabasz index is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The higher CH score is the indication of better cluster quality. In our work ensemble models have higher CH score.

Davies-Bouldin Index is a measure of the average similarity between clusters. A lower Davies-Bouldin index relates to a model with better separation between the clusters. Ensemble models have lower DB index value.

The C Index compare the dispersion of clusters of data relative to the total dispersion in a dataset. The low value indicates better cluster performance. The boosted ensemble model gives lower C index value.

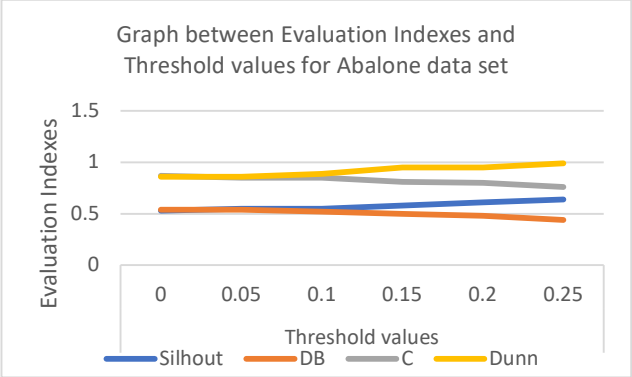
In this research work, the outlier detection is done by mainly two methods. Both the methods give better results. The performance of the algorithms is checked by removing the outliers from the set. We check the cluster evaluation indexes before removing the outliers. Then we remove the outliers and again the values of evaluation indexes are checked. The values of indexes indicate the performances of the resultant clusters when remove the outliers from data set. In both methods, the cluster evaluation results show better cluster results. The assembling of the cluster algorithm can also bring better cluster results.

We have conducted the experiment of detecting the outliers using small and large data sets. We use different thresholds for the outliers. Almost all results shows that the detection of outliers are done very accurately. The index values are checked to analyse the performance of the method. Improved results of indexes by Abalone data set using the Distance methodare given in table2.

Table2. Improvements of indexes after outlier removal in Abalone

Indexes	Threshold values					
	0	. 05	. 1	. 15	. 2	. 25
Silhout	. 53	. 55	. 55	. 58	. 61	. 64
DB	. 54	. 54	. 52	. 50	. 48	. 44
C	. 87	. 85	. 85	. 81	. 8	. 76
Dunn	. 86	. 86	. 89	. 95	. 95	. 99

The cluster validation indexes of the Abalone data sets for different threshold valueare given in table2. When the threshold value increases, more outliers are removed and the cluster performance are increased. The improved cluster evaluation index is the indication of the cluster quality. Figure 2 is the performance improvement of cluster evaluation indexes in Abalone data set. Here Silhouette value and Dunn index values are increases when threshold value increase. DB index and Cindex decreases with increase of threshold.



Figur2. Graph of improvement in indexes after outlier removal in Abalone

2) Distance method

We conduct the experiment by changing the threshold values for different data sets. When we analyse the results, we can see that in almost all the data sets the index values improved when removing more outliers from the data sets. The improvement silhouettes are given in table 3.

Table3 Silhouette improvement in Distance method

Data sets						
	0	. 05	. 1	. 15	. 2	. 25
	Silhouette values					
Abalone	. 53	. 55	. 55	. 58	. 61	. 64
IRIS	. 43	. 46	. 46	. 5	. 53	. 55
Covid19	. 55	. 57	. 59	. 62	. 62	. 65
Yeast	. 3	. 3	. 34	. 37	. 39	. 41
ETF	. 48	. 49	. 52	. 54	. 56	. 59

The graphical representation of the silhouette improvement for different data sets are shown in figure3. The improved values for higher threshold values are clear indication of the cluster quality after removing outliers from the data set.

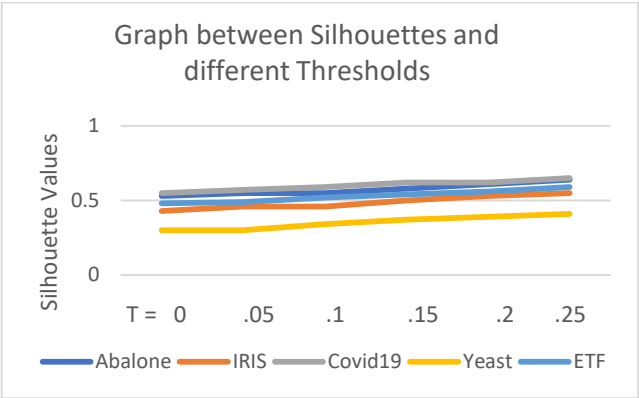


Figure3 Graph of Silhouette improvements in different data sets

The improvements in Dunn index value is given in table4 and the graphical representation of the improvement in Dunn index is given in Figue4. Here we can see that the Dunn index vale is improved when detecting more outliers using higher threshold value.

Table3. Dunn index improvement in Distance method

Data sets	Threshold values					
	0	. 05	. 1	. 15	. 2	. 25
	Dunn Index values					
Abalone	. 86	. 86	. 89	. 95	. 95	. 99
IRIS	. 73	. 73	. 79	. 83	. 83	. 85
Covid19	. 98	. 98	1. 1	1. 2	1. 2	1. 3
Yeast	1. 2	1. 3	1. 3	1. 5	1. 5	1. 6
ETF	. 78	. 79	. 83	. 87	. 89	. 93

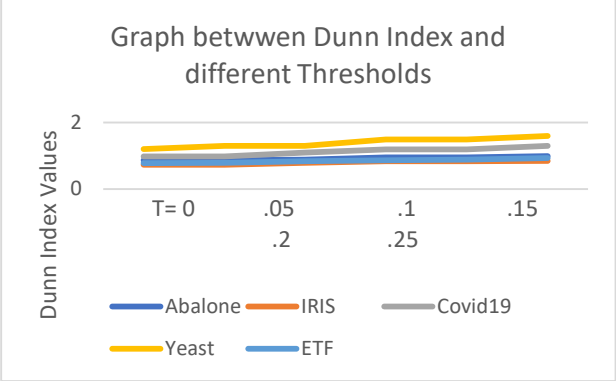


Figure4. Graph of Dunn index improvement in Distance method

3) Silhouette method

Here we use 5 different type of data sets to analyse the performance of Silhouette method for outlier detection. In almost all data sets the better detection of the outliers possible . In case of negative silhouette and 0zero indexes very small number of outliers are removed. When threshold is increased more outliers are removed. The performance of this method analysed using different data sets. The improvement of Silhouette index is given in table 5. Figure 5 gives the graphical representation of improvements in silhouette value for different datasets.

Table5 Silhouette improvements in different data sets using Silhouette Method

Data sets	Threshold values					
	-ve	0	. 05	. 1	. 15	. 2
	Silhouette values					
Abalone	. 53	. 55	. 57	. 59	. 62	. 63
IRIS	. 43	. 45	. 46	. 49	. 52	. 55
Covid19	. 55	. 57	. 59	. 61	. 64	. 65
Yeast	. 31	. 33	. 35	. 37	. 39	. 42
ETF	. 49	. 52	. 54	. 56	. 58	. 6

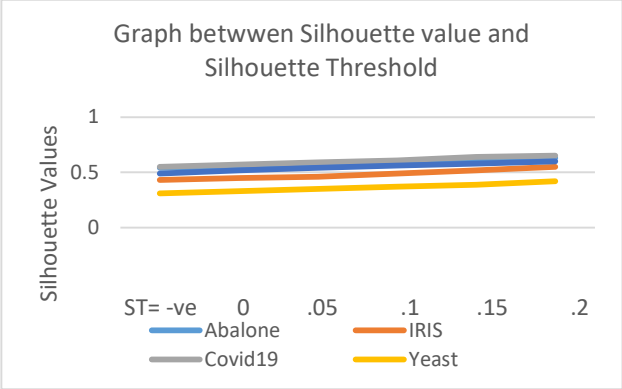


Figure5 Silhouette improvements using Silhouette method

Table 6. Dunn index improvements using Silhouette Method

Data sets	Silhouette Threshold Values					
	-ve	0	.05	.1	.15	.2
	Dunn Index values					
Abalone	.86	.89	.91	.94	.96	.99
IRIS	.73	.75	.77	.81	.83	.85
Covid19	.98	.99	1.1	1.2	1.3	1.4
Yeast	1.2	1.3	1.4	1.5	1.6	1.7
ETF	.78	.8	.82	.86	.89	.92

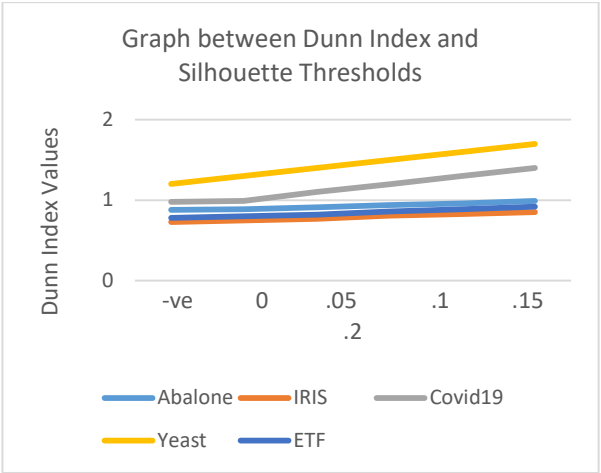


Figure 6. Dunn index improvement graph using Silhouette Method

The Dunn index improvements of different data sets using Silhouette outlier detection method is given in table6. The graphical representation of the Dunn index improvement is

shown in Figure6. The Dunn index value is improved for all data sets. The improved value of the index is clear indication of performance of Silhouette index outlier detection method.

VII. Conclusion.

This paper discusses different outlier detection methods to find the outliers from the data sets. Both the methods convincingly remove the outliers. The limit of the distances of the outliers from the centroids is calculated using different threshold values. The quality of the cluster after removing the outliers are analyzed using different cluster evaluation tools. The different internal and external cluster evaluation indexes such as Silhouette index, Dunn Index, Calinski Harbasz Index, Davies Bouldin Index and C are used for the study. The work revealed that both the silhouette method and Distance method remove the outliers efficiently. Almost the same results are obtained for the larger data sets as well as small data sets. Proper cluster count selection as well as the ensemble cluster can bring more accurate results.

References

[1] RShi Na ; Liu Xumin ; 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,2010,10. 1109/IITSI. 2010. 74

[2] Baolin Yi ; Haiquan Qiao ; Fan Yang ; Chenwei Xu,"An Improved Initialization Center Algorithm for K- Means Clustering", IEEE Xplore: 30 December 2010

[3] 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, pp 781– 792, 2011.

[4] Wang Yintong ; Li Wanlong ; Gao Rujia,"An improvedk-means clustering algorithm,IEEE,World Automation congress 2012:ISBN,978-1-889334-47-9

[5] Balamurugan, K. , Latchoumi, T. P. , & Ezhilarasi, T. P. (2022). Wearables to Improve Efficiency, Productivity, and Safety of Operations. In Smart Manufacturing Technologies for Industry 4. 0 (pp. 75-90). CRC Press.

[6] Caiquan Xiong ; Zhen Hua ; Ke Lv ; Xuan Li,"An Improved K-means Text Clustering Algorithm by Optimizing Initial Cluster Centers,IEEE2016 ,7th International Conference on Cloud Computing and Big Data (CCBD)

[7] 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;

[8] Anwesha Barai (Deb), Lopamudra Dey,"Outlier Detection and Removal Algorithm in K-Means and Hierarchical Clustering World Journal of Computer Application and Technology ', 5(2): 24-29, 2017 http://www. hrpub. org DOI: 10. 13189/wjcat. 2017

[9] Mengxing Huang ; Hongjing Lin,"A New Method of K-Means Clustering Algorithm with Events Based on Variable Time Granularity",IEEE Explore,: 2016 13th Web Information Systems and Applications Conference)

[10] Min Wang , Zachary B Abrams , Steven M Kornblau , Kevin R Coombes ," Thresher: determining the number of clusters while removing outliers", BMC Bioinformatics,. 2018 Jan 8;19(1):9. doi: 10. 1186/s12859-017-1998-9.

[11] Artur Starczewski & Adam Krzyżak , "Performance Evaluation of the Silhouette Index", International Conference on Artificial Intelligence and Soft Computing, ICAISC 2015: Artificial Intelligence and Soft Computing pp 49–58, Part of the Lecture Notes in Computer Science book series (LNAI,volume 9120

[12] Chawla S, Gionisy A (2013) κ-means-: A unified approach to clustering and outlier detection. Proceedings of the (2013) SIAM International Conference on Data Mining, SDM 2013, pp 189–197

[13] Nadeem Iftikhara,∗ , Thorkil Baattrup-Andersenb, Finn Ebertsen Nordbjerga , Karsten Jeppesena," Outlier Detection in Sensor Data using Ensemble Learning", Elsevier B. VProcedia Computer Science 176 (2020) 1160–1169 1877-0509 © 2020

[14] Ahmed M, Mahmood AN (2013) A novel approach for outlier detection and clustering improvement," In Proceedingsofthe (2013)IEEE8thConferenceonIndustrialElectronicsand Applications, ICIEA 2013, pp 577–582

[15] Chakraborty D, Narayanan V, GhoshA (2019) Integration ofdeep feature extraction and ensemble learning for outlier detection. Pattern Recognition 89:161–171

[16] Qadri YA, Nauman A, Bin Zikria Y, Vasilakos AV, Kim SW (2020) The future of healthcare internet of things: a survey of emerging technologies. IEEE Commun Surv Tutorials

17] Jiang SY, An QB (2008) Clustering-based outlier detection method. In: Proceedings – 5th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2008, vol 2, pp 429–433

18] Meshal Shutaywi and Nezamoddin N. Kachouie , " Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering", Entropy2021,23(6),759;https://doi. org/10. 3390/e2306075

[19] Akash Saha,Agneet Chatterjeeand Soulib Ghosh, "An ensemble approach to outlier detection using some conventional clustering algorithms", https://link. springer. com/article/10. 1007%2Fs11042-020-09628-5.

[20] Kumar, B. P. , & Latchoumi, T. P. (2022). Dataset identification for prediction of heart diseases. International Journal of Cloud Computing, 11(5-6), 415-424.

[21] Balamurugan, K. , Latchoumi, T. P. , & Satla, S. Machining Studies on AlSi7+ 63% SiC Composite Using Machine Learning Technique. In Metal Matrix Composites (pp. 139-166). CRC Press.

[22] Balamurugan, K. , Latchoumi, T. P. , Deepthi, T. , & Ramakrishna, M. Optimization Studies on Al/LaPO4 Composite Using Grey Relational Analysis. In Metal Matrix Composites (pp. 29-48). CRC Press.

[23] Wang YF, Jiong Y, Su GP, Qian YR (2019) A new outlier detection method based on OPTICS. Sustain Cities Soc 45:197–212.

[24] Yan H, Wang L, Lu Y (2019) Identifying cluster centroids from decision graph automatically using a statistical outlier detection method. Neurocomputing 329:348–358.