IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Revue d'Intelligence Artificielle
Vol. 38, No. 4, August, 2024, pp. 1183-1192
Journal homepage: http://iieta.org/journals/ria

A Comprehensive Study of Ensemble Models to Improve the Performance of Cluster Algorithms

Abdul Nassar Ayyaril Abdulla*, Latha Ravindran Nair

Computer Science and Engineering, School of Engineering CUSAT Cochin, Cochin 682022, India

Corresponding Author Email: nasrishabaa@gmail.com

https://doi.org/10.18280/ria.380412

Received: 3 October 2023
Revised: 20 December 2023
Accepted: 28 January 2024
Available online: 23 August 2024

Keywords:
cluster, data mining, ensemble, Kmeans, silhouette and Dunn index

ABSTRACT

The study analyzed the individual performance of partition cluster algorithms and selected Kmeans, Kmeans^{9+}, Kmedoid, and Fuzzy Cmeans algorithms as base algorithms for the ensemble. The cluster performance is assessed using UCI data sets as well as other common public data sets. The quality of cluster results depends on the base cluster algorithm used. The efficiency of base algorithms is added based on the ensemble models. We developed two ensemble models: a simple hard voting ensemble and a soft boosting ensemble based on the bagging and boosting ensemble technique. Ensemble of different cluster algorithms can generate the most accurate clusters. Both models show better cluster results than their base cluster algorithms for the small and big data sets. When using most data sets, the Soft Boosting Ensemble model achieves 100% cluster accuracy. The cluster evaluating functions are the benchmark for assessing the quality of the cluster. All the cluster evaluating indices show better performance for developed ensemble models. Internal cluster-evaluating indices as well as external cluster-evaluating indices are used to compare the cluster quality of the individual cluster algorithm and generated ensemble cluster models. The work establishes that the developed ensemble methods improved the quality of the generated clusters.

1. INTRODUCTION

Data has an immense role in the day-to-day events of the society. The importance of data mining activities has become very important. The process of examining hidden patterns in data from multiple angles to classify them into useful information is known as data mining. In many applications, the size of the data to process is very large, and handling these data with conventional algorithms becomes difficult. The growth of the data leads to an increase in the cost and complexity of the machine learning operations. In data mining, the two main machine-learning techniques employed are classification and clustering [1].

One of the most popular methods for revealing the data's hidden structure is clustering. The process of finding subgroups in the data so that data points in the same subgroup are highly similar and data points within various subgroups are highly different can be used to characterize it. When there are homogeneous subgroups, every cluster's samples are as similar as possible based on a similarity metric like correlation or Euclidean distance. By dividing samples into feature-based subgroups, cluster analysis can be performed. It attempts to maximize the distance between inter-cluster data points and minimize the similarity between intra-cluster data points.

It is an unsupervised learning problem to analyze clusters [2]. To discover interesting patterns in data, e.g. groups of customers based on their actions, it has been frequently employed as an analytical tool. In the feature space of input data, it is an unsupervised problem to find natural groups. The dataset is divided into a set of k groups by the Partition Cluster method, where k is the total number of pre-defined groups. Another name for it is the centroid-based approach [3]. The most used partitioning clustering method is the K-means Clustering algorithm.

An ensemble in machine learning technique is generally termed as the system of working individual models in parallel to attain a common solution to the problem. In the context of supervised learning, ensemble methods were effectively implemented to improve the classification stability and accuracy techniques. To improve quality and consistency over the output of individual clustering algorithms, a clustering ensemble combines multiple clustering models [4]. Ensemble models are widely used in unsupervised learning applications. The success of ensemble applications in the classification field leads the researchers to apply the same in the cluster field.

The resultant cluster obtained using different cluster algorithms may be different. In many of the partition models, the cluster count value, initial centroid selection, and the algorithm itself can reduce the quality of the resultant cluster. The major issue in an unsupervised model is that there was no prior information on the underlying structure of the data. The cluster analysis helps us to select the most suited and accurate algorithms for the ensemble. Bagging and boosting are well-known Ensemble techniques used in many applications.

Majority of the researchers used simple methods such as averaging to ensemble cluster algorithms. The bagging and

boosting ensemble approaches are very simple and effective to develop new ensemble models. The boosting models give more appropriate cluster results. One of the crucial issues that is needed to make cluster models a success is validation. There are two types of cluster validation: internal and external clustering validation. The cluster quality can be measured by checking the compactness of the samples in the cluster.

There are various functions and indices with which we can evaluate the performance of clustering algorithms. The cluster quality can be measured, and different cluster results can be compared using functions such as the Rand Index, Fowlkes-Mallows Score, Purity, and Sum of Square Distance (SSD).

2. ENSEMBLES

Cluster and classification accuracy can be improved using the ensemble technique. The classification ensemble is made up of using different base classifiers. The class label prediction done by the ensemble method is more accurate than their base classifier component. In the case of the ensemble cluster, the composite cluster gives a more efficient cluster than the base cluster models. The imbalance of data in classification applications and outlier issues in cluster models can affect the performance of ensemble models. Ensemble methods can be done parallel by allotting different base classifiers to separate processors.

M base algorithms are used in the ensemble model as given in Figure 1. Each base algorithm gives its cluster outputs. Each cluster results are ensembled using the ensemble cluster model to get the resultant cluster.

Ensemble models are formed by combining base models using a single base learning algorithm known as a homogeneous ensemble model. It can also be created as multiple base learning algorithms for each model known as a heterogeneous ensemble model. Ensemble models are more efficient than any base models. The prediction reliability of the ensemble model is very high [5]. The main types of ensembles used in Machine learning models are given below

(1) Bayes optimal classifier.
(2) Bayesian model combination.
(3) Stacking.
(4) Bayesian model averaging.
(5) Random Forest.
(6) Bucket of models.
(7) Bootstrap aggregating (bagging).
(8) Boosting (Ada Boost algorithm).

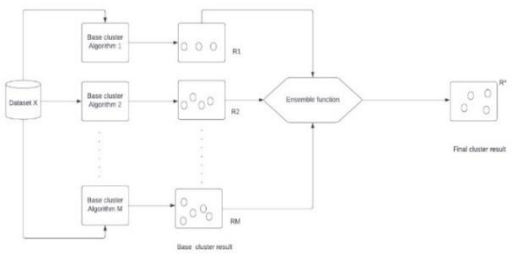

Figure 1. Ensemble cluster model

2.1 Bayes optimal classifier

It is the probabilistic model that is mostly used in prediction applications. The model is framed using the Bayes Theorem. The Bayes theorem was first introduced by Thomas Bayes, who developed the algorithm that uses evidence to compute the limits of an unknown parameter using conditional probability for the first time. It's a principled way to calculate the conditional probability. In the Classifier of Bayes Optimal, the outcome is selected with maximum probability. The computational price of the Bayes Optimal Classifier was high. The model gives the best result in almost all applications. Bayesian classification is a statistical classification based on Bayesian probability principles.

2.2 Bayesian model averaging

Bayesian model averaging (BMA) averages the weights of the posterior probability of each model to predict the results. BMA ensembles generally give better results than individual models. Bayesian Model Averaging was originally developed to combine results and predictions from multiple statistical models. It was widely utilized in applications where statistical linear regression and related models are used. BMA has been used in many deterministic prediction models as the statistical post-processing model to forecast the results. It gives better results than Bayes optimal classifier in classification problems.

2.3 Bayesian model combination

An algorithmic improvement to Bayesian Model Averaging is called Bayesian Model Combination. Rather than taking a sample from every model in the ensemble separately, it takes a sample from the space for possible ensembles. With this adjustment, BMA's possibility to converge toward giving one model the maximum weight is overcome. In terms of computation, BMC is more costly than BMA. The result from BMC is better than bagging and BMA. The probability of the data given for every model must be calculated to determine the model weights, which is made easier by the Bayes law. Since the training data were not generated from any of the models in the ensemble, each model appropriately receives a value close to 0 for this term. If the ensemble were large enough to sample the whole model space, this would work effectively. As a result, the ensemble weight will move in favor of the model in the ensemble that most closely resembles the distribution of the training data for each pattern in the training set. It boils down to an overly complicated model selection procedure. Cross-validation can be used to choose the optimal ensemble combination from a random sample of potential weights, roughly approximating the results from BMC. This model gives better classification results compared to other Bayesian models.

2.4 Bucket of models

It is an ensemble method where the optimal model for every issue is selected using a model selection algorithm. When examined on a single problem, a bucket of models cannot outperform the best model in the set; however, when tested on numerous problems, on average, it will outperform every model in the set. This model is suitable for the classification and clustering of multi-level applications.

2.5 Stacking

Algorithms for ensemble machine learning are called

stacking or stacked generalization. It gains the ability to integrate the forecasts from several effective machine learning methods. A stacked model operates by passing the results of several models through a meta-learner, which is typically a linear classifier. Each model's strengths are maximized, and its weaknesses are minimized by the meta-learner. This model is more suitable when different models are combined with a meta-learner.

2.6 Random Forest

The popular random subspace method Random Forest was used in 1995 to create the first algorithm for random decision forests. During training, many decision trees are built using this ensemble learning technique for classification, regression, and several tasks. The mean or average prediction made by each tree is returned for regression tasks. The strength of each classifier and the degree of dependence on them determine Random Forest's accuracy. The number of attributes chosen for every split's consideration affects the random forest. For classification tasks, every tree votes, and the major popular class is returned as a result. Hence, the Random forests generally outperform decision trees. Random forests require minimal configuration and produce reasonable predictions in a variety of data applications.

2.7 Bagging

Bootstrap Aggregation is referred to as "bagging" in short. It is an ensemble technique for reducing the prediction model's variance. Bagging is a parallel training approach that trains individual students apart from one another. Bagging generates training data by random sampling of original data. Certain observations can be repeated in each new training data set when sampling with replacement is used. Every element in the data set has an equal probability of being included in the new dataset. Bagging works as in the majority vote principle. This method is very appropriate to ensemble similar types of cluster models. This concept can be used to develop ensemble models to improve the cluster quality.

2.8 Boosting

Using several weak classifiers, this general ensemble approach creates a strong classifier. Using the training data, it constructs a model and then builds a second model that fixes the errors in the initial model. Up until the training set is perfectly predicted, models are added. Boosting is a sequential ensemble technique that modifies the weight of observations based on the most recent classification iteratively. The weight of an observation is increased if it is classified incorrectly. It creates robust predictive models and reduces bias error. During training, the Boosting algorithm gives weights to each of the generated models. A learner may be given a higher weight if their predictions of the training data are accurate.

2.8.1 Adaptive boosting

Adaptive boosting is a statistical classification meta-algorithm known as Ada Boost. To increase performance, it may be combined with a variety of other learning algorithm types. The final output of the boosted classifier is represented by a weighted sum that is created by combining the output of the other learning algorithms. The method is adaptive since the impact due to weak learners is adjusted by misclassifying the instances of previous classifiers. In some applications, the method is low susceptible to the overfitting issue than other learning algorithms. Even though the individual learners are ineffective, their performances are converged to strong learners. The best applications for Ada Boost are machine learning and binary classification tasks, where it improves decision tree efficiency. This model helps to consider the accurate base model with a weight. It is more applicable when considering base models with varying output results. Emphasis on certain models is possible by giving suitable weights to the results of that model and can improve the ensemble cluster results.

In the case of Bagging, the result is obtained by averaging the base cluster results. By adding the weight to the base models, the most appropriate results are obtained. Stacking model ensembles the results using a meta-learner. The random forest method is better than the decision tree models.

3. LITERATURE REVIEW

An ensemble in the context of machine learning is a model built using several independent methods operating concurrently, the output of which is integrated with a strategy for making decisions to generate a solution to an issue that is more accurate. The ensemble methods were initially used in supervised learning fields. The success of applying the ensemble method leads to using it in unsupervised applications such as cluster applications. By imposing a specific structure on the data, various clustering algorithms may yield varying clustering results for the same data. There are no set rules for selecting specific clustering algorithms for a given issue, nor is there a single clustering algorithm that consistently performs well for a variety of issues.

Several updated models are proposed to enhance the K-means cluster method. The researchers [1] developed a modified Kmeans cluster method. The distance calculation between samples and centroids for different iterations is limited in this method. To save needless calculations, the intermediate distances are stored in a data structure. The outcomes of earlier iterations in this technique can be applied to subsequent iterations. The method improves the clustering speed by lowering the frequent distance calculations in each iteration. Experimental outcomes show better accuracy and speed for this method.

The optimal use of random projection for clustering high-dimensional data is a crucial area of research in this field [2]. The random projection in a cluster ensemble approach improves the cluster performance. The proposed approach is mostly used for high dimensional data, which brings better, and more consistent clustering performance compared to individual algorithms.

A cluster count value collection is very important in the fast converging of the k means algorithm [3]. Utilizing the iris datasets from UCI, they examined the effectiveness of the cluster count value prediction algorithms Gap statics, Canopy, Elbow, and Silhouette. The work emphasizes the importance of conducting research works to analyze the performance of the Silhouette and Elbow technique utilizing standard data sets [4].

Taiyuan's power consumption data is categorized using the Python-based K-means plus clustering algorithm [5]. After classifying the electricity consumption data, the Kmeans plus clustering algorithm yields five distinct user types. They were

aware that Kmeans plus clustering algorithm is quicker than KMeans, and clustering results are more accurate. Optimization of the first cluster centers has led to an improved Kmeans Text Clustering Algorithm [6].

A novel ensemble classifier is developed by analyzing the performances of different base classifiers [7]. They proposed a Fusion classifier and Base classifier, two stages in a new ensemble model. By acquiring knowledge of the cluster boundaries, the base classifier generates the cluster confidences. The fusion classifier combines the cluster confidences to make decisions. The UCI data repository's dataset is used for the work, and two-tailed sign tests are used to determine the models' efficiency.

To use tumour clustering based on bimolecular data, fuzzy theory has been developed for the framework of cluster ensembles and proposed four different kinds of clusters that can be mixed [8]. To create a set of fuzzy matrices in the ensemble, they use various ensemble methods. Particularly with bio-molecular data, the suggested hybrid fuzzy cluster ensemble frameworks perform admirably on actual datasets. The suggested methods have the potential to produce more reliable, accurate, and stable results.

A new cluster-oriented ensemble classifier built on old concepts like the learning of clusters' boundaries by base classification methods and mapping of Cluster Confidences to Class Decision, is another significant progress in this area [9]. The data set that has been categorized is divided into multiple categories and supplied to various unique base classifiers. Cluster confidence vectors are generated, and cluster boundaries are learned by the base classifiers. The cluster trust is combined by a second-level fusion classifier, which maps to class decisions. To enable effective learning, the suggested ensemble classifier changes the base classifiers' learning domain. To determine the effect of multi-cluster boundaries on classification accuracy and classifier learning, the approach is tested on benchmark data sets from the repository for machine learning at UCI.

The analogous technique for ensemble clusters is suggested by the researchers [10]. They combine the dissimilar clusters to improve the cluster accuracy. They examine the potential of using the ensemble method by comparing the performances in medical diagnosis. They demonstrate several ensemble generation and integration techniques and assess each one using several fictitious and real-world datasets.

A technique for using ensemble cluster analysis to identify representative trends in ensemble weather forecasts within a chosen spatiotemporal area is presented [11]. This method helps to improve the performance of cluster formation.

An ensemble cluster method is proposed to link multiple clustering models to improve the consistency and quality of the cluster [12]. They introduced the Adaptive Clustering Ensemble model which gives cluster similarity and membership similarity. The improved final cluster is generated in three stages. They tested the methods using various benchmark datasets and the model is more accurate and efficient than the Meta-Clustering Algorithm and the Co-association method.

A Fuzzy Possibilistic Cmeans model is proposed to improve the quality of the cluster [13]. It generates both typicality and membership values when clustering unlabelled data. The new model they suggest is known as the possibilistic fuzzy Cmeans model. It generates the cluster centers for every cluster concurrently with possibilities and memberships. It is a cross between fuzzy Cmeans and possibilistic Cmeans. The technique fixes the FCM's noise sensitivity flaw. PFCM prototypes are a great option for fuzzy rule-based systems since they can prevent coincident clusters and are less susceptible to outliers [14].

The features of various extensions of the Kmeans algorithm were compared with modified versions of the algorithm [15]. The different extensions of Kmeans are used in machine learning and pattern recognition. The Kmeans method and its extensions are always influenced by initializations centroids and cluster count. They assemble an unsupervised learning schema for the k-means method, and it does not require initial seed selection. With no parameter or initialization selection required, they present a novel unsupervised Kmeans (U-Kmeans) clustering algorithm that automatically determines the ideal number of clusters. The computational complexity of the planned U-Kmeans clustering method and the existing algorithms shows improved results in the proposed algorithm.

The performances of Bagging and Boosting ensemble methods with classifiers are compared with neural networks and decision trees in 32 standard datasets [16]. The outcomes show that the Bagging ensemble was more accurate than the individual base classifier. Boosting is more accurate than Bagging and the individual base classification results. Analysis indicates that Boosting results vary with the datasets. Classification results of Bagging and Boosting vary with the number of base classifiers used in both neural network and decision tree classification models. The concept of bagging and boosting can be used to design ensemble models and can improve the cluster quality.

By combining several distinct partitions made from the same data, data with excellent-quality partitions are used to create the ensemble clusters [17]. Methods may be used to assess and select a subset of partitions, which provides an ensemble result superior to that obtained from the entire set of partitions to improve diversity and quality characteristics. This work investigates several partition evaluation and selection techniques, the majority of which are based on relative clustering validity indexes. The partitions with the best quality are chosen by these indexes to be a part of the ensemble. After combining the various relative indices, a final assessment is produced that is typically resistant to modifications made to the application. A useful design strategy for the ensemble selection was developed through a comparative analysis of data from many experiments.

The cluster validation criteria are compared with the silhouette index tool [18]. In cluster analysis, the silhouette index is frequently used to determine the optimal number of clusters. It is also used as a final clustering validation and evaluation tool. Relative compactness and separation of the cluster are well reflected in this index. Its straightforward interpretation rules and minimal computational complexity are its advantages. The performance analysis of this index with other cluster-validating indexes suggests considering the silhouette index as the base for cluster validation.

The studies to analyse the cluster quality based on internal cluster evaluation indexes are done [19]. The study focuses on the evaluation of the internal clustering impact and suggests an enhanced index called the Peak Weight Index (PWI) that is based on the Silhouette and Calinski-Harabasz indices. PWI takes the highest value of the two indexes as an impact point and assigns the appropriate weight within a given range. It integrates the features of the Calinski-Harabasz index and the Silhouette index.

The work is done to improve the different cluster evaluation

indexes [20]. Validity indices are a popular method for assessing clustering results. Two criteria are available for use in clustering validity methods: internal and external criteria. Various types of indexes are utilized to solve different types of issues. The results of this study denote that internal indexes are most accurate in cluster determining in each clustering structure. A general review of improving the cluster quality using different methods is done [21]. They reviewed different cluster quality-improving methods such as ensemble methods, summarization, and consensus clustering and realized that the cluster ensembles give better results.

Given a set of data to be examined, this literature review assists users in resolving the conundrum of choosing an appropriate technique and the corresponding parameters. It produces a variety of ensemble generation techniques as well as summarization and member representation. This review helps to select proper base classifiers in ensemble methods and suggest improvements in the ensemble strategies used in applications.

When we review the literature, we can understand that the boosting methods are more accurate than other methods. It reflects the efficiency of the base algorithm and many researchers use this method. Bagging, Stacking, and Boosting methods are very simple to implement. Even though many cluster-evaluating indexes are available, the majority of the researchers used the Calinski-Harabasz index, Silhouette index, C index, Dunn index, and DB index for evaluating cluster performance.

4. PROPOSED METHODOLOGY

In this section, we develop different ensemble models to cluster the data. We have used different cluster algorithms for ensemble modelling. Even though there are various partition algorithms used in cluster applications, none of the algorithms are perfect. In this work, we ensemble the base cluster algorithms Kmeans, Kmeans^{9+}, Fuzzy Cmeans, and Kmedoid. We can see that the efficiency of the ensemble algorithm is more effective than that of base cluster algorithms.

The combinations of hard and soft cluster algorithms are used for developing ensemble models. In the case of hard clustering techniques, the sample is assigned to a single cluster. In soft cluster techniques, the samples are given probabilities to belong to multiple clusters. The sample was assigned to the cluster in which the probability was greatest. We build several models that link the results of several methods and produce the result. We have used several ensemble methods to improve the performance of base cluster models.

The results of base classifiers and ensemble models are compared using the result of cluster validation indexes. The performances of the index values are given in tables. in the result section. The improvement in this cluster evaluating indexes are clear indication of improved cluster quality in ensemble models.

4.1 Simple hard voting ensemble

The simple hard voting is a hard cluster method. A dataset consisting of N samples is what we are working with. The data samples are represented by the letters S1, S2, ..., Sn. Assume for the moment that there are K clusters. C1, C2, ..., Ck are the cluster centres. It is computed what the distance Dj is between samples Si and Cj. Based on the lowest value of Dj, the sample Si is placed in the cluster with centre Cj. Data object Si in this instance of hard clustering is associated with the jth cluster, which has centre Cj. The cluster in which the sample resides is given a score of 1 and all other gives a cluster score of zero. We can do the cluster operation using various cluster models and can sum the scores corresponding to each base model. The resultant scores of a sample are used to determine the exact position of a sample in a cluster.

Here the Bagging ensemble technique is used to find the resultant cluster. Averaging of the resultant values obtained when using different base clusters is used to decide the position of the sample. The resultant ensemble cluster algorithm gives better results than that of base cluster models. We are using different cluster evaluation indices to analyse the performance of the ensemble cluster.

4.2 Soft boosting ensemble

It is another ensemble technique used to raise the quality of the cluster. The cluster models are ensembled using the weighted ensemble technique. The Figure 2 shows the flow diagram of a simple ensemble model. The distances between the sample and the various cluster centroids are used to compute the sample's probability of belonging to every cluster. We use the Eq. (a) to compute the probabilities of sample Si belonging to clusters 1, 2, ..., K.

$$\text{PY[i][j]} = \frac{\exp\left(-Eudist(S_i, c_j)\right)}{\sum exp(-Eudist(S_i, C_k)}, j = 1,2, \dots, k \qquad \text{(a)}$$

For the soft clustering method, we already had the probability of data object Si fitting to the jth cluster. Probabilities attained for the resulting ensemble technique can be formulated as shown in the below Equation.

$$\text{PY[i][j]} = \frac{\pi_z \text{PY[i][j][z]}}{\sum_j \pi_z \, PY[i][j][z]}, \; z = 1,2,.,m \; and \; j = 1,2,.,k \qquad \text{(b)}$$

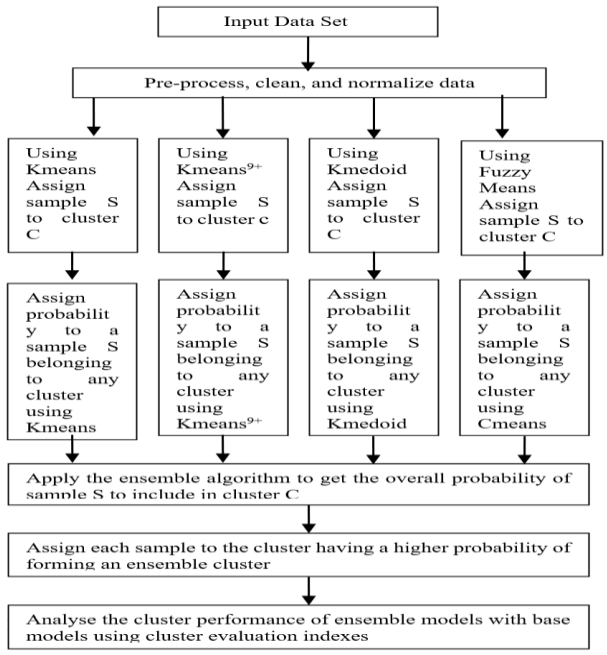

Figure 2. Simple ensemble model

Sample positions can be ascertained by calculating the resultant probabilities of the samples to include in each cluster.

5. CLUSTER ALGORITHMS

In this paper, we develop different ensemble models to cluster the data. We have used different cluster algorithms for ensemble modelling. Even though there are various partition algorithms used in cluster applications, none of the algorithms are perfect. In this work, we ensemble the base cluster algorithms Kmeans, Kmeans^{9+}, Fuzzy Cmeans, and Kmedoid. All these algorithms are partition algorithms. The conceptual approaches of these algorithms are similar. We also used different combinations of the selected cluster algorithms in the ensemble. More accurate cluster results are obtained when we use these base cluster algorithms. We can see that the efficiency of the ensemble algorithm is more effective than that of base cluster algorithms.

5.1 Kmeans algorithm

Machine learning clustering problems are solved via Kmeans clustering algorithms, which is an unsupervised learning algorithm [4]. The unlabelled dataset is divided into k distinct clusters with similar properties by the technique. The algorithm is centroid-based. A centroid is connected to every cluster. This algorithm's main aim is to reduce the total distance between each data point and its associated cluster center. The unlabelled dataset is the input used by the algorithm. The dataset is divided into k number of clusters by the algorithm. In this algorithm, the cluster count value is initially predicted.

5.2 Kmeans^{9+} algorithm

To calculate the cluster centroid, the Kmeans^{9+} algorithm is adjusted utilizing the statistical measures Mean, Median, and Partition Center. To determine whether to include a sample in a cluster using the traditional K-means algorithm, the samples are compared with each of the partition centroids. The samples in the Kmeans^{9+} method are only compared with the current cluster partition's centroids and the eight closest neighbouring cluster partitions. It increases the algorithm's performance and lessens the number of pointless comparisons between samples and cluster centroids. Kmeans^{9+}, the nine nearest neighbour uniform partition cluster model, enhances the Kmeans algorithm's efficiency and reduces the number of iterations required to obtain the natural cluster results. By using this method, the samples are not needlessly checked against the centroids of non-neighbour clusters. When applied to larger data sets with higher cluster count values, the model performs better.

5.3 Kmedoid algorithm

An unsupervised clustering algorithm called Kmedoid Clustering groups objects in unlabeled data. The distance between each data point in cluster I and every other data point is calculated and added. The medoid for an ith cluster is designated as the point in which the total distance calculated from other points is the least. An unsupervised clustering technique called K-medoid clustering groups objects in unlabeled data. K-means the sensitivity of clustering to outliers is high. Kmedoid Clustering provides an answer to this problem. Using this technique, the medoid serves as the reference point rather than the centroid of the objects in the cluster. An object in a cluster that is most centrally located is a medoid. Its average difference from every object is negligible. Compared to the K-means algorithm, the Kmedoid algorithm is more resilient to noise. The three algorithms used in this method, are CLARA (Clustering LARge Applications), CLARANS ("Randomized" CLARA), and PAM (Partitioning around medoids). The PAM was a widely used powerful algorithm.

5.4 Fuzzy Cmeans algorithm

Fuzzy Cmeans clustering (FCM) is one of the popular soft clustering methods. Here, each sample is assigned a probability score to belong to that cluster [8]. Every sample in the Fuzzy Cmeans clustering process has a weight assigned to it. Unlike the Kmeans algorithm, the Fuzzy Cmeans algorithm won't overfit the data for clustering. It will be more beneficial to mark the sample to multiple clusters rather than just one than to only one cluster [13, 14]. It is important to understand that fuzzy Cmeans are essentially Kmeans in which the probability function is set to 1 in the case of a data point that is closest to a centroid and 0 in all other cases.

6. IMPLEMENTATION RESULTS

There are various approaches used to develop and implement the ensemble models. In our model we mainly used bagging and boosting concepts. The generally used partition algorithms are analysed based on their cluster results, using cluster evaluation indexes and most appropriate combinations of cluster algorithms are selected to ensemble. Our own developed cluster algorithm, Kmeans^{9+} is also used to form ensemble model.

This section deals with the description of data sets, Cluster evaluation indexes, and results based on different ensemble methods.

6.1 Datasets

To analyze and assess the efficacy of various ensemble techniques, we employed a variety of data sets. The utilized datasets are IRIS, Abalone, Covid_19_clean_complete, Mall_Customers, All_india_po_list, Yeast, and Shuttle.

(1) 50 samples from each of the three species of iris-Iris virginica, Iris versicolor, and Iris setosa-make up the IRIS data set. Four measurements were taken from every sample: the petals' and sepals' lengths and widths, expressed in centimeters. To differentiate the species from one another, Fisher created a linear discriminant model based on a combination of these four characteristics.

(2) The Abalone data collection comprises 4177 data instances with eight different properties.

(3) Mall-Customer data consists of 201 instances and 5 attributes.

(4) Covid19_clean_complete data set consists of 49069 instances and 15 attributes.

(5) All_india_po_list gives the total listing of the post office details in the country. The data set consists of 15 attributes and 154798 instances.

(6) The shuttle dataset contains 9 attributes all of which are numerical with the first one being time. There are 58000 instances for the shuttle data set.

(7) Yeast dataset.

6.2 Evaluation metrix for cluster algorithms

Clustering algorithms are evaluated based on performance and quality using evaluation measures. There are two types of clustering assessment methods: unsupervised evaluation using an internal criterion and supervised evaluation using an external criterion, Calinski-HarabaszI index, Dunn Index, Silhouette Index, DB index, and C index are widely used clustering evaluation metrics.

6.2.1 Silhouette index

The higher Silhouette Index is the indication of well-defined clusters. For a single sample, the silhouette coefficient is given

$$SI(i) = \frac{NC(i) - SC(i)}{\max(SC(i), NC(i))}$$

The sample's mean distance from every other sample in the same cluster is shown here by SC(i). This score represents the degree of closeness between points in a similar cluster. While NC(i) represents the mean distance between the sample and every other point in the next closest cluster. This score is used for assessing the distance among samples in various clusters. The average of the silhouette Index for each sample determines the silhouette coefficient for a sample group. The range of the score is -1 for improper clustering and +1 for extremely dense clustering. Clusters that overlap are indicated by scores that are almost zero. For convex clusters, the Silhouette Coefficient is typically greater.

6.2.2 Dunn index

In 1974, J. C. Dunn presented the Dunn index (DI), the statistic for evaluating clustering techniques. Improved cluster performance is indicated by a higher index value. The formula for calculating it is to divide the largest intra-cluster distance by the minimal inter-cluster distance, or the minimum separation of any two cluster centroids.

6.2.3 Calinski-Harabasz index

The CH Index compares an object's isolation from different clusters to its cohesion within its cluster [19]. Another name for this measure is the Variance Ratio Criterion (VRC). It was computed as the ratio of the total dispersion between clusters to the inter-cluster dispersion for every cluster. Better performance is indicated by the high score.

6.2.4 Davies-Bouldin index

The average similarity between clusters is assessed by this index. A model with a stronger separation between the clusters is associated with a lower Davies-Bouldin index. It is very simple to compute. Only values and attributes that are inherent to the dataset are used to generate the index.

6.2.5 C index

This index represents the average degree of similarity amongst clusters. Its goal is to assess how dispersed individual data clusters are about the overall dispersion of the dataset. Its goal is to assess how dispersed individual data clusters are about the overall dispersion of the dataset. A better performance in clusters can be seen by a lower index value.

6.3 Implementation details of the work

We have developed and implemented the cluster algorithms and cluster evaluation indexes in Python. Sklearn library functions are mainly used to implement the cluster evaluation indexes. Sklearn library provides many built-in functions for clustering such as Kmeans, Kmedoid, and cluster evaluating functions. The Python libraries numpy has been used as an object to arrange the samples in a cluster to find the silhouettes of the cluster. The panda's object is used to read the values from CSV files using library functions and to store the data sets in list format.

The matplotlib is used similarly to a graphical interface and many graphical library built-in functions are used to plot and show the pictures according to the user's requirement. The standard library functions are used to design and implement the cluster evaluation indexes. Ensemble models such as Simple Voting Ensemble and Boosting Ensemble Methods are developed in Python and implemented using small and big data datasets.

7. PERFORMANCE EVALUATION

The Cluster operations are performed using different cluster models and the performances of the cluster are evaluated using different cluster evaluation indexes. Different ensemble algorithms are used to ensemble the base cluster algorithms and the results of cluster operations performed are compared with the results obtained using individual base cluster algorithms.

In the case of hard and soft clustering, we use the ensemble approach Bagging and Boosting. All the cluster results using different cluster algorithms are summed and the average value is taken to decide the position of cluster for samples. In the case of soft clustering, the probability of occurrence of a sample in a cluster is summed and that value is utilized to determine the membership of samples in a cluster. In almost all data sets the ensemble cluster results are better than the individual base cluster results.

7.1 Results and discussions

We use Kmeans, Kmeans^{9+}, Kmedoid, and Cmeans algorithms for base clusters for performance evaluation.

These cluster algorithms are used to develop both the hard and soft ensemble models. Various cluster performance factors such as Compactness, Cohesion, and Separation similarity are given by different evaluating indexes.

We have separately calculated these indexes for all base cluster models and the developed hard and soft ensemble models using 7 data sets. The results of these indexes for the base and ensemble models indicate the high performance in ensemble models. The results of the ensemble models are denoted as Ens-1, for Simple Hard Voting Ensemble and Ens-2 for Soft Boosting Ensemble.

Table 1 gives the cluster evaluation indexes of the basic models as well as the developed ensemble models. The performance of the clusters is increased by around twenty four percent in ensemble models.

Figure 3 gives the pictorial view of the cluster performance of the IRIS data set using different cluster evaluation indexes. The developed ensemble models give better indexes and hence better cluster results. Since the Calinski-Harabasz index (CH index) values are very high, the CH index is not considered when drawing a graph.

Table 2 gives the comparison of base algorithms and

developed ensemble algorithms.

Figure 4 gives the graphical view of the cluster performances for the Abalone data sets.

Table 1. Results using the IRIS dataset

Index	K Means	K Means^{9+}	K Medoid	C Means	Ens1	Ens2
Silhouette	.55	.53	.50	.53	.56	.58
CH	562	559	555	560	572	586
DB	.66	.68	.70	.66	.64	.62
C	.72	.72	.75	.70	.68	.66
Dunn	.339	.339	.34	.37	.41	.43

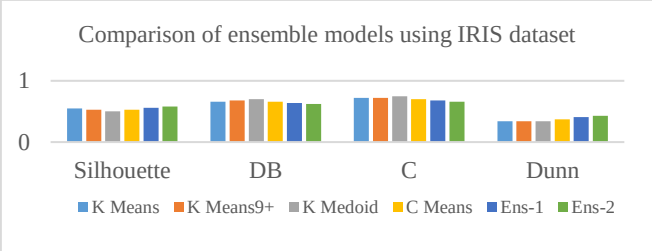

Figure 3. Comparison of ensemble models using IRIS data

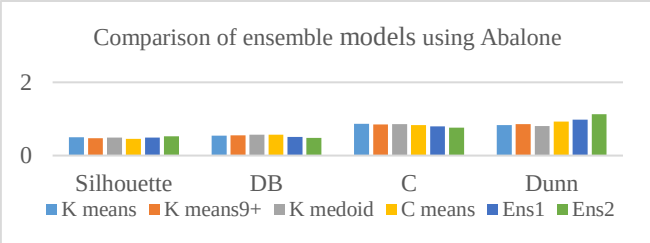

Figure 4. Comparison of ensemble models using Abalone

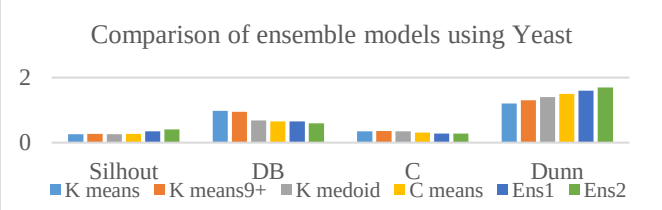

Figure 5. Comparison of ensemble models using Yeast data

Table 2. Results using Abalone dataset

Index	K Means	K Means^{9+}	K Medoid	C Means	Ens1	Ens2
Silhout	.5	.47	.49	.46	.49	.53
CH	23361	23550	23571	23460	23640	23680
DB	54	.55	.57	.57	.51	.48
C	.87	.85	.86	.83	.8	.76
Dunn	.83	.86	.81	.93	.98	1.13

Table 3. Results using Yeast dataset

Index	K Means	K Means^{9+}	K Medoid	C Means	Ens1	Ens2
Silhout	.26	.27	.26	.27	.35	.41
CH	215	210	220	236	256	263
DB	.98	.95	.68	.65	.65	.59
C	.35	.36	.35	.31	.28	.28
Dunn	1.2	1.3	1.4	1.5	1.6	1.7

Table 3 shows the cluster performances of two developed ensemble models. Silhouette values and Dunn index values are improved very much in ensemble models.

Figure 5 shows the pictorial version of the performance of the ensemble models using Yeast data sets. The boosting Ensemble Method gives better cluster results.

Table 4 gives the comparisons of the performance of developed ensemble models using a small Mall dataset.

Figure 6 shows the performances of developed ensemble models. The results show that the ensemble models give very good cluster improvement sin small data sets.

Table 5 gives the performance comparison of ensemble models using the Covid_19_clean_complete dataset.

Figure 7 gives the graphical representation of cluster evaluation indexes of base and ensemble models.

Table 4. Results using Mall dataset

Index	K Means	K Means^{9+}	K Medoid	C Means	Ens1	Ens2
Silhout	.42	.45	.49	.45	.53	.56
CH	245	269	167	248	276	288
DB	.84	.57	.56	.76	.55	.52
C	.86	.84	.86	.85	.79	.76
Dunn	.65	.67	.65	.66	.71	.73

Table 5. Results using Covid_19 _clean _ complete dataset

Index	K Means	K Means^{9+}	K Medoid	C Means	Ens1	Ens2
Silhout	.54	.49	.5	.51	.54	.57
CH	120633	117210	118310	118567	124670	129455
DB	.61	.69	.56	.55	.48	.44
C	.34	.36	.36	.36	.32	.29
Dunn	.98	1.2	1.2	1.1	1.4	1.5

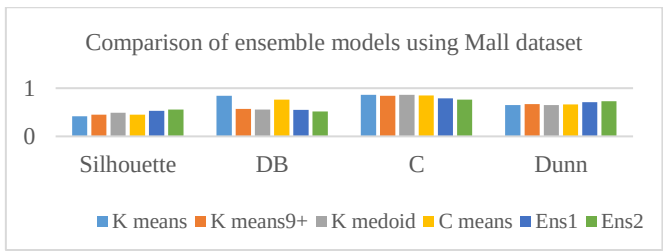

Figure 6. Comparison of ensemble models using Mall dataset

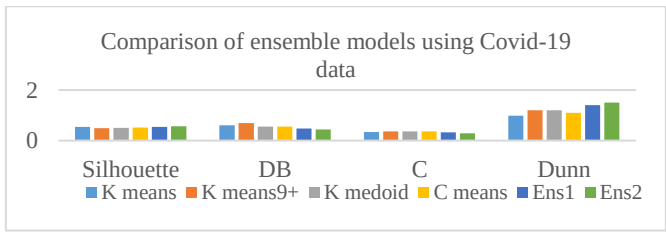

Figure 7. Comparison of ensemble models using covid-19 data

Table 6. Results using All_India_PO_List data set

Index	K Means	K Means^{9+}	K Medoid	Ens1	Ens2
Silh	.6	.6	.49	.63	.65
CH	1785159	1785208	1785222	1785881	1786444
DB	.49	.57	.52	.47	.45
C	.31	.32	.31	.29	.29
Dunn	1.55	1.58	1.6	1.82	1.87

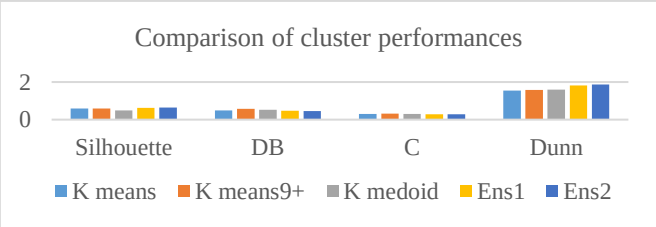

Figure 8. Comparison of ensemble models using All-India-PO-List data set

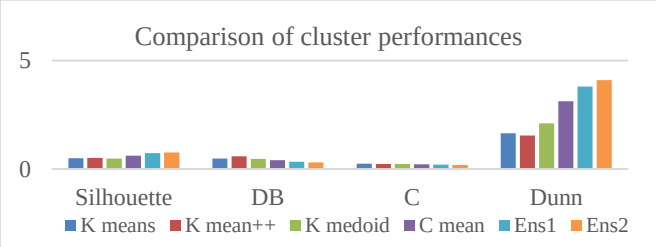

Figure 9. Comparison of cluster performances

Table 7. Results using Shuttle data set

Index	K Means	K Means^{9+}	K Medoid	C Means	Ens1	Ens2
Silhout	.50	.51	.49	.62	.73	.77
DB	.49	.59	.47	.41	.34	.31
C	.24	.23	.23	.22	.21	.19
Dunn	1.65	1.55	2.1	3.12	3.8	4.1

The performance comparison of the very big data set, All_India PO_List is given in Table 6.

Figure 8 shows the pictorial view of the performance of ensemble models for large data sets. The *Calinski-Harabasz index value or* CH value obtained using the All_India_PO_List data set is very high. The cohesion of samples in the generated clusters is indicated by the high CH value. Soft Boosting Ensemble Model gives better cluster performance.

Table 7 gives different cluster evaluation indexes of base and developed models using the Shuttle data set.

The pictorial representation of the Cluster evaluation indexes is shown in Figure 9. Since the CH value is very high, it is not shown with other indexes in the graph. In almost all cases, Soft Boosting Ensemble Model gives better performance.

For developed ensemble models, almost all performance evaluation indices yield superior results. The cluster assessment Indices' results are displayed in tables with unique results for every data set.

8. CONCLUSION

This paper discusses various ensemble methods used to improve the performance of cluster algorithms for Machine learning applications. The performance of various cluster algorithms is analyzed and selected Kmeans, Kmeans^{9+}, Kmedoid, and Fuzzy Cmeans algorithms to the ensemble. Data sets from UCI and other widely available public data sources are used to evaluate the cluster performance. The study's findings indicate that ensemble cluster models outperform all the individual base models in terms of accuracy and cluster quality. The Soft Boosting Ensemble model gives almost 100 percent accuracy in the majority of data sets. The simple Hard Voting Ensemble model also gives better cluster results than the base models. The improved values in cluster evaluation indexes in ensemble models using different data sets establish that the developed ensemble models are very effective in improving the quality of the cluster. There is future scope to improve the efficiency of cluster using different types of base cluster algorithms and modified ensemble models.

REFERENCES

[1] Na, S., Xumin, L., Yong, G. (2010). Research on k-means clustering algorithm: An improved k-means clustering algorithm. In 2010 Third International Symposium on Intelligent Information Technology and Security Informatics, IEEE, Jian, China, pp. 63-67. https://doi.org/10.1109/IITSI.2010.74

[2] Fern, X.Z., Brodley, C.E. (2003). Random projection for high dimensional data clustering: A cluster ensemble approach. In proceedings of the 20th international conference on machine learning (ICML-03), Washington DC, pp. 186-193.

[3] Yuan, C., Yang, H. (2019). Research on K-value selection method of K-means clustering algorithm. Multi-Disciplinary Scientific Journal, Graduate Institute, Space Engineering University, Beijing 101400, China, 2(2): 226-235. https://doi.org/10.3390/j2020016

[4] Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A.Y., Foufou, S., Bouras, A. (2014). A survey of clustering algorithms for big data: Taxonomy and empirical analysis. IEEE Transactions on Emerging Topics in Computing, 2(3): 267-279. https://doi.org/10.1109/TETC.2014.2330519

[5] Zhao, Z., Wang, J., Liu, Y. (2017). User electricity behavior analysis based on K-means plus clustering algorithm. In 2017 International Conference on Computer Technology, Electronics and Communication (ICCTEC), IEEE, Dalian, China, pp. 484-487. https://doi.org/10.1109/ICCTEC.2017.00111

[6] Xiong, C., Hua, Z., Lv, K., Li, X. (2016). An improved K-means text clustering algorithm by optimizing initial cluster centers. In 2016 7th International Conference on Cloud Computing and Big Data (CCBD), Macau, China, pp. 265-268. https://doi.org/10.1109/CCBD.2016.059

[7] Rahman, A., Verma, B. (2011). Novel layered clustering-based approach for generating ensemble of classifiers. IEEE Transactions on Neural Networks, 22(5): 781-792. https://doi.org/10.1109/TNN.2011.2118765

[8] [8] Yu, Z., Chen, H., You, J., Han, G., Li, L. (2013). Hybrid fuzzy cluster ensemble framework for tumor clustering from biomolecular data. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 10(3): 657-670. https://doi.org/10.1109/TCBB.2013.59

[9] Verma, B., Rahman, A. (2011). Cluster-oriented ensemble classifier: Impact of multicluster characterization on ensemble classifier learning. IEEE Transactions on Knowledge and Data Engineering, 24(4): 605-618. https://doi.org/10.1109/TKDE.2011.28

[10] Greene, D., Tsymbal, A., Bolshakova, N., Cunningham, P. (2004). Ensemble clustering in medical diagnostics. In Proceedings. 17th IEEE Symposium on Computer-Based

Medical Systems, Bethesda, MD, USA, pp. 576-581. https://doi.org/10.1109/CBMS.2004.1311777
[11] Kumpf, A., Tost, B., Baumgart, M., Riemer, M., Westermann, R., Rautenhaus, M. (2017). Visualizing confidence in cluster-based ensemble weather forecast analyses. IEEE Transactions on Visualization and Computer Graphics, 24(1): 109-119. https://doi.org/10.1109/TVCG.2017.2745178
[12] Alqurashi, T., Wang, W. (2019). Clustering ensemble method. International Journal of Machine Learning and Cybernetics, 10: 1227-1246. https://doi.org/10.1007/s13042-017-0756-7
[13] Pal, N.R., Pal, K., Keller, J.M., Bezdek, J.C. (2005). A possibilistic fuzzy c-means clustering algorithm. IEEE Transactions on Fuzzy Systems, 13(4): 517-530. https://doi.org/10.1109/TFUZZ.2004.840099
[14] Sreenivasarao, V., Vidyavathi, S. (2010). Comparative analysis of fuzzy C-mean and modified fuzzy possibilistic C-mean algorithms in data mining. International Journal of Computer Science and Technology, 1(1): 104-106.
[15] Sinaga, K.P., Yang, M.S. (2020). Unsupervised K-means clustering algorithm. IEEE Access, 8: 80716-80727. https://doi.org/10.1109/ACCESS.2020.2988796
[16] Opitz, D., Maclin, R. (1999). Popular ensemble methods: An empirical study. Journal of Artificial Intelligence Research, 11: 169-198. https://doi.org/10.1613/jair.614
[17] Naldi, M.C., Carvalho, A.C.P.L.F., Campello, R.J. (2013). Cluster ensemble selection based on relative validity indexes. Data Mining and Knowledge Discovery, 27: 259-289. https://doi.org/10.1007/s10618-012-0290-x
[18] Starczewski, A., Krzyżak, A. (2015). Performance evaluation of the silhouette index. In Artificial Intelligence and Soft Computing: 14th International Conference, ICAISC 2015, Zakopane, Poland, June 14-18, 2015, Proceedings, Part II. Springer International Publishing, 14: 49-58. https://doi.org/10.1007/978-3-319-19369-4_5
[19] Wang, X., Xu, Y. (2019). An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index. In IOP Conference Series: Materials Science and Engineering. IOP Publishing, 569(5): 052024. https://doi.org/10.1088/1757-899X/569/5/052024
[20] Rendón, E., Abundez, I., Arizmendi, A., Quiroz, E.M. (2011). Internal versus external cluster validation indexes. International Journal of Computers and Communications, 5(1): 27-34.
[21] Boongoen, T., Iam-On, N. (2018). Cluster ensembles: A survey of approaches with recent extensions and applications. Computer Science Review, 28: 1-25. https://doi.org/10.1016/j.cosrev.2018.01.003