J. Integr. Sci. Technol. 2024, 12(5), 824

Journal of Integrated
SCIENCE & TECHNOLOGY

Combined clustering with classification in a semi-supervised context: An efficient data partitioning

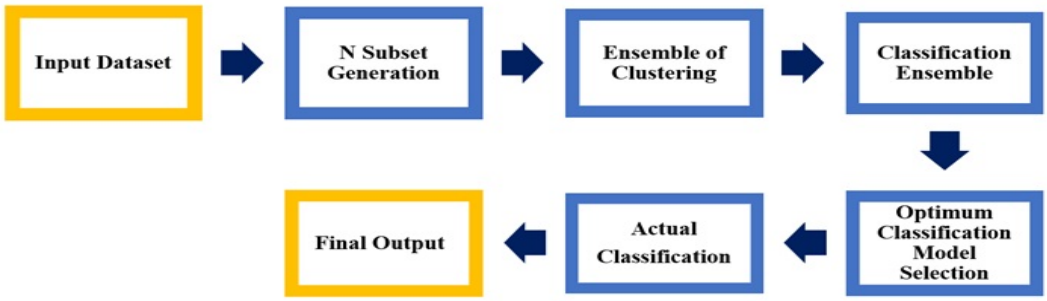
Govind Pole,* Pradeepini Gera

Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Andhra Pradesh, Guntur, India.

Received on: 14-Jan-2024, Accepted and Published on: 07-May-2024

Article

ABSTRACT



The studies on utilizing the output of unsupervised clustering techniques with a supervised classifier are pushing the concept of semi-supervised learning. Current ensemble models give the basic results and the consensus for the applied function. Very few attempts have been made to combine clustering methods with fundamental classifiers. The proposed method uses supervised clustering to partition the data into groups. The next step is to pairwise combine clusters from different groups to construct a number of training subsets. Each subgroup of the training set is given a unique base classifier, and he outputs of these base classifiers are consolidated through a specialized Consensus function. The weight given to a base classifier is based on how well it classifies the data. The experimental findings demonstrate that, compared to base classifiers and traditional ensemble classification methods, the proposed method 'Ensemble of Clustering and Classification (ECC).' gives a general performance upshift up to 10%. Furthermore, it provides base classifiers with enhanced diversity and accuracy, thereby enhancing the data analysis process.

Keywords: Cluster-based ensemble, classification ensemble, categorical data, bagging, boosting.

INTRODUCTION

Semi-supervised clustering and ensemble clustering have recently become significant paradigms in traditional clustering. The goal of ensemble clustering is to combine several clustering findings from various approaches or from the same approaches but with different parameters. In semi-supervised clustering, the learning process makes use of a tiny amount of class membership information in certain samples. In this case, a consensus function based on the cluster clustering strategy configures the target partition, whereas the primary partitions are formed using different hierarchical clustering algorithms.

The majority of studies used random sampling to create training subsets to build the fundamental classifiers. As a result, their diversity cannot be guaranteed, which may reduce the categorization performance as a whole.[1,29] A novel strategy that makes use of certain facts is suggested, such as the fact that a group's properties are comparable to those of its constituent items. There is a greater likelihood that two objects will belong to the same class if they are members of the same basic groups. The class distribution of an object resembles the average class distribution that was determined by a number of base classifiers. A group's class distribution resembles the average class distribution of the items that make up the group.

Algorithms for classification and clustering have each been shown to be effective in a variety of settings. Both have advantages and disadvantages of their own. For instance, despite the fact that classification algorithms are more effective than clustering approaches at predicting an object's class label, they do not perform well in the absence of a substantial amount of valid hand-labelled data. However, despite not producing label information for objects, clustering algorithms offer additional constraints; eg. If two objects

*Corresponding Author: Govind Pole, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Andhra Pradesh, Guntur, India
Email:govindpole@gmail.com

Cite as: *J. Integr. Sci. Technol., 2024, 12(5), 824.*
URN:NBN:sciencein.jist.2024.v12.824

are associated with each other, they will likely be assigned the same label, which can be used to predict the unknown object's label.[2,14,18] As a result, systematic use of both types of algorithms together can improve forecast accuracy.

Certain complex data cannot always be fit into the ensemble prediction methods that are currently in use, such as bagging[5] and boositing.[6] As part of an ensemble approach called "bagging," several models are trained separately on arbitrary subsets of the data, and their predictions are then combined by averaging or voting. In contrast, boosting is an ensemble learning technique that reduces training errors by turning a group of weak learners into a strong learner. Performance improvement in the classification process is not guaranteed by the widespread use of bagging and boosting. A tailor-made approach for a particular data type yields superior results. When analyzing data, categorical data require extra care. Furthermore, the accuracy of the clustering cannot be guaranteed by the unpredictability of the clustering process. Thus, there is a great deal of room for study to increase the predictive accuracy of the ensemble model, the adaptability to complicated data, and the capacity to lessen the effect of randomness on clustering accuracy.

Different classifier models are created in the suggested system by training samples. Next, the prediction results of the basis classifiers in the verification set are clustered using the Kmeans method to increase the difference between the base classifiers. Ultimately, each cluster's base classifier with the best generalization performance is chosen for ensemble learning, and the bagging method is used to attain the desired outcome. To strike a balance between competitiveness and robustness in the proposed ensemble model, a clustering-based regular partition assignment technique is suggested. This strategy helps create appropriate partitioning and mitigates the influence of outliers.

REVIEW OF THE LITERATURE

In recent years, multiple studies have been carried out to find substitute technique for creating clustering and classification models. Chakraborty et al.[3] proposed a technique that supports both binary and multiclass classification by combining classification with clustering. The ensemble approaches outperform other baselines for a single dataset. Both supervised[4-6] and unsupervised[7-12] learning has already demonstrated the efficacy of ensembles. Supervised and unsupervised models have both used ensemble learning. In order to enable more precise and superior decision making, machine learning ensemble approaches integrate the insights acquired from many learning models. The ensemble-based unsupervised clustering,[5,,7,9] and concentrate on modern supervised ensemble techniques as Bagging,[5] Boosting,[6] and XG Boost[12] are derivations of a variety of base classifiers. By treating the output of base methods as features, meta-feature generating techniques like stacking and adaptive combination of experts create a meta-learning method on top of already existing base methods. Techniques such as bagging, random forest,[16] which train base models on training data and use majority voting to reach agreement, while the Bayesian averaging model[15] uses training data to learn both base models and how their outputs are combined, which requires a lot of labeled data.

Again, it is observed that other studies focused on how effectively select appropriate base models which are as diverse as feasible in order to reduce the generalized error.[17, 18] Incorporating clustering into the classification model in a semi-supervised way has been attempted using a variety of methods[19] and popular techniques like SemiBoost,[20] etc. Their main objective was to learn from a lot of unlabelled data and a little bit of labelled data. It took into account one basic classifier and one base clustering method. Various groups of the training set were created using 'Ensemble classification based on super-vised clustering,[21] and clusters of pair-wise classes were joined to form a number of training samples. Each classifier is trained using a specific training sample and the results are combined using a weighted voting method. Most frequently, a graph or a multidimensional array[16] are used to summarise the base results.

Cluster-based ensembles operate in two steps. Initially, alternative initializations for the same technique or distinct clustering algorithms are used to generate multiple partitions as ensemble members. The various partitions are eventually combined into the consensus partition via a particular consensus procedure [22]. Consensus functions have been created, including voting-based consensus functions,[8,9] attention-based boosting ensemble method,[21] hyper-graphic partitioning,[22] etc. Multiple clustering algorithms have been used to generate input partitions,[22] a single algorithm for clustering has been used with various initializations and parameter settings,[23] a single algorithm for clustering has been used with various features extracted from the original data,[24] and a stand-alone base clustering algorithm has been applied to the sampled partition of the original dataset.

By connecting the created clusters with the appropriate class labels, M. Halder et al.[27] have suggested a method to improve the classifier's performance by creating a connection between the supervised classification task and unsupervised clustering.

For experimental analysis, three prominent algorithms are utilized: Classification & Regression Tree (CRT), Iterative Dichotomizer 3 (ID3), an extension of which is known as C4.5, and the Naive Bayes classifier. Additionally, ensemble classifiers such as Random Forest, Bagging, and Boosting are employed.[27] D. Yuan et al.[28] proposed a novel method for clustering ensemble selection utilizing a random forest approach with the Dunn index. Furthermore, they developed a random forest technique to address the limitations of the hierarchical clustering algorithm's irrevocable merging strategy, selecting a hybrid clustering ensemble based on hierarchical clustering and partition clustering of k-medoids.. To enhance the generalization performance of the naive Bayes classifier's generalization performance, T. Liu et al.[29] suggests a technique based on Kmeans++ clustering technology to improve the integration difference of the classifier. In an effort to increase the accuracy and generalizability of the ensemble and leaching models, G. Hu and F. Yang[30] have developed a unique selected ensemble approach which takes into account both the creation of submodels as well as the combinations of submodels. Initially, the bagging algorithm-based multimodel ensemble hybrid model (MEHM) is suggested. The data model and the mechanism model make up the sub-models. The suggested-based vector bootstrap sampling procedure is used by the data model to create training

subsets. Next, a novel binary particle swarm optimization (PSO)-based selective multimodel ensemble hybrid model (NSMEHM) is introduced. In this model, the group of MEHMs that reduces error and improves diversity is determined using the binary PSO optimization algorithm.[30]

In their work, J. Yuan et al.[31] provide a clustering-based dynamic ensemble forecasting technique for non-stationary oil prices. In particular, the ensemble forecasting framework incorporates clustering, by which historical observations for the given period are automatically categorized into many clusters based on the attributes of the data. This classification offers a strong foundation for the focused dynamic evaluation of individual forecasting models. A unique similarity measure and stratified feature sampling are the foundations of Hui Shi et al.'s[32] semi-supervised hierarchical ensemble clustering system. The TWC-EL model, created by Xunjin Wu et al.[33] is a multivariate prediction model that makes use of ensemble learning and three-way clustering (TWC). To improve clustering accuracy, we initially partition the sample set using the k-means clustering algorithm. Following this, we conduct a secondary division of the sample set using the same method. The outcomes of these dual clustering processes are then amalgamated, considering both the variance in intersection points and the distance from samples to each cluster's centroid. This integration gives rise to a novel TWC (Two-Step Weighted Clustering) method. Subsequently, we delineate the core and fringe regions within each cluster from the initial clustering results. These regions are then classified into three categories—low, medium, and high—based on their interrelation. Finally, leveraging the strengths of the Elman neural network model, the Extreme Learning Machine (ELM) model, and the back-propagation neural network (BPNN) model, we construct an ensemble prediction model.[33] In contrast to an individual clustering method, the Baohua Shen et al.'s technique[34] combines the outcomes of multiple output partitions to produce a consensus that is more accurate.

After reviewing the latest advancements in classification and clustering models, several facts become apparent. Working with huge datasets, or "big data," makes the labor-and resource-intensive process of classifying data extremely difficult. As a result, it is possible to identify examples' commonalities rather than labels with professional annotations. The adage "the more base classifiers, the better the classification effect" isn't necessarily accurate because a model with a higher number of underperforming base classifiers may end up producing a worse classification resultOne of the main areas of study in machine learning is multivariate data analysis, which focuses on making advantage of the inherent relationship between feature variables and target variables. Nevertheless, current single prediction methods frequently fall short of optimal outcomes in complex multivariate prediction contexts.

Cluster-based ensemble : The benefits of distributed processing include the high performance of less expensive computers.[27] As a result, the data mining process uses a distributed clustering system implementation for high performance. There is often a conundrum as to which algorithm or combination of algorithms should be taken into consideration to address problems with the best results because several algorithmic solutions are offered for the same data mining activities. Cluster-based ensembles combine several iterations of various algorithms on a dataset to produce a single, aggregated clustering, hence optimising the clustering outcome. Based on KMeans, Expected Maximization-EM, and Hierarchical Clustering -HC techniques, a number of studies on ensemble-based distributed clustering have been conducted. A distributed hierarchical clustering approach that requires numerous iterations of message passing for categorical data. It clusters categorical data using various clustering techniques, such as 1. k means one hot encoding, which first converts categorical data to numeric data before performing k-means clustering. 2. K-modes: When grouping categorical data, it uses the modes of categorical attributes rather than the means in k-means. 3. Rock: The Rock method, which was created exclusively for categorical variables, will be applied to the categorical data. Categorical data are transformed to Booleans in this instance.

Ensemble classification based on supervised clustering : nterdependencies between items are not taken into account when tracking distributions because items are often distributed all at once, assuming they are selected from independent and uniform distributions.[8] Additionally, due to limited label information, it may be difficult to classify and predict the labels of unknown products. On the other hand, unsupervised clustering techniques increase this by taking product relationships into account, thereby imposing additional constraints on product distribution. For example, pairs of items that are close together in a given location are more likely to receive the same label than items that are further apart. Especially when there is a small amount of data, these additional parameters can help improve the overall ability of the final classifier.They can also be helpful when building learning systems if the distributions of the training and testing data differ noticeably. Combining classification and clustering algorithms can produce improved classification results, as recent work has demonstrated. classification results are presented in terms of well-known performance metrics such as AUC score.[31] However, how to mix classification with clustering is the key issue.

From the previous work it is observed that the idea of semi-supervised learning is anchored in the ongoing researches on how to use the results of unsupervised clustering techniques with a supervised classifier. In order to provide a single, precise solution, in this work a method is designed with the mixture of several supervised - unsupervised algorithms. Current ensemble models are unable to combine clustering methods with a number of fundamental classifiers, hence in this work an attempt is made to use supervised clustering to divide the data into a number of groups.

DESIGNED ENSEMBLE METHODOLOGY

The large categorical data set is divided into multiple subsets and distributed to different machines (slave) through coordinators as shown in Figure 1. The slave can act as coordinator by further distributing data to multiple machines if needed. Thus, every machine can act as a coordinator and slave. Once the data are received by the slave machine, the clustering using different values of K (number of clusters) starts. All clusters created by each participating node are evaluated for compactness and quality of clustering. The best value of K (number of clusters) is selected from all results by the coordinator machine. The best value of K is used

for further processing. This best value is achieved with the help of the algorithm 'Cluster Based Ensemble-member Selection (CBES)'.[37] Primary clusters are created with the best K-value derived by CBES. The created clusters are further used as training data sets to prepare the different classifier models. Different models derived from classification ensemble are evaluated by using Class membership probability (Cmp) for all existing classes in the datasets. The user-defined threshold δ1,δ2 are used to convey the lables of the Class membership (Cm) for all created classifiers. The optimum classification model is derived from this process and used for the final classification to produce the optimised final output.

A common system combining classification and association is readily available to demonstrate that the amalgamation of classification and clustering models can improve prediction results. It is based on two basic assumptions; (i) the final prediction should not differ too much from the majority votes of the classifier; (ii) if two objects are combined with different grouping algorithms, they will most likely belong to the same category. class. You must ensure that each group or group of products is made of similar products (i.e. the products in the group will have the same characteristics). products. An optimization problem was created for this task. Although many techniques have been developed for various tasks and processes, unsupervised learning still has limitations such as group formation and sample selection[2]. The latter focuses on bringing items together to form a group, while the former looks for a solution that can address the number of groups in the space in depth.

Optimization Principles

The following optimization principles create the motivation for ensemble creations and form the fundamental base for the Ensemble of Clustering and Classification (ECC) algorithm.The four components constituting our final goal function are derived from the following hypotheses:

(i) The similarity between the members of a group and the group itself: If an object is part of a group, then the class distributions of both the object and the group should be comparable.

(ii) Similarity between two objects inside a group: The "co-occurrence principle" asserts that the likelihood of two items belonging to the same class increases with the frequency with which they are assigned to the same categories. The co-occurrence principle operates using a method that searches the records for two or more concepts that frequently occur together. When two or more entities appear together in a set of records more often than not and infrequently individually in any other records, they are considered to have strongly co-occurred. (iii) The alignment between an object's final class distribution and its average class distribution, as determined by the base classifiers, follows the consensus principle. It is expected that the final class distribution of an object closely mirrors its average class distribution. (iv) The similarity between a group's final class distribution and the average class distribution of its constituent objects is paramount. A group's final class distribution ought to mirror more closely the average class distribution of the objects it comprises.

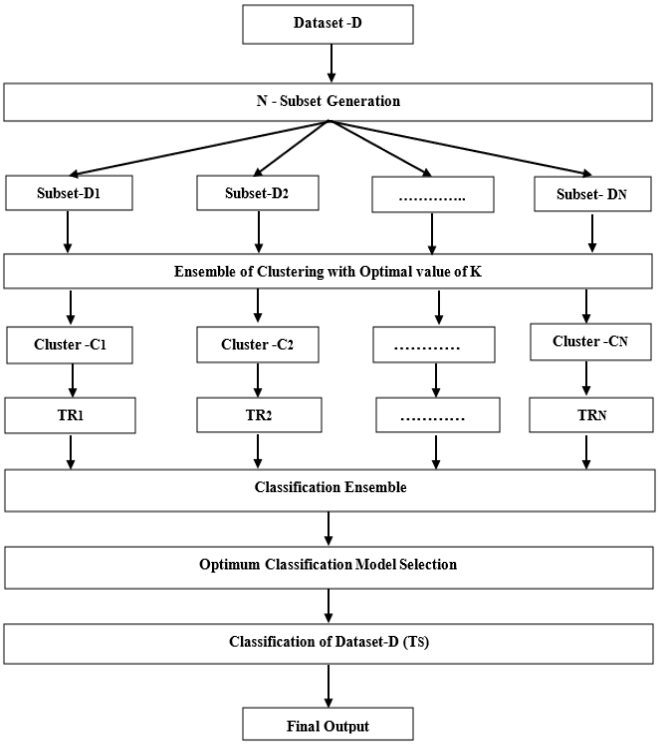


Figure 1. Proposed Ensemble of Clustering and classification.

Algorithm 1: Ensemble of Clustering and Classification (ECC)

Input: Training set (TR), Testing set (TS)

Process:

1: Apply clustering algorithms with optimized value of K on entire dataset (TR+TS) and obtain optimized set of clusters C. // CBES is used to find the optimized value of K.

2: Apply classification algorithms on the training dataset (TR) and find the classification model.

3: **For** each Sample S in TS do

3.1: Cmp (S, Cm) = Classification Output //Membership probability of sample 'S' in class 'Cm'.
Cmp* = Max Cmp (S, Cm) for all existing classes in dataset

3.2: Cm* = arg_max (Cmp (S, Cm)) // Class Membership where maximum probability is achieved.

3.3: **If** Cmp* >= (δ1), Where δ1 is a user-defined threshold
Then assign the labels of class Cm: Cm[S] = Cm*.
Otherwise, Cm[S] = Unlabeled U.

4: **While** Cm [S] == Unlabeled U **do**

4.1: Cs = cluster of S, where Cs ϵ C.

4.2: IF the samples greater than δ2 samples in Cs are labelled with a class Cm
Then Cm[s] = Cm.
Else Cm[S] = Cs.

5: **Return** Cm[S].

Output: Labels in Testing Dataset (TS)

EXPERIMENTAL

The experimental setup comprises the datasets employed, a selection of standalone classifiers, and clustering algorithms whose outputs are amalgamated. Additionally, a set of benchmark methods is utilized to compare against the proposed method

Base Classifiers: Several stand-alone base classifiers such as Decision Tree algorithm (DT),[41] Naive Bayes algorithm (NB),[39] K-nearest neighbour algorithm (KNN),[42] Linear Discriminant Analysis (LDA)[41], and Support Vector Machine (SVM)[42] with linear kernel are a few examples of decision making algorithms. Later, these algorithms are used as independent baseline classifiers in contrast to our ensemble techniques.

Base Clustering Methods: Here are some cutting-edge clustering techniques to consider: K-Means, K-Means with One Hot Encoding, and K-Mode Clustering. The Silhouette method is used to calculate the value of K in K- Means clustering. To achieve optimal performance, the additional parameters of the methods are systematically tweaked.

Baseline ensemble classifiers: The proposed approach contrasts with the already mentioned solo classifiers. And also contrasted them with cutting-edge ensemble classifiers: XGBoost[29] (XGB): a tree boosting approach, bagging, adaptive boosting, gradient boosting (GBST), Random Forest, etc.

RESULTS AND DISCUSSION

This section describes the procedure that was used to select the parameters and then compares the results to the baselines. The procedure that follows explains how it handles unbalanced input into basis solutions and how it depends on base techniques. Additionally, it examines the impact of the components of the objective function on the model's performance, as well as how the runtime of the model is influenced by various dataset characteristics such as the number of objects, classes, and base procedures.

Table 1. Comparison of the AUC of different algorithms with different dataset.

Dataset		Mushroom	US Census	Diabetes	Titanic	Creditcard
Standalone Classifier	**DT**	0.94	0.87	0.67	0.65	0.67
	NB	0.92	0.85	0.60	0.65	0.69
	K-NN	0.94	0.82	0.66	0.66	0.63
	SVM	0.94	0.80	0.69	0.66	0.62
	LDA	0.91	0.79	0.72	0.67	0.67
Ensemble Classifier	**BAG**	0.96	**0.90**	0.65	0.65	0.66
	BOOST	0.97	0.89	0.75	0.66	0.74
	XGBST	0.97	0.87	0.73	0.65	0.74
	GBST	0.94	0.88	0.75	0.69	0.74
	RF	**0.98**	**0.90**	0.77	0.65	0.71
Proposed Ensemble	**ECC**	**0.98**	**0.90**	**0.88**	**0.78**	**0.85**

Performance of ensemble classification methods : AUC and F-Score (F1-Score) are the two metrics used to compare the performance of the competing approaches. The harmonic mean of the precision and recall of a classification model is known as the F1 score, the F measure, or the F score. To ensure that the F1 measure accurately represents the reliability of a mode, the two metrics make equal contributions to the score. The range of values for both measures is 0 to 1, with a higher number indicating greater measurement accuracy. The AUC score and the F-score are used to compare the accuracy of the opposing approaches. Table 1 shows the AUC score for each approach for various datasets. Table 2 shows the F-Score of each approach for various datasets.

Table 2. Comparison of the F1 Score of different algorithms with different dataset.

Dataset		Mushroom	US Census	Diabetes	Titanic	Creditcard
Standalone Classifier	**DT**	0.89	0.83	0.68	0.62	0.64
	NB	0.92	0.86	0.59	0.63	0.67
	K-NN	0.89	0.79	0.64	0.65	0.61
	SVM	0.91	0.78	0.65	0.64	0.59
	LDA	0.89	0.76	0.71	0.68	0.64
Ensemble Classifier	**BAG**	0.94	**0.87**	0.62	0.61	0.69
	BOOST	**0.95**	0.84	0.71	0.64	0.71
	XGBST	0.94	0.82	0.69	0.61	0.70
	GBST	**0.95**	0.85	0.73	0.65	0.72
	RF	**0.95**	0.87	0.74	0.61	0.69
Proposed Ensemble	**ECC**	**0.95**	**0.87**	**0.85**	**0.73**	**0.81**

Performance evaluation plays an important role in machine learning. Area Under the Curve - AUC and Receiver Operating Characteristic - ROC curves are important when it comes to competing classifications. AUC score is one of the most important parameters that determine whether the classification model is valid or not. Another way to say this is "Area Under Receiver Operating Specifications" or AUROC. The AUC-ROC curve is used to evaluate the performance of classification problems based on variables. ROC is the probability curve, while AUC is the index or degree of separation. Explain to what extent the model is capable of discriminating within the class. The larger the AUC, the better the model predicts class 0 as 0 and class 1 as 1.It is noted that no single baseline consistently performs optimally across all datasets. Depending on the datasets, DT, NB, LDA,K-NN, and SVM are the popular baseline classifiers used. The best classification results of each dataset are shown in boldface as shown in Tables 1-2. Here, it is seen that the ECC is the only algorithm that delivers the best performance. Figure 2 and Figure 3 show the comparison of classification AUC Score and F-Score of the existing standalone classifier and the ensemble classifier with the proposed ensemble method.

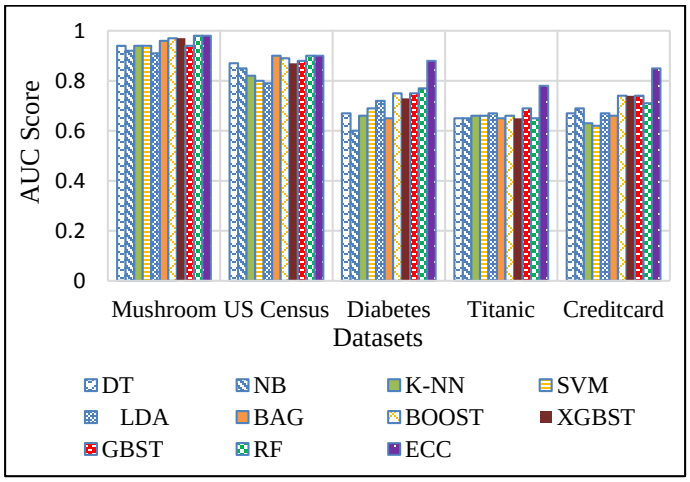


Figure 2. Comparision of AUC Score

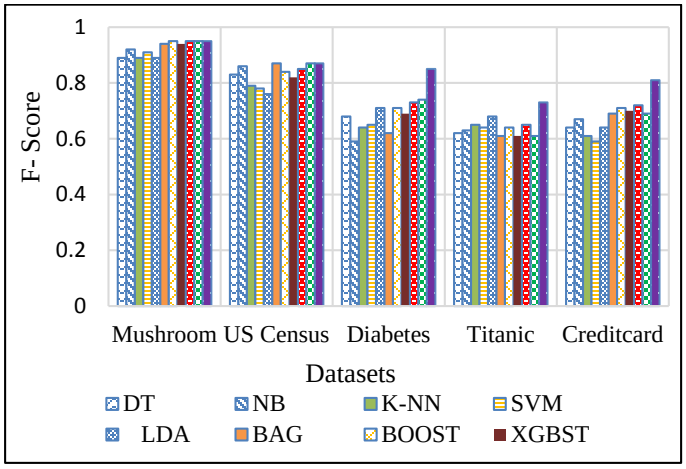


Figure 3. Comparison of the F-Score

Table 3. Summary of the data set from UCI Repository

Datasets	Instances	Attribute	Attribute Type	Classes	Class Distribution
Mushroom	8124	22	Categorical	2	4208,3196
US Census	1458285	68	Categorical	NA	NA
Diabetes	520	17	Categorical	2	320,200
Titanic	891	07	Categorical	2	342,549
Credit Approval	690	15	Mixed	2	307,383

DATASET

The experimental datasets utilized are sourced from the widely recognized UCI machine learning library.[30] Four categorical and one mixed dataset is used for experimentation. In terms of size, number of characteristics, and distribution of items in various classifications, these datasets are widely used and very diverse in nature.[36-43] There are descriptions of fictitious samples that correspond to 23 different types mushrooms in the Mushroom data set.[39,40] The Diabetes dataset comprises 17 variables and sign and symptom data from patients who are recently diagnosed or who are at risk of developing diabetes.[42] The credit approval database contains credit card applications and has a good mix of attributes.[34] The entire available data set samples are used for experimentation except the US Census where only a 1% sample of the 1990 Census is taken from the USCen-sus1990 raw data set due to its large size.[41] The datasets consist of low as wll as high dimension data where the number of categorical attributes vary from 07 to 68. All the above datasets are summarized in Table 3.

CONCLUSION AND FUTURE SCOPE

To ensure the production of more reliable results, the proposed algorithms leverage the complementary constraints provided by a range of classifiers and clustering techniques. The proposed method is further examined for efficiency on 5 different University of California Irvine (UCI) datasets, ECC outperformed against the popular base classifiers like DT, NB,K-NN,SVM, LASSO, LDA by overll 10 %. It also shows the same performance upshift when compared with existing popular ensembles models. The challenges of selecting acceptable base methods and algorithmic parameter choices are also investigated, but which base model set to use to get optimum out-comes are still up for debate. Through correlation research on the machine learning-based technique, it might be able to keep only the most crucial basic models. To make decisions more rapidly, the clustering process is accelerated by a distributed clustering system, which allows one to complete data analysis parallely. Scattered clustering enables computations i.e cluster values to be retrieved from one level to another instead of directly from raw data, leading to a significant reduction in processing time. Various algorithms are available for clustering categorical data. Each has particular benefits and drawbacks. The most popular clustering method has many drawbacks, such as dependence on initialization variables, locally optimal clusters, and starting conditions. The experimental finding demonstrates that the proposed ensemble technique (ECC) optimizes the clustering outcome while overcoming the drawbacks of individual clustering algorithms. Algorithms that work more efficiently should be included in the further clustering ensemble process. Consequently, the robust members of the ensemble clustering process actively engage in subsequent steps, thereby elevating the overall quality of clustering. In the future, the aim is to study selective and increasing member selection methods to find the better ensembles members for performance and robustness of the data analysis process.

REFERENCES AND NOTES

1. M. Anila, G. Pradeepini. Study of prediction algorithms for selecting appropriate classifier in machine learning. *J. Adv. Res. Dyn. Control Syst.* **2017**, 9 (Special Issue 18), 257–268.
2. S. Khedairia, M.T. Khadir. A multiple clustering combination approach based on iterative voting process. *J. King Saud Univ. - Comput. Inf. Sci.* **2022**, 34 (1), 1370–1380.
3. T. Chakraborty, D. Chandhok, V.S. Subrahmanian. MC3: A multi-class consensus classification framework. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* **2017**, 10234 LNAI, 343–355.

4. S.W. Purnami, J.M. Zain, A. Embong. Reduced support vector machine based on k-mode clustering for classification large categorical dataset. *Commun. Comput. Inf. Sci.* **2011**, 180 CCIS (PART 2), 694–702.
5. S. González, S. García, J. Del Ser, L. Rokach, F. Herrera. A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities. *Inf. Fusion* **2020**, 64, 205–237.
6. A. Mayr, H. Binder, O. Gefeller, M. Schmid. The evolution of boosting algorithms: From machine learning to statistical modelling. *Methods Inf. Med.* **2014**, 53 (6), 419–427.
7. Y. Yang, J. Jiang. Hybrid Sampling-Based Clustering Ensemble with Global and Local Constitutions. *IEEE Trans. Neural Networks Learn. Syst.* **2016**, 27 (5), 952–965.
8. Y. Yang. Unsupervised Ensemble Learning and Its Application to Temporal Data Mining: Keynote Address. Manchester, U.K 2021, pp 1–1.
9. Y. Yang, K. Chen. Unsupervised learning via iteratively constructed clustering ensemble. *Proc. Int. Jt. Conf. Neural Networks* **2010**, 1–8.
10. T. Alqurashi, W. Wang. Clustering ensemble method. *Int. J. Mach. Learn. Cybern.* **2019**, 10 (6), 1227–1246.
11. Y. Yang, K. Chen. Temporal data clustering via weighted clustering ensemble with different representations. *IEEE Trans. Knowl. Data Eng.* **2011**, 23 (2), 307–320.
12. T. Chen, C. Guestrin. XGBoost: A scalable tree boosting system. *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.* **2016**, 13-17-Aug, 785–794.
13. G. Pole, P. Gera. A recent study of emerging tools and technologies boosting big data analytics. In *Advances in Intelligent Systems and Computing*; Springer, **2016**; Vol. 413, pp 29–36.
14. M. Hinne, Q.F. Gronau, D. van den Bergh, E.J. Wagenmakers. A Conceptual Introduction to Bayesian Model Averaging. *Adv. Methods Pract. Psychol. Sci.* **2020**, 3 (2), 200–215.
15. J.E. van Engelen, H.H. Hoos. A survey on semi-supervised learning. *Mach. Learn.* **2020**, 109 (2), 373–440.
16. E. Bauer, R. Kohavi. Empirical comparison of voting classification algorithms: bagging, boosting, and variants. *Mach. Learn.* **1999**, 36 (1), 105–139.
17. P.K. Mallapragada, R. Jin, A.K. Jain, Y. Liu. SemiBoost: Boosting for Semi-Supervised Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* 31, 2000–2014.
18. V. Singh, L. Mukherjee, J. Peng, J. Xu. Ensemble clustering using semidefinite programming with applications. *Mach. Learn.* **2010**, 79 (1–2), 177–200.
19. X. Ji, S. Liu, P. Zhao, X. Li, Q. Liu. Clustering Ensemble Based on Sample's Certainty. *Cognit. Comput.* **2021**, 13 (4), 1034–1046.
20. R. Cao, J. Wang, M. Mao. Feature-wise attention-based boosting ensemble method for fraud detection. *Eng. Appl. Artif. Intell.* **2023**, 126, Part C, 106975.
21. L. Breiman. Bagging predictors. *Mach. Learn.* **1996**, 24 (2), 123–140.
22. J. Gao, F. Liang, W. Fan, Y. Sun, J. Han. A graph-based consensus maximization approach for combining multiple supervised and unsupervised models. *IEEE Trans. Knowl. Data Eng.* **2013**, 25 (1), 15–28.
23. X. Ao, P. Luo, X. Ma, et al. Combining supervised and unsupervised models via unconstrained probabilistic embedding. *Inf. Sci. (Ny).* **2014**, 257, 101–114.
24. G.S. Pole, M.A. Potey. A highly efficient distributed indexing system based on large cluster of commodity machines. *IFIP Int. Conf. Wirel. Opt. Commun. Networks, WOCN* **2012**.
25. M. Priya, A. Anitha, M.K. Nallakaruppan, et al. Heart Disease Prediction Using Machine Learning Algorithms. *2023 Innov. Power Adv. Comput. Technol. i-PACT 2023* **2023**, 8 (5), 270–272.
26. Y. Qian, F. Li, J. Liang, B. Liu, C. Dang. Space Structure and Clustering of Categorical Data. *IEEE Trans. Neural Networks Learn. Syst.* **2016**, 27 (10), 2047–2059.
27. M. Halder, S. Shopnil, Y. Arafat, et al. Clustering as a Catalyst for Big Data Classification (CC-BC). In *2023 26th International Conference on Computer and Information Technology, ICCIT 2023*; **2023**; pp 1–6,.
28. D. Yuan, J. Huang, X. Yang, J. Cui. Improved random forest classification approach based on hybrid clustering selection. In *Proceedings - 2020 Chinese Automation Congress, CAC 2020*; Shanghai, China, **2020**; pp 1559–1563.
29. T. Liu. Research on Naive Bayes Integration Method based on Kmeans++ digital teaching clustering. *Proceedings - 2023 8th International Conference on Information Systems Engineering, ICISE 2023*. **2023**, pp 302–305.
30. G. Hu, F. Yang. Selective Hierarchical Ensemble Modeling Approach and Its Application in Leaching Process. In *2015 10th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*; IEEE, **2015**; pp 554–561.
31. J. Yuan, J. Li, J. Hao. A dynamic clustering ensemble learning approach for crude oil price forecasting. *Eng. Appl. Artif. Intell.* **2023**, 123, 106408.
32. H. Shi, Q. Peng, Z. Xie, J. Wang. A semi-supervised hierarchical ensemble clustering framework based on a novel similarity metric and stratified feature sampling. *J. King Saud Univ. - Comput. Inf. Sci.* **2023**, 35 (8), 101687.
33. X. Wu, J. Zhan, W. Ding. TWC-EL: A multivariate prediction model by the fusion of three-way clustering and ensemble learning. *Inf. Fusion* **2023**, 100, 101966.
34. B. Shen, J. Jiang, F. Qian, et al. Semi-supervised hierarchical ensemble clustering based on an innovative distance metric and constraint information. *Eng. Appl. Artif. Intell.* **2023**, 124, 106571.
35. M.M.F. Islam, R. Ferdousi, S. Rahman, H.Y. Bushra. Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques. In *Advances in Intelligent Systems and Computing*; **2020**; Vol. 992, pp 113–125.
36. K. Kousalya, B. Krishnakumar, S. Boomika, N. Dharati, N. Hemavathy. Edible Mushroom Identification Using Machine Learning. In *2022 International Conference on Computer Communication and Informatics, ICCCI 2022*; India, **2022**; pp 1–7,.
37. H. Cao, S. Rawls, P. Natarajan. 1990 US Census Form Recognition Using CTC Network, WFST Language Model, and Surname Correction. In *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*; Kyoto, Japan, **2017**; Vol. 1, pp 977–982.
38. G. Pole, P. Gera. Cluster-Based Ensemble Using Distributed Clustering Approach for Large Categorical Data. *Lect. Notes Networks Syst.* **2021**, 154, 671–680.
39. P.H. Swain, H. Hauska. Decision Tree Classifier: Design and Potential. *IEEE Trans Geosci Electron* **1977**, GE-15 (3), 142–147.
40. T. Mladenova, I. Valova. Comparative analysis between the traditional K-Nearest Neighbor and Modifications with Weight-Calculation. In *ISMSIT 2022 - 6th International Symposium on Multidisciplinary Studies and Innovative Technologies, Proceedings*; Ankara, Turkey, **2022**; pp 961–965.
41. J.H. Xue, P. Hall. Why does rebalancing class-unbalanced data improve AUC for linear discriminant analysis? *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, 37 (5), 1109–1112.
42. O.A. Montesinos López, A. Montesinos López, J. Crossa. Support Vector Machines and Support Vector Regression. *Multivariate Statistical Machine Learning Methods for Genomic Prediction*. Springer 2022, pp 337–378.