# An Adaptive Cluster-Based Ensemble Learner for Computational Biology

**Niti Jain and Ambar Maini**

**Abstract** In quantitative biology, discovering a class when presented with a large bimolecular dataset poses a big problem. However, ensemble learning approach has been helpful in various complex areas of decision-making. So, in this paper, we propose a cluster-based ensemble learner called adaptive cluster-based ensemble learner (ACEL) which incorporates the prior knowledge of the datasets into the cluster ensemble framework. ACEL computes the cluster boundaries using three diverse clustering algorithms to obtain clusters for classification decision. ACEL learns by transforming the obtained clusters into rules and performing adaptive rule tuning to optimize the classification decision. The cluster-based classification results are then processed using majority voting algorithm. The proposed approach is compared with other supervised benchmark algorithms using seven problems from the field of biology. The experiments performed on benchmark datasets show that ACEL works effectively in classifying datasets.

**Keywords** Clustering · Adaptive learner · Ensemble · Supervised · Classification

## 1 Introduction

Adaptive learning is more effective than traditional non-adaptive learning algorithms and is better suitable for large-scale data [3]. In any expert system, before arriving at a conclusion, opinions from all the experts are taken into consideration and then the final decision is made. This is the principle behind ensemble learning [7]. In applications where the size of data is too large for a single classifier to analyse, ensemble

N. Jain (✉) · A. Maini
National Institute of Technology, Raipur, India
e-mail: nitijain305@gmail.com

A. Maini
e-mail: ambarmaini1993@gmail.com

© Springer Nature Singapore Pte Ltd. 2018

S. Bhattacharyya et al. (eds.), *Advanced Computational and Communication Paradigms*, Advances in Intelligent Systems and Computing 706,
https://doi.org/10.1007/978-981-10-8237-5_56

systems partition the data into subsets where each classifier works on a subset of dataset and further combines the results using the existing approaches like majority voting, weighted majority voting, etc. [5]. There are two ways of combining the classifiers: classifier fusion and classifier selection [6]. In classifier fusion approach, all individual classifiers are trained on the whole dataset. Examples of this include bagging predictors and boosting [8]. In classifier selection approach, each individual classifier performs its best in some part of total dataset. There are two major components of any ensemble system. The first component is making a diverse ensemble. The second component is used to combine the output of decisions of the single classifiers.

In the real world, data consist of classes with overlapping boundaries. Excessive training will help solve this problem, but it will result in overfitting which will lead to misclassifications of testing data. Whereas learning generalized boundaries will not lead to overfitting but it will misclassify the overlapping patterns. Therefore, we opt to use clustering. Clustering makes it easy to learn the decision boundaries. Organization of data into groups is one of the fundamental methods of learning.

For each problem, let $x = \{x_1, x_2 \dots x_n\}$ be a set of input vector in $R^p$ and $y = \{y_1, y_2 \dots y_n\}$, for a system given by $S$, where $S$ transmutes $x$ to $y$

$$y = S(x) \tag{1}$$

Here, $x = \{x_1, x_2 \dots x_p\} \in R^p$ is an input vector and $y = \{y_1, y_2 \dots y_r\} \in R^r$ is the output vector of a system. The purpose of our experiment is to identify a classification system that builds $S$ to explain the given input–output data $(x, y)$.

We present adaptive cluster-based ensemble learner (ACEL) in the following sections. Section 2 gives the literature review. Section 3 gives the description about proposed approach. Section 4 summarizes the experimental setup. Section 5 summarizes the results. Section 6 presents the conclusions and the future work in this field.

## 2 Literature Review

Clusters are dense regions which are separated by low-density regions in feature space. Several Bayesian approaches are used for data clustering like undirected graphical model. Ensemble classifier combines the result of various diverse base classifiers [13]. Diversity is a property used to define ensemble classifiers. Greater diversity is observed when incorrect decisions made by one of the classifiers are handled by the other classifiers. This results in uniform distribution of errors. To combine the results of base classifiers, various methods have been proposed including majority voting, weighted majority and decision template. Among the several existing approaches for ensemble learning, boosting and bagging have been used to a greater extent [9, 10]. In bagging, the base classifiers learn on data subsets drawn randomly from entire training set, and the results are combined by majority voting [2]. In boosting, re-sampling of data instances is performed. The new learners work on the instances that are difficult to classify by the previous number of the ensemble. This mechanism encourages the construction of complementary learners.

Lately novel cluster ensemble technique, CE-GMDH was brought forward that comprises three parts: one initial approach, one conveying function and one external condition [11]. Experimentations were performed by CE-GMDH on artificial and real data. Yu et al. [14] has suggested a feature assortment oriented semi-supervised cluster ensembling technique for clustering of tumour obtained from instances of biomolecular datasets. A progressive semi-supervised clustering ensemble technique with arbitrary subspace method, limitation propagation and progressive ensemble member selection technique was brought forward by [15]. Alves et al. [1] has developed a methodology of ensembling by using multiple particle swarm optimization and demonstrated its ability to solve problems of computation biology.

## 3 Proposed Approach

In our proposed approach, we have performed homogeneous clustering for partitioning the patterns belonging to a single class only. Fixed number of rules for each class (here, one rule for each class) is used. Every class is represented by a combination of rules. To generate initial rule base, the training data are clustered using three different categories of clustering method k-means, fuzzy c-means (FCM) and particle swarm optimization (PSO) based clustering algorithm [4]. In order to catch each aspect of data learning process, three different varieties of clustering algorithms which are of different nature and can cluster using different approaches have been used.

Every single cluster represents a thick region in the input dataset which is depicted by the related cluster centroid. Every individual cluster is thereafter transformed into a rule, after which we perform adaptive rule tuning process to minimize the error function. The proposed cluster-based classification produces three diverse base classifiers. These base classifier's results are joined using majority voting algorithm which is used for class prediction. Majority voting technique is used to predict the ultimate classification decision. The majority voting algorithm can be represented as

$$\sum_{t=1}^{T} d_{t,J}(x) = max_{j=1,2,\ldots,c} \sum_{t=1}^{T} d_{t,j} \tag{2}$$

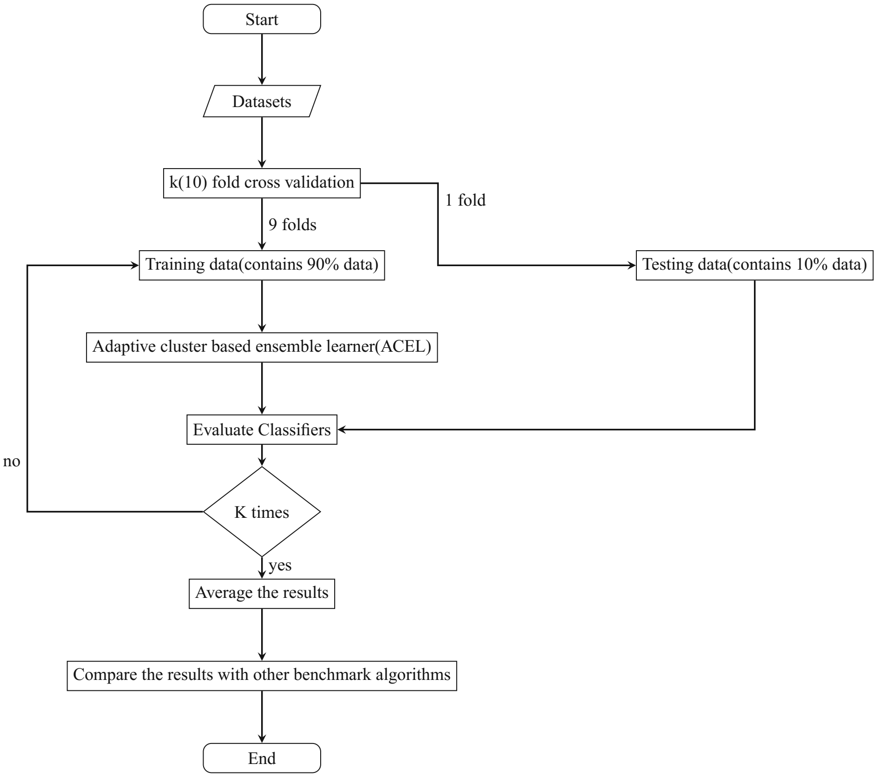

Start
Datasets
k(10) fold cross validation
1 fold
9 folds
Training data(contains 90% data)
Testing data(contains 10% data)
Adaptive cluster based ensemble learner(ACEL)
Evaluate Classifiers
no
K times
yes
Average the results
Compare the results with other benchmark algorithms
End

**Fig. 1** Proposed approach

In situations where individual classifier decisions are not dependent on each other, it can be observed that majority voting combination technique will lead to a performance and accuracy improvement. The classification results are compared with other standard algorithms J48, Adaboost, SMO, Naive Bayes and Random Forest. Figure 1 depicts the proposed approach.

Adaptive learning should be stopped when the training error (or misclassification) reaches an acceptable level. The main goal of rule tuning is to remove irrelevant data associated with the cluster and to include the data which belong to the cluster. Rule tuning potentially increases the predictive power of the rule, helping to avoid overfitting to the training data. As soon as the misclassification reaches to zero or maximum iteration gets completed, the construction of the final rule, i.e. the rule tuning procedure completes.

The strategy for the rule tuning procedure is based on the concept of best centroid for minimizing sum of squared error (SSE), and it can be obtained by the mean of the points in the cluster.

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} (c_i - x)^2 \tag{3}$$

Let $C_i$ be the $i$th cluster, $x$ is a point in $C_i$ and $c_i$ is the mean of the $i$th cluster. In order to find the best centroid which minimizes sum of squared error(SSE) to zero can be performed by the following differentiation for $k$th centroid $c_k$. $m_k$ is the number of objects in $k$th cluster

$$\begin{aligned} \frac{\delta}{\delta c_k} SSE &= \frac{\delta}{\delta c_k} \sum_{i=1}^{K} \sum_{x \in C_i} (c_i - x)^2 \\ \frac{\delta}{\delta c_k} SSE &= \sum_{i=1}^{K} \sum_{x \in C_i} \frac{\delta}{\delta c_k} (c_i - x)^2 \\ \frac{\delta}{\delta c_k} SSE &= \sum_{x \in c_k} 2 * (c_k - x_k) \end{aligned} \tag{4}$$

Equating sum of squared error (SSE) to zero,

$$\frac{\delta}{\delta c_k} SSE = 0 \tag{5}$$

Now combining (4) and (5)

$$\sum_{x \in c_k} 2 * (c_k - x_k) = 0 \tag{6}$$

$$m_k c_k = \sum_{x \in c_k} x_k \tag{7}$$

$$c_k = \frac{1}{m_k} \sum_{x \in c_k} x_k \tag{8}$$

Hence, it can be observed that the best centroid for minimizing the SSE of a cluster is mean of all the points in a cluster.

We have followed the same principle and tuned centre accordingly in order to minimize the misclassification accuracy. The training and prediction methodology of rule tuning for adaptive learning is presented in Algorithm 1.

```
Algorithm: Rule Tuning(C,N) ;
input  : Centres obtained from diverse clustering algorithm C, Number
         of features in the training set N
output: Rule Tuned Centers
// Tuning process is repeated until
   classification error is satisfactory for
   each feature of training set
for $i \leftarrow 1$ to $N$ do
    Compute mean $m_i$ and standard deviation $sd_i$ corresponding to $i^{th}$
    feature of training set;
    Choose tuning parameter $n_m$;
    Set initial value $I_v$ to Minimum ($m_i$-$sd_i$,$c_i$) and Max to (m+$sd_i$);
    Compute the error $E_0^x$ and misclassification $M_0^x$ for initial rule;
    while ($I_v < Max$) do
        $I_v = I_v + n_m$;
        Compute the error for the new rule base;
        if $E_t^x > E_{t-1}^x$ then
            $R^t \leftarrow R^{t-1}$ // since the error is increased,
                we restore the values corresponding
                to the base rule
        end
        if $M_t^x = 0$ or $E_t^x \approx 0$ then
            Stop
        end
        t ← t + 1 ;
    end
end
```

**Algorithm 1:** Rule tuning

## 4 Experimental Setup

In order to validate the proposed approach, the experiments were conducted on the benchmark datasets obtained from the UCI machine learning repository [12]. The ACEL algorithm has been applied on seven benchmark datasets. The datasets used are Iris, Thyroid, Balance Scale, Vertebral Column, Haberman's Survival, Liver Disorder and Diabetes.

**Table 1** Datasets used

| Dataset | Instances | Attributes | Classes |
|---|---|---|---|
| Iris | 150 | 4 | 3 |
| Thyroid | 215 | 5 | 3 |
| Balance Scale | 625 | 4 | 3 |
| Vertebral Column | 310 | 6 | 3 |
| Haberman’s Survival | 306 | 6 | 2 |
| Liver Disorder | 345 | 6 | 2 |
| Diabetes | 768 | 8 | 2 |

The result of the proposed approach has been compared with benchmark algorithms like J48, Adaboost, SMO, Naive Bayes and Random Forest. Table 1 gives a summary of datasets. The experiment was performed in MATLAB.

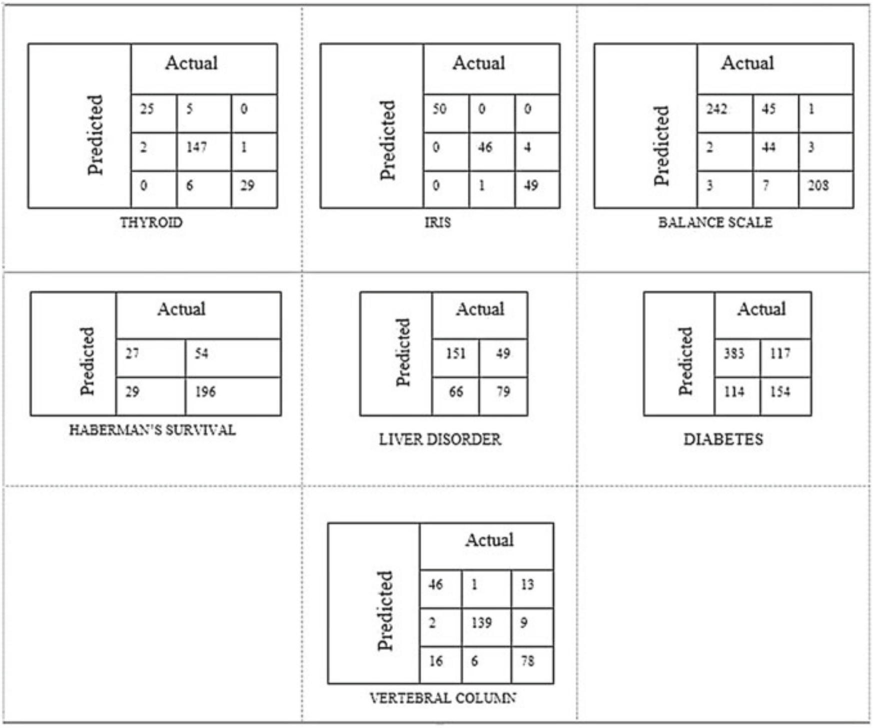
## 5 Results and Discussion

We have trained our learner ACEL on seven biological datasets taken from UCI machine learning repository [12]. To test cluster-based ensemble learner, we compare the classification results by ACEL and by other standard algorithms J48, Adaboost, SMO, Naive Bayes and Random Forest. Table 2 gives the accuracy corresponding to each dataset.

For the Thyroid dataset, ACEL gives 93.46 accuracy, performing similar to other state-of-art algorithms J48, Adaboost and outperforming SMO by almost 5%. For the Iris dataset, ACEL outperformed other algorithms and performed similar to SMO with an accuracy of 96.67. For the dataset Liver disorder, ACEL performs better than Naive Bayes and SMO by at least 15%. Haberman’s Survival dataset using ACEL achieved 72.84 accuracy performing better than Random Forest by at least 8%. The algorithm gives an accuracy of 84.83 for the dataset Vertebral Column, which is

**Table 2** Results in terms of classification accuracy

| Algorithm | Thyroid | Iris | Liver disorder | Haberman’s survival | Diabetes | Vertebral column | Balance scale |
|---|---|---|---|---|---|---|---|
| ACEL | 93.46 | 96.67 | 66.6 | 72.84 | 69.91 | 84.83 | 79.04 |
| J48 | 92.09 | 95.33 | 68.4 | 72.87 | 73.83 | 81.61 | 76.64 |
| Adaboost | 93.48 | 97.33 | 66.66 | 75.16 | 74.35 | 77.42 | 72.32 |
| RF | 95.35 | 94 | 74.49 | 66.99 | 74.349 | 84.19 | 81.6 |
| NB | 96.74 | 95.33 | 54.2 | 74.5 | 76.3 | 83.22 | 90.4 |
| SMO | 89.76 | 96.67 | 57.97 | 73.53 | 77.34 | 74.52 | 87.68 |



| | Actual | | |
|---|---|---|---|
| Predicted | 25 | 5 | 0 |
| | 2 | 147 | 1 |
| | 0 | 6 | 29 |

THYROID

| | Actual | | |
|---|---|---|---|
| Predicted | 50 | 0 | 0 |
| | 0 | 46 | 4 |
| | 0 | 1 | 49 |

IRIS

| | Actual | | |
|---|---|---|---|
| Predicted | 242 | 45 | 1 |
| | 2 | 44 | 3 |
| | 3 | 7 | 208 |

BALANCE SCALE

| | Actual | |
|---|---|---|
| Predicted | 27 | 54 |
| | 29 | 196 |

HABERMAN'S SURVIVAL

| | Actual | |
|---|---|---|
| Predicted | 151 | 49 |
| | 66 | 79 |

LIVER DISORDER

| | Actual | |
|---|---|---|
| Predicted | 383 | 117 |
| | 114 | 154 |

DIABETES

| | Actual | | |
|---|---|---|---|
| Predicted | 46 | 1 | 13 |
| | 2 | 139 | 9 |
| | 16 | 6 | 78 |

VERTEBRAL COLUMN

**Fig. 2** Confusion matrices

better than Adaboost by almost 9% and SMO by more than 12%. In Balance Scale dataset, our proposed learner performs better than J48 and Adaboost algorithm.

The results obtained after performing ACEL algorithm suggest that applying clustering on the datasets and transforming the cluster to rules followed by adaptive tuning on these clusters optimizes the classification decision. The experiments performed on benchmark datasets show that ACEL works effectively in classifying datasets.

Figure 2 shows the confusion matrix when proposed ACEL algorithm is applied to each dataset. The actual and predicted labels give the number of instances that are correctly classified or misclassified. Confusion matrix can be used to describe the performance of a classification model. In confusion matrix, the count of true positive and true negative indicates how well a classification model works. In the figure, we can see that count of TP and TN is high for Iris, Thyroid, Balance Scale, Haberman's Survival and Vertebral Column suggesting high performance of ACEL.

## 6 Conclusions and Future Work

We have presented a novel cluster-based ensemble learner (ACEL) based on the principle of cluster ensemble learning along with rule tuning, which gives better adaptability and improved accuracy. The proposed algorithm has been evaluated on biological experimental datasets. The results of the experiments have shown that our ensemble approach has given comparable results to individual learners. The evidences from the experimental results show that adaptive cluster ensemble learning process using tuning improves accuracy to a greater extent. In our future research, we would like to focus on finding the optimal number of clusters and other critical issues in ensemble classification like integration mechanism and robustness.

## References

1. Alves, P., Liu, S., Wang, D., Gerstein, M.: Multiple-Swarm ensembles: improving the predictive power and robustness of predictive models and its use in computational biology. IEEE/ACM Trans. Comput. Biol. Bioinform. **5963**, 1–1 (2017). https://doi.org/10.1109/TCBB.2017.2691329
2. Dudoit, S., Fridlyand, J.: Bagging to improve the accuracy of a clustering procedure. Bioinformatics **19**, 1090–1099 (2003). https://doi.org/10.1093/bioinformatics/btg038
3. Giotis, I., Petkov, N.: Cluster-based adaptive metric classification. J. Neurocomputing **91**, 33–40 (2012). https://doi.org/10.1016/j.neucom.2011.10.018
4. Jain, A.: Data clustering: 50 years beyond K-means. Pattern Recognit. Lett. **31**, 651–666 (2010). https://doi.org/10.1016/j.patrec.2009.09.011
5. Kittler, J.: Combining classifiers: a theoretical framework. Pattern Anal. Appl. **1**, 18–27 (1998). https://doi.org/10.1007/BF01238023
6. Kuncheva, L., Ll, G., Kingdom, U., Duin, R.: Decision templates for multiple classifier fusion: an experimental comparison. Pattern Recognit. **34**(2), 299–314 (2001). https://doi.org/10.1016/S0031-3203(99)00223-X
7. Polikar, R.: Ensemble based systems in decision making. IEEE Circuits Syst. Mag. **6**(3), 21–45 (2006). https://doi.org/10.1109/MCAS.2006.1688199
8. Quinlan, J.: Bagging, boosting, and C4.5. Proc. Thirteen. Natl. Conf. Artif. Intell. **5**, 725–730 (1996). https://doi.org/10.1023/A:1018054314350
9. Rodrguez, J., Maudes, J.: Boosting recombined weak classifiers. Pattern Recogn. Lett. **29**, 1049–1059 (2008). https://doi.org/10.1016/j.patrec.2007.06.019
10. Skurichina, M., Duin, R.: Bagging, boosting and the random subspace method for linear classifiers. Pattern Anal. Appl. **5**, 121–135 (2002). https://doi.org/10.1007/s100440200011
11. Teng, G., He, C., Xiao, J., He, J., Zhu, B., Jiang, X.: Cluster ensemble framework based on the group method of data handling. Appl. Soft Comput. **43**, 35–46 (2016). https://doi.org/10.1016/j.asoc.2016.01.043
12. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml/
13. Xu, L., Krzyzak, A., Suen, C.: A method of combining multiple classifiers and their applications to handwriting recognition. IEEE Trans. Syst. Man Cybern. **22**, 418–435 (1992). https://doi.org/10.1109/34.368145

14. Yu, Z., Chen, H., You, J., Wong, H., Liu, J., Li, L., Han, G.: Double selection based semi-supervised clustering ensemble for tumor clustering from gene expression profiles. IEEE/ACM Trans. Comput. Biol. Bioinforma. **11**, 727–740 (2014). https://doi.org/10.1109/TCBB.2014.2315996

15. Yu, Z., Member, S., Luo, P., You, J., Wong, H., Leung, H., Wu, S.: Incremental semi-supervised clustering ensemble for high dimensional data clustering. Tkde **28**, 701–714 (2016). https://doi.org/10.1109/TKDE.2015.2499200