Power System Technology
ISSN:1000-3673

Received: 06-11-2024 *Revised: 15-12-2024* *Accepted: 05-01-2025*

Pattern Ensemble Learning Method for Clustering Ensemble using Incremental Genetic-Based Algorithm

Reza Ghaemi
Department of Computer Engineering, Quchan Branch, Islamic Azad University, Quchan, Iran.
r.ghaemi@iauq.ac.ir

Abstract: The clustering ensemble has emerged as a prominent method for improving clustering accuracy of unsupervised classification. It combines multiple partitions generated by different clustering algorithms into a single clustering solution. Genetic algorithms are known as methods with high ability to solve optimization problems including clustering. To date, significant progress has been contributed to find consensus clustering that will yield better results than existing clustering. This paper has proposed an Incremental Genetic-Based Algorithm for Clustering Ensemble (IGCE) to perform the search task, but has replaced its traditional crossover operator with a Pattern Ensemble Learning Method (PEL). Therefore, IGCE-PEL is capable to avoid the problems of clustering invalidity and context insensitivity from the traditional crossover operator of genetic algorithms. IGCEs have been evaluated on twelve benchmark datasets based on different recombination operators used. The experimental results have demonstrated that IGCE using PEL is able to achieve better clustering accuracy when compared with several other existing genetic-based clustering ensemble algorithms.

Keywords: Clustering Ensemble, Genetic Algorithms, Incremental Genetic Algorithms, Ensemble Learning, Clustering Accuracy.

1. Introduction

Knowledge reuse is one of the primary motivations for developing cluster ensembles. In several applications, a variety of cluster partitions for the samples under consideration may already exist. However, still there is a need either to integrate these cluster partitions into a single solution, or uses this information to influence a new cluster solution of these samples [1,2,3]. Several analogous approaches exist in supervised learning as knowledge reuse, but seldom applied to totally unsupervised settings [4,5].

Genetic Algorithms (GAs) are well known methods with high ability to solve optimization problems such as clustering [6,7]. The disadvantages of clustering algorithms

Power System Technology
ISSN:1000-3673
Received: 06-11-2024 *Revised: 15-12-2024* *Accepted: 05-01-2025*

have motivated the application of more powerful search methods such as GAs in the clustering and clustering ensemble. In many studies, standard GAs (generational GAs) and traditional crossover operators are utilized for the clustering and clustering ensemble. Typically, these GAs often have common problems such as the loss of population diversity, clustering invalidity, and context insensitivity [8,9,11]. The above challenges have motivated the application of more heuristic search methods such as GAs, particularly incremental GAs in the clustering ensemble. Numbers of recent studies have demonstrated that the clustering and clustering ensemble using GAs are often able to identify a better clustering solution [10,12].

The traditional crossover operator in genetic-based clustering algorithms suffers from clustering invalidity and context insensitivity, which will significantly degrade the search capability of GAs [7,13]. One of the existing approaches to solve clustering invalidity is by penalizing any unfeasible clustering solutions within the population. Similarly, the context insensitivity may be avoided by removing the recombination operator from the genetic-based clustering algorithms and remaining the mutation operator from perturbing the population [14]. Nonetheless, without the recombination operator, the search capability in GAs will again be significantly weaken.

To against such limitations, the clustering ensemble is viable [15,16]. This approache is based on the premise that the exploratory nature of clustering would benefit from combining the strengths of many individual clustering algorithms. The main goal of clustering ensembles is to improve the overall accuracy through leverages the consensuses of the best features across multiple clustering solutions [7]. For example, in the case of classification, the best feature would be the class label and in the case of regression, the best feature would be the desired value [17]. Whereas the problem of clustering combination bears some traits of a classical clustering problem, it struggles for two major problems including diversity of clustering and consensus function. The major hardship in clustering ensemble is consensus functions and partitions combination algorithm to produce final partition.

This paper has proposed an Incremental Genetic-Based Algorithm for Clustering Ensemble using Pattern Ensemble Learning Method called IGCE-PEL to perform the search task. There are two main phases in IGCE-PEL. In the first phase, IGCE-PEL utilizes a number of clustering partitions as the population. In the second phase, IGCE-PEL combines the clustering partitions generated in the previous phase to yield the best clustering solution.

The remainder of this paper proceeds as the following: Section 2 deliberates on all related works on the incremental genetic-based algorithms for the clustering ensemble using the ensemble learning methods. Section 3 illustrates the methodology of this research and then mechanism of IGCE-PEL and its components is explained in Sections 3.1 and 3.2, respectively. Section 4 presents the experiment results by IGCE-PEL and other IGCEs with

Received: 06-11-2024 Revised: 15-12-2024 Accepted: 05-01-2025

five different recomination opertors on the twelve benchamrk datasets. Finally, Section 5 will conclude the work with indication of future works.

2. Related Works on the Clustering Ensemble

This Section presents the clustering ensemble problem and its steps in Section 2.1 and then introduces the related works to clustering ensemble learning which has focused on the ensemble learning strategies in Section 2.2. Finally, Section 2.3 explains the existing genetic-based clustering ensemble algorithms which has concentrated on genetic operations for clustering ensemle.

2.1. Clustering Ensemble Problem

Because different clustering algorithms exert different results on a dataset, the results of different clustering algorithms can be combined and the final clusters are calculated by the results of the obtained combination [13]. The clustering ensemble is usually a two-staged algorithm. In the first stage, it stores the results of some independent runs of different clustering algorithms such as K-Means. In the second stage, it uses a specific consensus function to find the best clustering solution from the stored results. However, there are some prolems in clustering ensemble such as the loss of population diversity, clustering invalidity, and context insensitivity [17,21,22,23].

There are different types of consensus function including the hypergraph partitioning, voting approach, mutual information, co-association-based functions, and finite mixture model [18,19]. In this paper, we focused on the evidence accumulation method as consensus function. Figure 1 illustrates an overview on the clustering ensemble process [5,24,25].

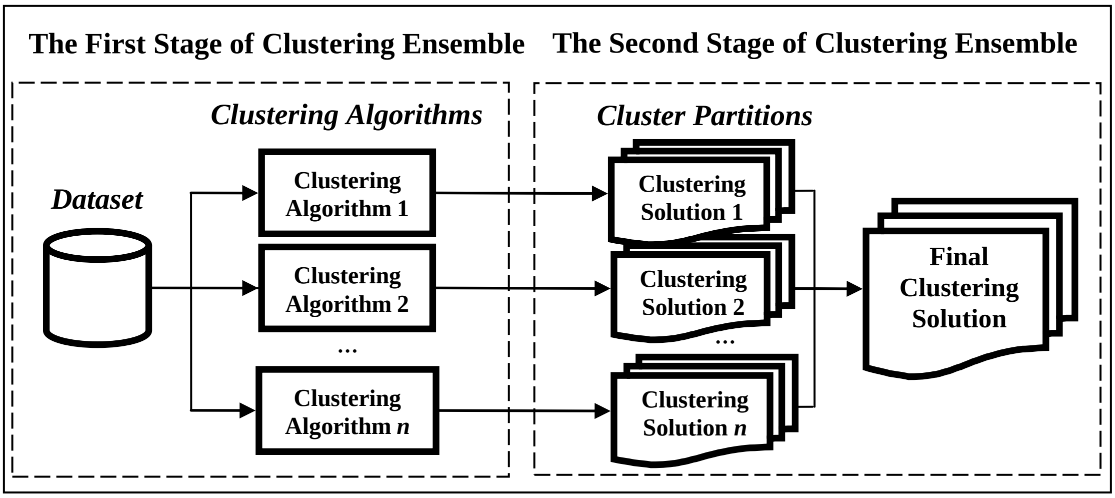


Figure 1: An Overview on the Clustering Ensemble Process.

Power System Technology
ISSN:1000-3673

Received: 06-11-2024 Revised: 15-12-2024 Accepted: 05-01-2025

The major hardship in the clustering ensembles is a combination algorithm to find a consensus clustering solution from the output clustering solutions generated by various clustering algorithms [9,19,25,26]. The problem of combining multiple clustering solutions is determined a single consensus clustering solution. The data clustering is defined as the following combinatory optimization problem. Let $D_{samples} = \{x_1, x_2, ... , x_n\}$ denotes a dataset containing $N_{sample} = n$ unlabeled samples, clustering algorithms work to classify these n samples into k groups, where k is the positive integer numbers, such that the optimal value of a predefined clustering criterion is achieved.

Provided that each sample x_i has m features $x_i = \{x_{i1}, x_{i2}, \ldots, x_{im}\}$, $i = 1, \ldots, n$. Each cluster is denoted by C_k, where $C_k = \{x_1, x_2, \ldots , x_{|Ck|}\}$ is k-th cluster, and $|C_k|$ is the cardinality of the aforementioned cluster, i.e., numbers of samples existing in cluster C_k. The clustering solutions set is denoted by $S = \{S_1, S_2, \ldots, S_{NCandidate}\}$, where $N_{Candidate}$ is number of candidate clustering solutions for combination. Each clustering solution is denoted by $S_i = \{C_1, C_2, \ldots, C_k\}$, where $i = 1, \ldots, N_{Candidate}$, and C_k presents k-th cluster. In the each clustering solution S_i [4,19,25]:

$$\bigcup_{k=1}^{K} C_k = D_{samples} \quad \text{and} \quad C_p \bigcap C_q = \phi \quad , \quad \forall \quad p, \quad q \quad = \quad 1, \quad \ldots, \quad k \qquad (1)$$

Each cluster is denoted by $C_k = \{x_1, \ldots , x_l\}$ as k-th cluster, where $l \in [1,|C_k|]$.

2.2. Clustering Ensemble Learning Methods

It can be clearly that the function of the ensemble learning method is somewhat similar to a recombination operator of GAs that works to aggregate different clustering solutions into a new better one. In the ensemble learning, a more reliable result can be achieved by combining the output of multiple experts. Nonetheless, the commonly used recombination operators of GAs such as the single-point crossover are not able to perform well enough due to the problems of clustering invalidity and context insensitivity. Because of the fact mentioned, the ensemble learning operator is only able to mix string blocks of different chromosomes, but not able to recombine clustering contexts of different chromosomes into new better solutions. The ensemble learning is able to mitigate the problems of the clustering invalidity and the clustering insensitivity [4,9,23].

For the first time, the ensemble learning method has been introduced by Dietterichl [27,28] which refers to a collection of methods that learn a target function by training a number of individual learners and combining their predictions. This ensemble learning method has been defined for the standard supervised learning problem to train samples. There are some effective reviews and book chapters which have investiged ensemble learning, its

Received: 06-11-2024 *Revised: 15-12-2024* *Accepted: 05-01-2025*

methods, applications and challenges. In the book chapter by Yang [23], the ensemble learning techniques have been explored from three aspects including ensemble learning algorithms, combining methods and diversity of ensemble learning. Furthermore, in-depth knowledge about unsupervised ensemble learning has been reviewed by discussing the consensus functions and objective functions of clustering ensemble approaches. Also, in another book chapter, Yang [2] has focused on both ensemble approaches to clustering tasks, and to present a hybrid sampling-based clustering ensemble by combining the strengths of Boosting and Bagging.

The ensemble learning method has been applied in different applications of clustering, for example, an ensemble method for cluster analysis based on dynamic cooperation by Kang and et. al. [29], a center matching scheme for constructing a consensus function in the K-Means cluster ensemble learning by Zhang and et. al. [30], a multiple K-Means clustering ensemble algorithm to find nonlinearly separable clusters by Bai and et. al. [21], and an ensemble clustering framework for categorical data using label information matrix by Yu and et. al. [31], and also the new researches such as, a multiple clustering combination approach based on iterative voting process by Soufiane and Tarek Khadir [24], a clustering framework using agglomerative hierarchical clustering methods based on ensemble approaches and other words, a meta-clustering ensemble scheme applied the bi-weighting policy to solve the model selection associated problem to improve ensemble clustering by Li and et. al. [19], and a self-directed learning framework for cluster ensemble including two models of predicting test set labels and detecting best results, to improve the traditional ensemble framework using assisting the consensus function in achieving the highest assessment of clustering performance by Kadhim and et. al. [25].

In light of the fact that Evidence Accumulation Clustering (EAC) can cluster data for arbitrary shapes and numbers of clusters, hence Wong and Tsuchiya [20] has been presented a variant of EAC using combinations of features to better cluster data. Proposed method on the existing EAC algorithm has been built by populating the clustering ensemble with clusterings based on combinations of fewer features than the original dataset at a time. Next, proposed method has been called in ording to pre whitening the recombined data and weighting the influence of each individual clustering by an estimate of its informativeness.

In view of the fact not be sufficiently of prior knowledge in the most semi-supervised ensemble clustering algorithms, a semi-supervised hierarchical ensemble clustering framework has been suggested by Shi and et. al. [32] based on a novel similarity metric and stratified feature sampling. Their algorithm has utilized the information of all primary partitions according to their strength to calculate the similarity between samples and it has been equipped with a stratified feature sampling mechanism that has been able to improve the diversity of primary partitions and has dealed with high-dimensional data. Primary partitions

Received: 06-11-2024 *Revised: 15-12-2024* *Accepted: 05-01-2025*

have been generated based on multiple hierarchical clustering techniques, and the target partition has been configured by a consensus function based on the clusters clustering policy.

Kaufman [33] has defined the notion of a mutation invariant function on a cluster ensemble with respect to a group action of the cluster modular group on its associated function fields. In other research by Aktaş and et. al. [34], a multi-objective optimization-based solution framework has been designed to produce consensus solutions. Proposed algorithm has selected representative clustering solutions from the preprocessed library with respect to size, coverage, and diversity criteria and has combined them into a single consensus solution, for which the true number of clusters has been assumed to be unknown. In a research by Shan and et. al. [35], an innovative and robust fuzzy self-consistent clustering ensemble model has been introduced to consider the scalable dummy variable representation of base clustering results as a novel feature attributes intrinsic to the original dataset. A fuzzy operator has been formulated, enabling the adjustment of coupling strength contingent upon the uncertainties inherent in practical problems.

Golalipour and et. al. [36] have presented a cluster ensemble selection method based on maximum quality-maximum diversity, in which two important factors including diversity and quality have been considered. By removing irrelevant and redundant clustering, diversity and quality have been increased simultaneously and also, based on the minimum redundancy-maximum relevance (mRMR) criterion, pair-wise and non-pair-wise methods have been proposed. In other research by Chakraborty and et. al. [37], considering the importance to solve problems of clustering articles' citation trajectories and citation time series due to their non-linear and non-stationary characteristics, a feature-based multiple K-Means cluster ensemble framework has been proposed, where multiple learners have been trained for evaluating the credibility of class labels, unlike single clustering algorithms.

A two-stage clustering ensemble algorithm applicable to risk assessment of railway signaling faults has been presented by Chang and Shiwu [38], in which knowledge graph modeling has constructed a connected network of hazard/fault events. The event information has been transformed with text for risk level prediction. In addition, text clustering technology has been utilized to intelligently divide the entity short text data set in the knowledge graph, assign standardized entity names to the cluster partitions, and then complete the calculation and analysis according to the characteristic parameter formula, greatly reducing the labor and time consumption of data annotation and approximate text repetitive processing.

2.3. Genetic-based Clustering Ensemle Algorithms

GAs are a class of heuristic search methods that loosely mimic the behavior of Darwinian evolution for solving large-scale complex optimization problems [8]. Major steps of GAs

Power System Technology
ISSN:1000-3673
Received: 06-11-2024 Revised: 15-12-2024 Accepted: 05-01-2025

include three genetic operators including the selection operator, the mating operator, and the mutation operator. GAs work with these three operators to explore and exploit the coded search space of the objective function (fitness function) [10,39].

Approaches using GA can be classified broadly into two basic categories, which are the generational GAs (standard GAs) and the incremental GAs (steady-state GAs) [40,41]. The first category is original version of GAs which uses typical parameters such as roulette selection, elitism, and generational replacement, where the entire population is replaced at each iteration. This is a method by which the fittest potential parents are selected from a population. However, this does not guarantee that the fittest member proceeds to the next generation [10]. In the generational GAs, offspring generated in each generation will replace population in the same generation. This causes to lose population diversity at a very fast rate due to converge to a local optimal clustering solution [9,14].

The second method is the incremental GAs that select two individual parents (sometimes all individuals) are selected [10] and individual parents are combined by algorithm to produce one offspring, thereby replacing the worst characteristics of a population with better characteristics. Unfortunately, the incremental GAs have the potential of premature convergence when convergence happens too early [8,9,10,39]. The major difference between the incremental and the generational GAs is that, for each parent of the population generated in the generational GA, there are two parents selected by the incremental GA. Combining the strengths of the various methods counteracts the weaknesses of each clustering system [39]. In the incremental GAs, population in successive two iterations significantly overlap and only one or two candidate clustering solutions are replaced at each generation. Therefore, the incremental GAs have a better performance for maintaining the diversity of the population and are more suitable for solving the problem of data clustering [10,39].

Numbers of clustering algorithms exist so far and their clustering solutions may be significantly different. There are several most basic approaches for combining multiple clustering results which have been introduced by Rogers and Prügel-Bennett [41], Dempster and et.al. [42], and Fred and Jain [43]. Many studies have been addressed by Hong and Kwong [9], and Hong and et. al. [26] to need into a GA-based algorithm for clustering using suitable recombination operators.

Due to this point, some researchers have proposed various GA-based algorithms for clustering, especially for clustering ensemble and they have utilized GA's search strategies in their papers. For example, a multi-objective GA-based clustering ensemble algorithm and its operators including a special mutation and one-point crossover, in combination with co-association consensus function by Azimi and Mohammadi [6], a GA-based ensemble learning for detecting community structure in complex networks with a multi-individual crossover operator based on ensemble learning by He and et. al. [40], a multi-objective GA-based

Received: 06-11-2024 *Revised: 15-12-2024* *Accepted: 05-01-2025*

clustering ensemble algorithm by Chatterjee and Mukhopadhyay [8], and a metaheuristic-based clustering ensemble method using an improved generation mechanism and a co-association matrix by Kuo and et. al. [46]. In the following and due to the large volume of articles, only the last few years researches and also, approaches that have compared with proposed GA-based clustering ensemble algorithm in this article have been reviewed.

A data clustering algorithm has been presented by Hong and Kwong [9], that has combined the incremental GA and the ensemble learning method. It has generated its population of candidate clustering solutions by using the random subspaces method. The average-linkage agglomerative clustering algorithm has been employed as clustering algorithm that has yielded a new clustering solution in the two steps based on the evidence accumulation method. First, the each clustering solution has transformed into a similarity matrix. Then all similarity matrices have been combined into a single consensus similarity matrix [44,45] using the evidence accumulation method as consensus function. Second, a new similarity matrix has been sampled from the above similarity matrices. A random number in the range of [0, 1] has been generated and the ensemble learning method has been oriented towards to generate random numbers, in such a way that the new similarity matrix corresponding to the new clustering solution has been constructed on basis of comparing between the frequency of two samples and a random value.

Considering to need automatic clustering to detect the appropriate clustering without a pre-defined number of clusters, Zhu and et. al. [47] offered enhanced an evolutionary multi-objective automatic clustering with quality metrics and ensemble strategy. They have resorted to quality metrics and ensemble strategy for the sake of explicit/implicit knowledge discovery. Quality and diversity of solutions have been defined in terms of cluster validities, as similar to performance indicator for multi-objective optimization, have been applied to assist in addressing automatic clustering problems and decreasing unnecessary computational overhead. Main components like initialization, reproduction operations, and environmental selection have been discussed and refined which involved during evolutionary multi-objective based automatic clustering. In addition, both quality metrics and cluster ensemble strategy have been considered for the determination of the final partitioning, to improve the retrieve system in the unsupervised way.

Yang and et. al. [12] proposed a hybrid genetic model for clustering ensemble which regarding of each base clustering as a new attribute of data, and the result of clustering ensemble evaluated by the objective function. In addition, proposed model could be inferred with the optimization, combination, and transcendence of base clustering results step by step, which has made it possible to maintain the diversity of population and provides more possibilities to avoid falling into local optimal solution. In other research by Hu and et. al. [48], with respect to problem to identify novel molecular subtypes to guide patient selection,

Received: 06-11-2024 Revised: 15-12-2024 Accepted: 05-01-2025

a multi-omics consensus ensemble clustering has offered for Molecular classification which has revealed the diverse genetic and prognostic features of gastric cancer.

Kordos and et. al. [49] investigated main difficulties and challenges in GA-based instance selection, which have high computational complexity and decreasing performance with the dataset size growth this has been caused by the fact that each instance has encoded in one chromosome position. The main contribution of this paper has addressed the above problems in a three-step process, hence, fuzzy clustering decomposition of genetic algorithm-based instance selection has proposed for regression problems. In the first step, the dataset has divided into several consistent regions by fuzzy clustering. Then GA-based instance selection has performed independently within each cluster. Finally, ensemble voting has provided seamless aggregation of the partial results from the overlapping clusters.

With respect to this fact that a small number existence of researches in the field co-clustering ensemble methods on basis of genetic models, in which fuzzy clustering and hard clustering have been combined, Zhong and et. al. [50] putted forward a multi-objective genetic model for co-clustering ensemble and and has been designed the corresponding objective function to process fuzzy samples and general samples more appropriately.

As findings from literature review, it can be mentioned that the main problem is performance of genetic-based clustering ensemle algorithms in term of accuracy which is needed to improve yet and it has special importance for researchers in this field. One additional reason is the need to explore the scalability of genetic-based clustering ensemble algorithms. Understanding how these algorithms perform and scale with various and larger datasets and more complex clustering tasks can provide valuable insights for improving their applicability in real-world scenarios.

3. Methodology

An important result gained by analyzing related works is that, avoiding the loss of population diversity, the clustering invalidity, and the context insensitivity are the basic idea to implement an incremental GA-based clustering ensemble algorithm using ensemble learning methods as recombination operator. Based on the above discussions, an architecture shown in Figure 2 for the clustering ensemble is proposed based on the incremental GA-based clustering ensemble algorithms that consist of two main phases of generation process and combination process.

Received: 06-11-2024 Revised: 15-12-2024 Accepted: 05-01-2025

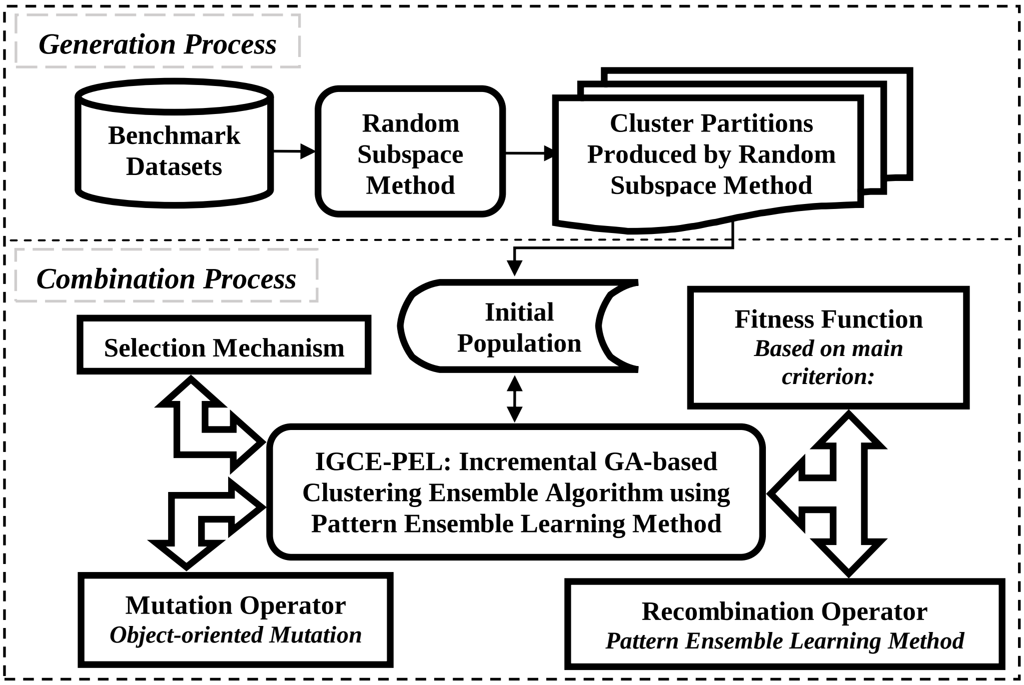


Figure 2: An Architecture for the Incremental GA-Based Clustering Ensemble Algorithm.

First, the generation process includes producing cluster partitions, as inaitial population for the Incremental GA-based Clustering Ensemble algorithm, using the random subspaces method which is able to create diverse cluster partitions by using various feature subsets to train them [51,52]. Since random subspaces method has simplicity and is mostly common methods [9,14], it is utilized to generate diveres cluster partitions in this paper. Second, the combination process contains combining cluster partitions belonging to initial population to discover the final clustering solution by the proposed the Pattern Ensemble Learning method, named IGCE-PEL as follow.

In the combination process, the each cluster partition is employed as initial population by the incremental GA-based clustering ensemble algorithm. A fitness function is utilized as objective function by the incremental GA-based clustering ensemble algorithm in which the compactness as main clustering criterion is applied to evaluate the fitness of clustering solutions using dissimilarity measure of total within-cluster variation. Moreover, the incremental GA-based clustering ensemble algorithm uses the pattern ensemble learning method as the proposed recombination operator. For the mutation, an object-oriented mutation operator is put to used by the incremental GA-based clustering ensemble algorithms. The proposed incremental GA-based clustering ensemble algorithm and the proposed pattern ensemble learning method are introduced in Sections 3.1 and 3.2, respectively.

Power System Technology
ISSN:1000-3673

Received: 06-11-2024 *Revised: 15-12-2024* *Accepted: 05-01-2025*

3.1. Incremental GA-based Clustering Ensemble Algorithm

The incremental GA-based clustering ensemble algorithms are developed using the basic steps of the incremental GAs and deals with five procedures as illustrated in Figure 3 (Algorithm 1), which include fitness evaluation, selection, recombination, mutation, and reinsertion. The incremental GA-based clustering ensemble algorithms first starts in by initializing population $P(t_g)$, where $t_g = 0$.

Algorithm 1 : Incremental Genetic-based Clustering Ensemble Algorithms

```
--------------------------------------------------------------------------------------------------------------
--
$t_g = 0$;
Initialize $P(t_g)$;
For all clustering solutions existing in population $P(t_g)$
    **Fitness Evaluation Procedure** (input: $p^i(t_g)$ ; output: $Fp^i(t_g)$)
        Compute fitness $p^i(t_g) \rightarrow Fp^i(t_g)$;
While termination condition = true
    **Selection Procedure** (input: $P(t_g)$ ; output: $MP(t_g)$)
        Select candidate solutions from $P(t_g)$ by the tournament selection $\rightarrow MP(t_g)$;
    **Recombination Procedure** (input: $MP(t_g)$ ; output: $p^{new}(t_g)$)
        Recombine $MP(t_g)$ using SPC, TPC, EL, EA, CCE, PEL $\rightarrow p^{new}(t_g)$;
    **Mutation Procedure** (input: $p^{new}(t_g)$ ; output: $p^{mutate}(t_g)$)
        Mutate $p^{new}(t_g)$ using the object-oriented mutation $\rightarrow p^{mutate}(t_g)$;
    **Fitness Evaluation Procedure** (input: $p^i(t_g)$ ; output: $Fp^i(t_g)$)
        Compute fitness $p^i(t_g) \rightarrow Fp^i(t_g)$;
    **Reinsertion Procedure**
        If fitness $p^{new}(t_g)$ and $p^{mutate}(t_g)$ < fitness of each $p^i(t_g) \in P(t_g)$
            Replace worst solutions $p^{worst}(t_g) \in P(t_g)$ with $p^{new}(t_g)$ and $p^{mutate}(t_g)$ into $P(t_g + 1)$;
$t_g = t_g + 1$;
```

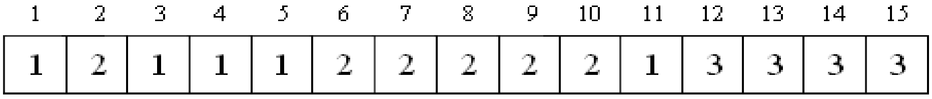
Figure 3: Incremental GA-based Clustering Ensemble Algorithms.

Genotype in chromosomes of incremental GA-based clustering ensemble algorithms are encoded to an array with length n, where n is number of samples and K is number of cluster (Genes' values), as shown in Figure 4 with n=15 and K=3. Second, the fitness of all clustering

Power System Technology
ISSN:1000-3673

Received: 06-11-2024 Revised: 15-12-2024 Accepted: 05-01-2025

solutions in the existing initial population are calculated during the fitness evaluation procedure, where $i = 1, \ldots, N_{population}$. Number of clustering solutions $p^i(t_g)$ is denoted by $N_{population}$. This procedure takes in a clustering solution $p^i(t_g)$ as input and returns the fitness of clustering solution $Fp^i(t_g)$ as output. After the population initialization and the fitness evaluation, the genetic cycle is started and is continued as long as the termination criterion remains true.



1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	2	1	1	1	2	2	2	2	2	1	3	3	3	3

Figure 4: Chromosomes encoding in incremental GA-based clustering ensemble algorithms.

Third, the selection procedure opts for the candidate clustering solutions by the tournament selection with a certain specified rate of selection and then, these candidate clustering solutions are placed in the mating pool $MP(t_g)$. Fourth, the recombination procedure takes in the current mating pool $MP(t_g)$ as input and returns the generated offsprings as output. Actually, the candidate clustering solutions existing in the mating pool $MP(t_g)$ are recombined by one of the six different strategies of recombination operators including Single-Point Crossover (SPC), Two-Point Crossover (TPC), Ensemble Learning method (EL) [9], Evidence Accumulation (EA) [20], Co-Clustering Ensemble (CCE) [50], and Pattern Ensemble Learning (PEL) as proposed method in this article. It should be mentioned in this paper has tried to utilize a few common traditional recombination operators such as SPC and TPC, and also new ensemble learning methods. As the result, one or more offsprings will be generated by the recombination procedure as new clustering solution $p^{new}(t_g)$.

As in Section 2.3 mentioned, EL proposed by Hong and Kwong [9] is an incremental GA-based clustering ensemble learning algorithm, with generated population by using the random subspaces method, which has been utilized the average-linkage agglomerative clustering algorithm on basis of the evidence accumulation method. In addition, EA suggested by Wong and Tsuchiya [20] is an evidence accumulation clustering algorithm, with combinations of fewer features, using pre whitening the recombined data and weighting the influence of each individual clustering. Furthermore, CCE offered by Zhong and et. al. [50] is a multi-objective GA-based co-clustering ensemble algorithm, which the processing of fuzzy samples and general samples have been applied as objective function, where chromosomes have been encoded as the membership.

Fifth, the mutation procedure executes the object-oriented mutation operator on the offsprings. Each new offspring and each mutated clustering solution are denoted by $p^{new}(t_g)$ and $p^{mutate}(t_g)$, respectively. Sixth, the fitness of each new offspring $p^{new}(t_g)$ and each mutated

Received: 06-11-2024 Revised: 15-12-2024 Accepted: 05-01-2025

offspring $p^{mutate}(t_g)$ are calculated by the fitness evaluation procedure. Finally, the reinsertion procedure compares the fitness of the new offsprings $p^{new}(t_g)$ and the mutated offsprings $p^{mutate}(t_g)$ with the fitness of the clustering solutions $p^i(t_g)$ existing in the current population $P(t_g)$ and then, the new offsprings $p^{new}(t_g)$ and the mutated offsprings $p^{mutate}(t_g)$ are replaced with a few of the worst clustering solutions $p^{worst}(t_g)$ existing in the current population $P(t_g)$. Finally, GA is moved to next generation, where $t_g = t_g + 1$.

3.2. Pattern Ensemble Learning Method

The pattern ensemble learning method (PEL) is proposed as a novel recombination operator that is applied by IGCE. Figure 5 presents the pattern ensemble learning procedure (Algorithm 2). As shown, number of $N_{Ensemble}$ individual parents including $\{p^1, \ldots, p^{NEnsemble}\} \in MP(t_g)$ as algorithm's input, and one new clustering solution $p^{(new)}(t_g)$ as algorithm's output have been considered.

```
Algorithm 2 : Pattern Ensemble Learning Procedure
------------------------------------------------------------------------------------------------------------------------
-
Input: $N_{Ensemble}$ individual parents $\{p^1, \ldots, p^{NEnsemble}\} \in MP(t_g)$
Output: one new clustering solution $p^{(new)}(t_g)$
Begin
  For each individual parents $p^e(t_g) \in MP(t_g)$ do
     Transform into a similarity matrix $SM^{(e)}$;
  Compute $SM_{avg} = (SM^{(1)} + \ldots + SM^{(NEnsemble)}) / N_{Ensemble}$;
  For each sample $x_i$ belonging to dataset do $(i = 1, \ldots, n)$
     $pat^{(i,z)} = \{(x_i, x_j) \mid (x_i, x_j) \in SM_{avg}$: maximum evidence, $\forall\, i, j = 1, \ldots, n, i \neq j\}$;
     $PTN^{(i)} = \{pat^{(i,1)}, \ldots, pat^{(i,z)}\}$;
     If $(x_i, x_j) \in PTN^{(i)} \mid PTN^{(j)}$ then $SM^{(new)}(x_i, x_j) = 1$ else $SM^{(new)}(x_i, x_j) = -1$;
     If $(x_i, x_j)$ has minimum evidence then $SM^{(new)}(x_i, x_j) = 0$ else $SM^{(new)}(x_i, x_j) = -1$;
  For each paired-sample $SM^{(new)}(x_{j1}, x_{j2}) = -1$ do $(j1, j2 = 1, \ldots, n)$
     $PTN^{(j1)} = \{(x_{j1}, x_{j1,1}), \ldots, (x_{j1}, x_{j1,z1})\}$;
     $PTN^{(j2)} = \{(x_{j2}, x_{j2,1}), \ldots, (x_{j2}, x_{j2,z2})\}$;
     $x_{j1,neighbors} = \{x_{j1,1}, \ldots, x_{j1,z1}\}$;
     $x_{j2,neighbors} = \{x_{j2,1}, \ldots, x_{j2,z2}\}$;
  For all paired-samples $\{(x_{j1}, x_{j2,1}), \ldots,(x_{j1}, x_{j2,z2})\}$ and $\{(x_{j2}, x_{j1,1}), \ldots,(x_{j2}, x_{j1,z1})\}$ do $(j1, j2 = 1, \ldots, n)$
     $Evidences_{neighbors}(x_{j1}, x_{j2}) = \frac{\sum_{w=1}^{z2} Evidences(x_{j1}, x_{j2,w}) + \sum_{w=1}^{z1} Evidences(x_{j2}, x_{j1,w})}{z1 + z2}$;
     If rand(1) < $Evidences_{neighbors}(x_{j1}, x_{j2})$ then $SM^{(new)}(x_{j1}, x_{j2}) = 1$ else $SM^{(new)}( x_{j1}, x_{j2}) = 0$;
  Transform $SM^{(new)}$ into a distance matrix $DM^{(new)} = 1 - SM^{(new)}$;
  Execute the average-linkage clustering algorithm on the $DM^{(new)}$;
     Generate the one new clustering solution $p^{(new)}(t_g)$;
```

Figure 5: Pattern Ensemble Learning Procedure.

Power System Technology
ISSN:1000-3673

Received: 06-11-2024 *Revised: 15-12-2024* *Accepted: 05-01-2025*

Notation: PEL utilizes the evidence accumulation clustering method [20] as consensus function that works to reproduce one new clustering solution $p^{new}(t_g)$ through combining $N_{Ensemble}$ individual parents $MP(t_g) = \{p^1, \ldots, p^{NEnsemble}\}$ in the several major steps, without accessing features of the samples. The similarity measure between samples is mapped into a similarity matrix *SM* that is a two-dimensional matrix.

Organization: PEL works to generate one new clustering solution $p^{new}(t_g)$ through combining $N_{Ensemble}$ individual parents $MP(t_g)$ in the three major steps. In the first step, taking the co-occurrences of the paired-samples in the same cluster as votes for their association, the each clustering solution $p^{(e)}(t_g)$ is transformed into an $n \times n$ similarity matrix $SM^{(e)}$ [9,44,45] by Equation (2):

$$SM^{(e)}(x_{j1}, x_{j2}) = \begin{cases} 1 & \text{if } p^{(e)}_{j1} = p^{(e)}_{j2} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

in such a way that $p^{(e)}_j(t_g)$ represents the cluster label in which the sample x_j is classified in the e-th clustering solution, where $e = 1, \ldots, N_{Ensemble}$, $j = 1, \ldots, n$, and n is the number of samples in dataset. In other words, a similarity measure between samples induced by inter-pattern relationship is mapped into a two-dimensional matrix. Accordingly, $SM = \{SM^{(1)}, \ldots, SM^{(NEnsemble)}\}$ can be yielded by $N_{Ensemble}$ clustering solutions. The elements of $SM^{(e)}(x_{j1}, x_{j2})$ represent the frequency that the samples x_{j1} and x_{j2} are classified into the same cluster in the e-th individual parent. Next, all the similarity matrices are aggregated into a single consensus similarity matrix $SM_{avg}(x_{j1}, x_{j2})$ by Equation (3):

$$SM_{avg}(x_{j1}, x_{j2}) = \frac{\sum_{e=1}^{N_{Ensemble}} SM^{(e)}(x_{j1}, x_{j2})}{N_{Ensemble}} \quad (3)$$

In this study, the average frequency that the samples x_{j1} and x_{j2} are classified into the same cluster in all the individual parents $MP(t_g)$ is called evidence corresponding to the paired-sample (x_{j1}, x_{j2}).

In the second step, a new similarity matrix $SM^{(new)}$ is initialized and then sampled using the above similarity matrix SM_{avg}. In the beginning of the second step, PEL uses shared patterns corresponding to each sample where these shared patterns are repeated in all or at least more the individual parents $MP(t_g)$. In other words, the shared patterns corresponding to each sample are the paired-samples which have maximum evidence in the similarity matrix

Power System Technology
ISSN:1000-3673

Received: 06-11-2024 *Revised: 15-12-2024* *Accepted: 05-01-2025*

SM_{avg}. These paired-samples with maximum evidence are appropriate shared patterns that are repeated in more ensemble members. In view of the fact that it is possible that several maximum evidence with equal values are assigned to each sample, each sample may possess several shared patterns. Therefore, there is a shared patterns set per each sample including a number of paired-samples that is denoted by $PTN^{(i)} = \{pat^{(i,1)}, \ldots, pat^{(i,z)}\}$, where $i = 1, \ldots, n$ and z is an integer positive number. In addition, $pat^{(i,z)}$ denotes the z-th shared pattern corresponding to the i-th sample that possesses maximum evidence. It is clear that $pat^{(i,1)}, \ldots, pat^{(i,z)}$ possess the same maximum evidence among the other evidences corresponding to the i-th sample. Indeed, the each shared pattern is defined by Equation (4):

$$pat^{(i,z)} = \{(x_i,\ x_j) \mid (x_i,\ x_j) \in SM_{avg} : \text{maximum evidence},\ \forall\ i,\ j = 1,\ \ldots,\ n,\ i \neq j\} \quad (4)$$

It is important that the appropriate shared patterns should not be ignored in the new similarity matrix $SM^{(new)}$. Therefore, to construct the new similarity matrix $SM^{(new)}$, all the detected paired-samples as the shared patterns corresponding to the each sample are assigned by 1 in the new similarity matrix $SM^{(new)}$ by Equation (5) as:

$$SM^{(new)}(x_{j1}, x_{j2}) = \begin{cases} 1 & \text{if } (x_i, x_j) \in PTN^{(i)} \text{ or } PTN^{(j)} \\ -1 & \text{otherwise} \end{cases} \quad (5)$$

where $i, j = 1, \ldots, n$. On the other hand, the unsuitable shared patterns are the paired-samples that possess minimum evidence in such a way their corresponding elements in the similarity matrix SM_{avg} are usually valued by 0. Hence, also it is significant that the unsuitable shared patterns should be removed in the new similarity matrix $SM^{(new)}$. It leads to construct the new similarity matrix $SM^{(new)}$ which in all the paired-samples with minimum evidence are assigned by 0 using Equation (6) as:

$$SM^{(new)}(x_{j1}, x_{j2}) = \begin{cases} 0 & \text{if } (x_{j1}, x_{j2}) \text{ has minimum evidence} \\ -1 & \text{otherwise} \end{cases} \quad (6)$$

If the paired-sample $(x_i,\ x_j)$ do not possess maximum or minimum evidence, the element corresponding to (x_i, x_j) in the new similarity matrix $SM^{(new)}$ is assigned by -1.

Received: 06-11-2024 Revised: 15-12-2024 Accepted: 05-01-2025

In the third step, the elements of the new similarity matrix $SM^{(new)}$ that have assigned by -1 from previous step, should be reassigned again. This step works on the basis of neighbor samples with two samples x_{j1} and x_{j2} that exist in the shared patterns, then this is indicative of similarity between two sample x_{j1} and x_{j2} on the basis of their neighbor samples. In other words, to assign the each paired-sample (x_{j1} , x_{j2}) that have been valued by -1 in the new similarity matrix $SM^{(new)}$, the evidences of all the neighbor samples belonging to two samples x_{j1} and x_{j2} are considered that exist in the shared patterns, where $j1, j2 = 1, \ldots, n$. Firstly, all the existing shared patterns corresponding to the samples x_{j1} and x_{j2} are considered that are denoted by $PTN^{(j1)} = \{(x_{j1}, x_{j1,1}), \ldots, (x_{j1}, x_{j1,z1})\}$ and $PTN^{(j2)} = \{(x_{j2}, x_{j2,1}), \ldots, (x_{j2}, x_{j2,z2})\}$, respectively. Simply put, two subsets $x_{j1,neighbors} = \{x_{j1,1}, \ldots, x_{j1,z1}\}$ and $x_{j2,neighbors} = \{x_{j2,1}, \ldots, x_{j2,z2}\}$ denote the neighbors of the samples x_{j1} and x_{j2} in the existing shared patterns, respectively. It can be seen that the samples x_{j1} and x_{j2} possess $z1$ and $z2$ shared patterns, respectively. Secondly, the evidences of sample x_{j1} with the neighbors of sample x_{j2} ($x_{j2,neighbors}$) belonging to the above shared patterns including $\{(x_{j1}, x_{j2,1}), \ldots,(x_{j1}, x_{j2,z2})\}$ are computed by the similarity matrix SM_{avg}. Similarly, the evidences of sample x_{j2} with the neighbors of sample x_{j1} ($x_{j1,neighbors}$) belonging to the above shared patterns including $\{(x_{j2}, x_{j1,1}), \ldots,(x_{j2}, x_{j1,z1})\}$ are computed by the similarity matrix SM_{avg}. Next, the computed evidences are averaged by Equation (7) that is denoted by $Evidences_{neighbors}(x_{j1} , x_{j2})$:

$$Evidences_{neighbors}(x_{j1}, x_{j2}) = \frac{\sum_{w=1}^{z2} Evidences(x_{j1}, x_{j2,w}) + \sum_{w=1}^{z1} Evidences(x_{j2}, x_{j1,w})}{z1 + z2} \quad (7)$$

where Evidences(x_{j1}, $x_{j2,w}$) and Evidences(x_{j2}, $x_{j1,w}$) denote the evidences of the sample x_{j1} with the neighbors of sample x_{j2} and the evidences of the sample x_{j1} with the neighbors of sample x_{j1}, respectively. Thirdly, the new similarity matrix $SM^{(new)}$ is sampled by Equation (8) as:

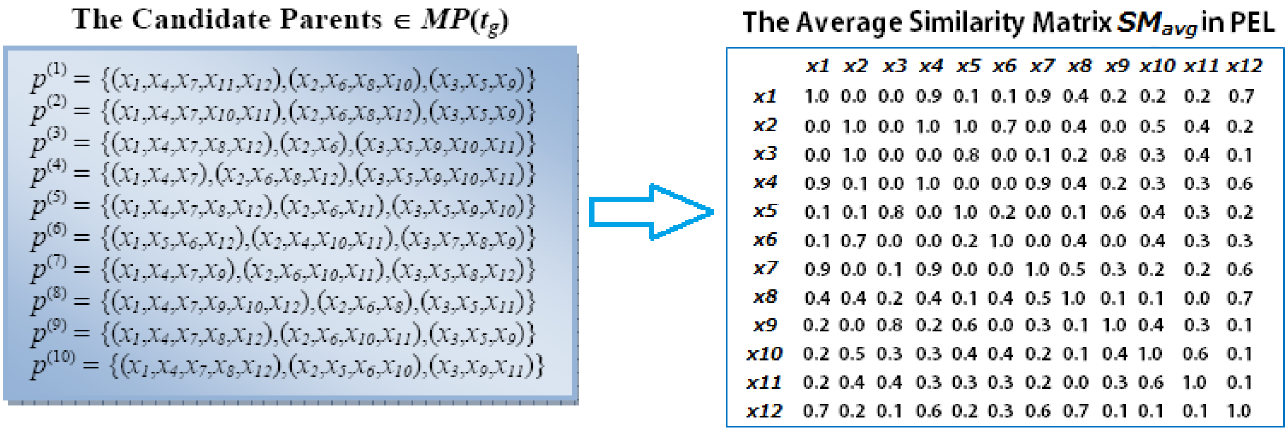
$$SM^{(new)}(x_{j1}, x_{j2}) = \begin{cases} 1 & \text{if rand (1)} < Evidences_{neighbors}(x_{j1} , x_{j2}) \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

To better understand the evidence computation based on neighbors of samples in proposed PEL method, a numerical example is given of proposed PEL method that illustrates all the three main steps of PEL which described above. As shown in Figure 6, it is assumed there are ten candidate patterns as cluster partitions $p^{(1)}$, $p^{(2)}$, …, $p^{(10)}$, belonged to mating pool

Power System Technology
ISSN:1000-3673

Received: 06-11-2024 Revised: 15-12-2024 Accepted: 05-01-2025

$MP(t_g)$, where numbers of samples is n=12 and number of cluster is k=3. In the first step of PEL, the first average similarity matrix SM_{avg} is generated by PEL, using Equations 2 and 3, and based on the existing cluster partitions.



The Candidate Parents $\in MP(t_g)$

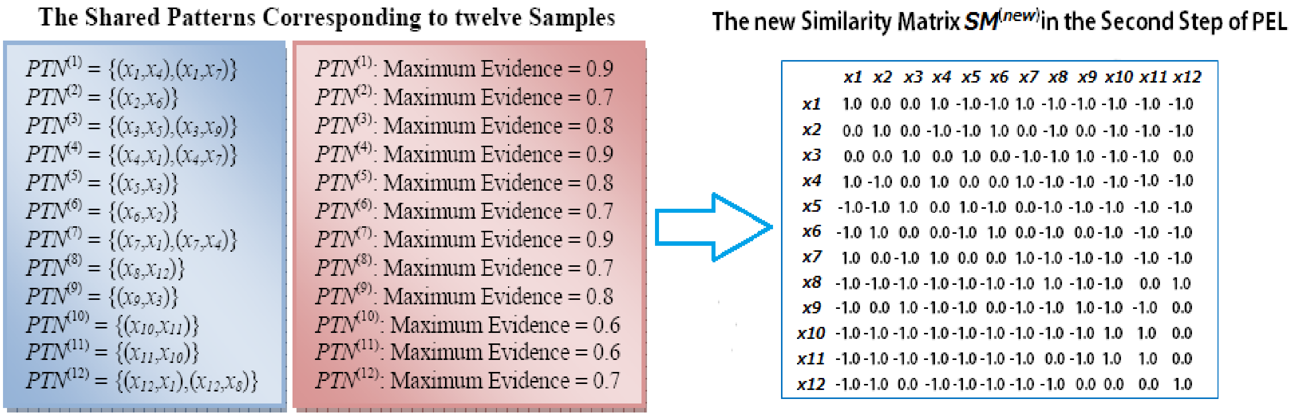
$p^{(1)} = \{(x_1,x_4,x_7,x_{11},x_{12}),(x_2,x_6,x_8,x_{10}),(x_3,x_5,x_9)\}$
$p^{(2)} = \{(x_1,x_4,x_7,x_{10},x_{11}),(x_2,x_6,x_8,x_{12}),(x_3,x_5,x_9)\}$
$p^{(3)} = \{(x_1,x_4,x_7,x_8,x_{12}),(x_2,x_6),(x_3,x_5,x_9,x_{10},x_{11})\}$
$p^{(4)} = \{(x_1,x_4,x_7),(x_2,x_6,x_8,x_{12}),(x_3,x_5,x_9,x_{10},x_{11})\}$
$p^{(5)} = \{(x_1,x_4,x_7,x_8,x_{12}),(x_2,x_6,x_{11}),(x_3,x_5,x_9,x_{10})\}$
$p^{(6)} = \{(x_1,x_5,x_6,x_{12}),(x_2,x_4,x_{10},x_{11}),(x_3,x_7,x_8,x_9)\}$
$p^{(7)} = \{(x_1,x_4,x_7,x_9),(x_2,x_6,x_{10},x_{11}),(x_3,x_5,x_8,x_{12})\}$
$p^{(8)} = \{(x_1,x_4,x_7,x_9,x_{10},x_{12}),(x_2,x_6,x_8),(x_3,x_5,x_{11})\}$
$p^{(9)} = \{(x_1,x_4,x_7,x_8,x_{12}),(x_2,x_6,x_{10},x_{11}),(x_3,x_5,x_9)\}$
$p^{(10)} = \{(x_1,x_4,x_7,x_8,x_{12}),(x_2,x_5,x_6,x_{10}),(x_3,x_9,x_{11})\}$

The Average Similarity Matrix SM_{avg} in PEL

	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12
x1	1.0	0.0	0.0	0.9	0.1	0.1	0.9	0.4	0.2	0.2	0.2	0.7
x2	0.0	1.0	0.0	1.0	1.0	0.7	0.0	0.4	0.0	0.5	0.4	0.2
x3	0.0	1.0	0.0	0.0	0.8	0.0	0.1	0.2	0.8	0.3	0.4	0.1
x4	0.9	0.1	0.0	1.0	0.0	0.0	0.9	0.4	0.2	0.3	0.3	0.6
x5	0.1	0.1	0.8	0.0	1.0	0.2	0.0	0.1	0.6	0.4	0.3	0.2
x6	0.1	0.7	0.0	0.0	0.2	1.0	0.0	0.4	0.0	0.4	0.3	0.3
x7	0.9	0.0	0.1	0.9	0.0	0.0	1.0	0.5	0.3	0.2	0.2	0.6
x8	0.4	0.4	0.2	0.4	0.1	0.4	0.5	1.0	0.1	0.1	0.0	0.7
x9	0.2	0.0	0.8	0.2	0.6	0.0	0.3	0.1	1.0	0.4	0.3	0.1
x10	0.2	0.5	0.3	0.3	0.4	0.4	0.2	0.1	0.4	1.0	0.6	0.1
x11	0.2	0.4	0.4	0.3	0.3	0.3	0.2	0.0	0.3	0.6	1.0	0.1
x12	0.7	0.2	0.1	0.6	0.2	0.3	0.6	0.7	0.1	0.1	0.1	1.0

Figure 6: An example for the first step of proposed PEL method.

In the second step of PEL, the shared patterns corresponding to all samples (n=12) including $PTN^{(1)}$, ..., $PTN^{(12)}$ are calculated by Equation 4 and then, maximum evidences are computed for all shared patterns, which is displayed in Figure 7. Next, considering to the all existing shared patterns, the second average similarity matrix $SM^{(new)}_{avg}$ is generated by PEL and using Equations 5 and 6.



The Shared Patterns Corresponding to twelve Samples

$PTN^{(1)} = \{(x_1,x_4),(x_1,x_7)\}$
$PTN^{(2)} = \{(x_2,x_6)\}$
$PTN^{(3)} = \{(x_3,x_5),(x_3,x_9)\}$
$PTN^{(4)} = \{(x_4,x_1),(x_4,x_7)\}$
$PTN^{(5)} = \{(x_5,x_3)\}$
$PTN^{(6)} = \{(x_6,x_2)\}$
$PTN^{(7)} = \{(x_7,x_1),(x_7,x_4)\}$
$PTN^{(8)} = \{(x_8,x_{12})\}$
$PTN^{(9)} = \{(x_9,x_3)\}$
$PTN^{(10)} = \{(x_{10},x_{11})\}$
$PTN^{(11)} = \{(x_{11},x_{10})\}$
$PTN^{(12)} = \{(x_{12},x_1),(x_{12},x_8)\}$

$PTN^{(1)}$: Maximum Evidence = 0.9
$PTN^{(2)}$: Maximum Evidence = 0.7
$PTN^{(3)}$: Maximum Evidence = 0.8
$PTN^{(4)}$: Maximum Evidence = 0.9
$PTN^{(5)}$: Maximum Evidence = 0.8
$PTN^{(6)}$: Maximum Evidence = 0.7
$PTN^{(7)}$: Maximum Evidence = 0.9
$PTN^{(8)}$: Maximum Evidence = 0.7
$PTN^{(9)}$: Maximum Evidence = 0.8
$PTN^{(10)}$: Maximum Evidence = 0.6
$PTN^{(11)}$: Maximum Evidence = 0.6
$PTN^{(12)}$: Maximum Evidence = 0.7

The new Similarity Matrix $SM^{(new)}$ in the Second Step of PEL

	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12
x1	1.0	0.0	0.0	1.0	-1.0	-1.0	1.0	-1.0	-1.0	-1.0	-1.0	-1.0
x2	0.0	1.0	0.0	-1.0	-1.0	1.0	0.0	-1.0	0.0	-1.0	-1.0	-1.0
x3	0.0	0.0	1.0	0.0	1.0	0.0	-1.0	-1.0	1.0	-1.0	-1.0	0.0
x4	1.0	-1.0	0.0	1.0	0.0	0.0	1.0	-1.0	-1.0	-1.0	-1.0	-1.0
x5	-1.0	-1.0	1.0	0.0	1.0	-1.0	0.0	-1.0	-1.0	-1.0	-1.0	-1.0
x6	-1.0	1.0	0.0	0.0	-1.0	1.0	0.0	-1.0	0.0	-1.0	-1.0	-1.0
x7	1.0	0.0	-1.0	1.0	0.0	0.0	1.0	-1.0	-1.0	-1.0	-1.0	-1.0
x8	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	1.0	-1.0	-1.0	0.0	1.0
x9	-1.0	0.0	1.0	-1.0	-1.0	0.0	-1.0	-1.0	1.0	-1.0	-1.0	0.0
x10	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	1.0	1.0	0.0
x11	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	0.0	-1.0	1.0	1.0	0.0
x12	-1.0	-1.0	0.0	-1.0	-1.0	-1.0	-1.0	-1.0	0.0	0.0	0.0	1.0

Figure 7: The second step of proposed PEL method.

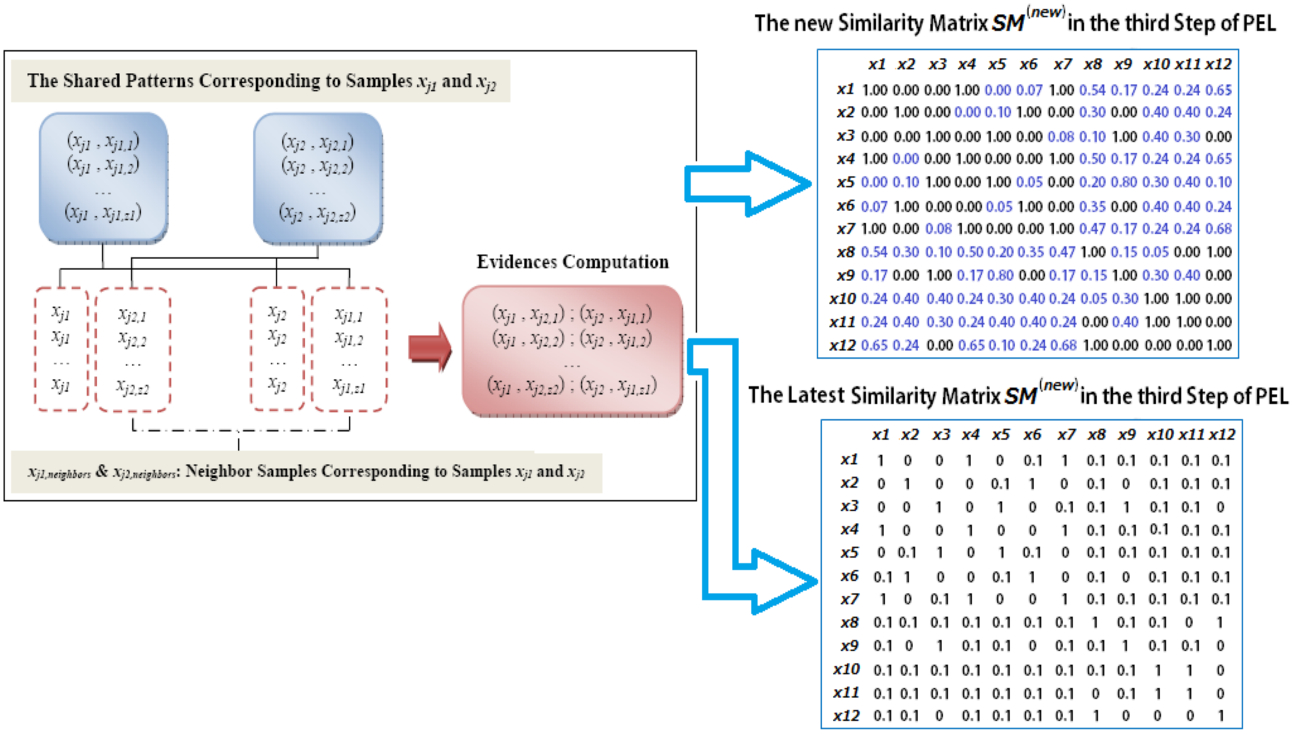
As illustrated in Figure 8, in the third step of PEL, the evidences of all samples are computed based on their neighbors by Equation 7 and their similarity matrix $SM^{(new)}_{avg}$ by Equation 8, in which the shared patterns corresponding to samples (x_{j1} , x_{j2}) are calculated in x_{j1}

Power System Technology
ISSN:1000-3673

Received: 06-11-2024 Revised: 15-12-2024 Accepted: 05-01-2025

and x_{j2}, separately. The evidence of paired-sample (x_{j1} , x_{j2}) is computed on basis of neighors of sample x_{j1} and neighors of sample x_{j2} and finally, neighors of samples corresponding to paired-sample (x_{j1} , x_{j2}) are considered in this method.

The pattern ensemble learning method resolves the correspondence problem by mapping a given set of cluster partitions to target cluster partition using similarity values [9,19,24,44,45]. Moreover, it should be mentioned that PEL solves the most common problems of the context insensitivity and the clustering insensitivity, because different clustering solutions with the same clustering context are transformed into one similarity matrix that represents one clustering context. In addition, since the new clustering solutions $p^{(new)}(t_g)$ are directly generated by the average-linkage clustering algorithm whose the number of clusters was fixed (k) already, PEL avoids from the clustering invalidity problem. PEL is also expected to increase the clustering accuracy resulted by the proposed incremental GA-based clustering ensemble algorithm (IGCE-PEL).


The Shared Patterns Corresponding to Samples x_{j1} and x_{j2}
(x_{j1} , $x_{j1,1}$)
(x_{j1} , $x_{j1,2}$)
...
(x_{j1} , $x_{j1,z1}$)
(x_{j2} , $x_{j2,1}$)
(x_{j2} , $x_{j2,2}$)
...
(x_{j2} , $x_{j2,z2}$)
x_{j1} x_{j1} ... x_{j1}
$x_{j2,1}$ $x_{j2,2}$ ... $x_{j2,z2}$
x_{j2} x_{j2} ... x_{j2}
$x_{j1,1}$ $x_{j1,2}$ ... $x_{j1,z1}$
Evidences Computation
(x_{j1} , $x_{j2,1}$) ; (x_{j2} , $x_{j1,1}$)
(x_{j1} , $x_{j2,2}$) ; (x_{j2} , $x_{j1,2}$)
...
(x_{j1} , $x_{j2,z2}$) ; (x_{j2} , $x_{j1,z1}$)
$x_{j1,neighbors}$ & $x_{j2,neighbors}$: Neighbor Samples Corresponding to Samples x_{j1} and x_{j2}

The new Similarity Matrix $SM^{(new)}$ in the third Step of PEL

	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12
x1	1.00	0.00	0.00	1.00	0.00	0.07	1.00	0.54	0.17	0.24	0.24	0.65
x2	0.00	1.00	0.00	0.00	0.10	1.00	0.00	0.30	0.00	0.40	0.40	0.24
x3	0.00	0.00	1.00	0.00	1.00	0.00	0.08	0.10	1.00	0.40	0.30	0.00
x4	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.50	0.17	0.24	0.24	0.65
x5	0.00	0.10	1.00	0.00	1.00	0.05	0.00	0.20	0.80	0.30	0.40	0.10
x6	0.07	1.00	0.00	0.00	0.05	1.00	0.00	0.35	0.00	0.40	0.40	0.24
x7	1.00	0.00	0.08	1.00	0.00	0.00	1.00	0.47	0.17	0.24	0.24	0.68
x8	0.54	0.30	0.10	0.50	0.20	0.35	0.47	1.00	0.15	0.05	0.00	1.00
x9	0.17	0.00	1.00	0.17	0.80	0.00	0.17	0.15	1.00	0.30	0.40	0.00
x10	0.24	0.40	0.40	0.24	0.30	0.40	0.24	0.05	0.30	1.00	1.00	0.00
x11	0.24	0.40	0.30	0.24	0.40	0.40	0.24	0.00	0.40	1.00	1.00	0.00
x12	0.65	0.24	0.00	0.65	0.10	0.24	0.68	1.00	0.00	0.00	0.00	1.00

The Latest Similarity Matrix $SM^{(new)}$ in the third Step of PEL

	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12
x1	1	0	0	1	0	0.1	1	0.1	0.1	0.1	0.1	0.1
x2	0	1	0	0	0.1	1	0	0.1	0	0.1	0.1	0.1
x3	0	0	1	0	1	0	0.1	0.1	1	0.1	0.1	0
x4	1	0	0	1	0	0	1	0.1	0.1	0.1	0.1	0.1
x5	0	0.1	1	0	1	0.1	0	0.1	0.1	0.1	0.1	0.1
x6	0.1	1	0	0	0.1	1	0	0.1	0	0.1	0.1	0.1
x7	1	0	0.1	1	0	0	1	0.1	0.1	0.1	0.1	0.1
x8	0.1	0.1	0.1	0.1	0.1	0.1	0.1	1	0.1	0.1	0	1
x9	0.1	0	1	0.1	0.1	0	0.1	0.1	1	0.1	0.1	0
x10	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	1	1	0
x11	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0	0.1	1	1	0
x12	0.1	0.1	0	0.1	0.1	0.1	0.1	1	0	0	0	1

Figure 8: The third step of proposed PEL method.

4. Experiment Results

In this Section, the proposed incremental GA-based clustering ensemble algorithm using the pattern ensemble learning method is evaluated by an experimental study runs on twelve benchmark datasets including one synthetic dataset, named X8K5D, and eleven real world datasets, named Diabetes, Glass, Heart, Ionosphere, Iris, Lymph, Promoters, Segmentation, Thyroid, Wine, and Zoo, where have been used by many reseaches such as [8,19,21,24,32,46,47]. The experimental results have been evaluated as the performance of the incremental GA-based clustering ensemble algorithms, in terms of the clustering accuracy that has been utilized by various articles such as [12,25,32,46,50].

To study the potentials of PEL as the proposed recombination operator, six desired incremental GA-based clustering ensemble algorithms using six recombination operators are implemented and tested in MATLAB R2020a, as follows: Incremental GA-based Clustering algorithm with Single-Point Crossover (IGCE-SPC), with Two-Point Crossover (IGCE-TPC), with Ensemble Learning method (IGCE-EL) [9], with Evidence Accumulation method (IGCE-EA) [20], with Co-Clustering Ensemble method (IGCE-CCE) [50], and with Pattern Ensemble Learning method (IGCE-PEL) as proposed learning method in this paper. In addition, the clustering accuracies obtained by the incremental GA-based clustering ensemble algorithms (Figure 3) are evaluated by the Rand index method [9,51]. The obtained average clustering accuracies are compared on the basis of different recombination operators applied. Table 1 summarizes genetic parameter setting.

Table 1: The genetic parameter setting.

Parameter	Value
Population Size	$N_{population}$ = 100
Probability $\mu_{selection}$	0.7
Tournament Size	2
Probability $\mu_{crossover}$	0.8
Probability $\mu_{mutation}$	0.005
Termination Condition	85% chromosomes without changing of fitness

This experiment set evaluates the clustering accuracy obtained by the incremental GA-based clustering ensemble algorithms on the basis of different recombination operators that are employed by them. Experimental results from the incremental GA-based clustering ensemble algorithms with two traditional crossover operators including SPC and TPC, and also four ensemble learning methods including EL, EA, CCE and PEL employed by IGCE-SPC, IGCE-TPC, IGCE-EL, IGCE-EA, IGCA-CCE and IGCE-PEL respectively, are presented on the twelve benchmark datasets consist of X8K5D, Diabetes, Glass, Heart, Ionosphere, Iris, Lymph, Promoters, Segmentation, Thyroid, Wine, and Zoo, and are analyzed that are shown in Table 2.

As indicated in the Table 2, the minimum and maximum increasing of average clustering accuracy obtained by the incremental GA-based clustering ensemble algorithms using two employed traditional recombination operators, named as IGCE-SPC and IGCE-TPC, respectively, are fluctuated in the range of [2.40,13.72] and [-1.12,6.49] by percentage on the twelve benchmark datasets as compared to the clustering accuracy of the initial population. In comparison to two traditional crossover operators, highest average clustering accuracy obtained IGCE-SPC is higher than IGCE-TPC on eleven out of twelve datasets, in other words, 90% of the times, IGCE-SPC is better than IGCE-TPC. Also, the highest average clustering accuracy obtained by IGCE-SPC are greater than the average clustering accuracy obtained by IGCE-TPC in the range of [5.37,11.32] by percentage on the twelve benchmark datasets.

Table 2: Highest average clustering accuracy by the incremental GA-based clustering ensemble algorithms on the basis of two traditional crossover (SPC and TPC) and four ensemble learning methods (EL, EA, CCE and PEL).

IGCE with	**Recomination Operators**						
Datasets	**Population(%)**	**SPC(%)**	**TPC(%)**	**EL(%)**	**EA(%)**	**CCE(%)**	**PEL(%)**
X8K5D	62.19	69.24	64.99	78.55	86.17	93.04	**95.62**
Diabetes	41.08	43.48	40.56	52.44	68.38	78.16	**81.34**
Glass	60.73	67.04	63.29	70.09	87.45	82.74	**90.47**
Heart	36.17	44.43	41.82	52.88	67.24	74.95	**78.09**
Ionosphere	38.03	40.87	37.69	44.21	64.67	63.38	**70.83**
Iris	59.87	70.44	63.53	76.96	84.36	92.18	**95.54**
Lymph	39.87	47.70	43.93	50.73	63.66	66.25	**70.59**
Promoters	40.13	44.24	46.60	56.11	72.50	78.33	**83.28**
Segmentation	53.10	64.33	59.59	69.19	83.29	90.06	**93.61**
Thyroid	58.79	64.51	61.27	80.25	88.85	92.80	**97.02**
Wine	62.93	68.23	61.81	76.59	93.97	91.11	**95.95**
Zoo	49.49	63.21	54.97	72.90	84.64	88.25	**90.92**

Similarly, in comparison to four ensemble learning methods, the minimum and maximum average clustering accuracy resulted by IGCE-EL, IGCE-EA, IGCE-CCE and IGCE-PEL are fluctuated in the range of [6.18,23.41], [23.79,35.15], [22.01,38.78], and [29.74,43.15] by percentage on the twelve benchmark datasets. Also in comparison to four ensemble learning methods, the average clustering accuracy resulted by IGCE-EL is always less than IGCE-EA,

Received: 06-11-2024 *Revised: 15-12-2024* *Accepted: 05-01-2025*

IGCE-CCE and IGCE-PEL on twelve out of twelve datasets, in other words, IGCE-EL is worsth operator among the four ensemble learning methods. Further, the average clustering accuracy resulted by IGCE-CCE is usually better than IGCE-EA on nine out of twelve datasets, in other words, 75% of the times, IGCE-CCE is better than IGCE-EA.

In addition, the average clustering accuracy obtained by IGCE-PEL are greater than all three average clustering accuracy obtained by IGCE-EL in the range of [16.77,28.90], IGCE-EA in the range of [1.98,12.96], IGCE-CCE in the range of [2.58,7.73] percentage on the twelve benchmark datasets, respectively. Consequently, PEL is regarded as the better candidate recombination operators among EL, EA and CCE. In comparison to six recombination operators, the highest average clustering accuracy resulted by clustering solutions are achieved by PEL as compared to other five recombination operators including SPC, TPC, EL, EA, and CCE on the twelve benchmark datasets. In other words, IGCE-PEL is best operator among the four ensemble learning methods and six recombination operations. Briefly, the minimum and maximum increasing of average clustering accuracy resulted by IGCE-PEL are in the range of [22.89,39.04], [26.66,40.78], [16.77,28.90], [1.98,12.96], and [2.58,7.73] by percentage on the twelve benchmark datasets, against the average clustering accuracy resulted by IGCE-SPC, IGCE-TPC, IGCE-EL, IGCE-EA, and IGCE-CCE, respectively.

On the other hand, the average clustering accuracy obtained by clustering solutions generated by TPC are minimum among other five recombination operators including SPC, EL, EA, CCE and PEL on the twelve benchmark datasets. Consequently, TPC is regarded as the worst candidate recombination operator among other mentioned recombination operators. Unlike the ensemble learning methods, considering that improvement of the average clustering accuracy by SPC and TPC is low as compared to the clustering accuracy of the population employed, it can be concluded that the good patterns are lost due to the problems of clustering invalidity and the clustering insensitivity.

As aforementioned, the average clustering accuracy obtained by the Pattern Ensemble Learning method (IGCE-PEL) is much better than the average clustering accuracy resulted by traditional recomination operators including Incremental GA-based Clustering algorithm with Single-Point Crossover (IGCE-SPC) and with Two-Point Crossover (IGCE-TPC), in the range of [22.89,39.4] and [26.66,40.78] respectively. Since most researchers acknowledge the fact that the traditional genetic recombination operators in the GA-based algorithms are much more ineffective than the smart recombination genetic operators such as ensemble learning in the optimal offspring reproduction stage, the result is logical.

In additoin, the average clustering accuracy obtained by the Pattern Ensemble Learning method (IGCE-PEL) is better than the average clustering accuracy resulted by two non-traditional recomination operators including Incremental GA-based Clustering algorithm with

Received: 06-11-2024 Revised: 15-12-2024 Accepted: 05-01-2025

Ensemble Learning method (IGCE-EL) [9] and with Evidence Accumulation method (IGCE-EA) [20], in the range of [16.77,28.90] and [26.66,40.78] respectively. These results can be due to the following reasons. Firstly, the elements of the new similarity matrix $SM^{(new)}$ sampled by EL and EA are generated only on the basis of the random numbers generated, whereas the elements of $SM^{(new)}$ sampled by PEL are generated on the basis of the shared patterns, the neighbor samples corresponding to the shared patterns, and the random numbers.

Secondly, in EL and EA, if the generated random numbers are greater than the evidences of the appropriate shared patterns, the appropriate shared patterns with high evidences are ignored in $SM^{(new)}$. In addition, in PEL, if the generated random numbers are less than the evidences of the unsuitable shared patterns, the unsuitable shared patterns with low evidences are kept in $SM^{(new)}$, while EL and EA do not operate like PEL and lose the appropriate gene shared patterns for the reproduction of the next generations. However, on the basis of any random numbers in PEL, the appropriate shared patterns with maximum evidences are not ignored in $SM^{(new)}$ and the unsuitable shared patterns with minimum evidences are removed in $SM^{(new)}$. Thirdly, in PEL, most elements of $SM^{(new)}$ are assigned on the basis of the average evidences of neighbor samples in the paired-samples belonging to the shared patterns. It means that not only the each paired-sample in $SM^{(new)}$ is assigned on the basis of the comparison between their evidences and the random numbers, but also it is assigned on the basis of the comparison between the average evidences of neighbor samples belonging to the each paired-sample and the random numbers. It is effective to seek desirable links between two samples in $SM^{(new)}$ that leads to generate the appropriate new clustering solutions. Therefore, it is possible that the appropriate patterns existing in the paired-samples with high frequency have ignored in the new similarity matrix by comparing with a random number with very high value. It has resulted that these appropriate patterns can not be transferred to the next generations.

Furtheremore, the average clustering accuracy obtained by the Pattern Ensemble Learning method (IGCE-PEL) is better than the average clustering accuracy resulted by Incremental GA-based Clustering algorithm with non-traditional recomination operator including Co-Clustering Ensemble method (IGCE-CCE) [50], in the range of [2.58,7.73]. As previously mentioned, CCE that is a multi-objective GA-based co-clustering ensemble algorithm, with the processing of fuzzy samples and general samples as objective function, and using encoded chromosomes as the membership of rows and columns, hence CCE cannot consider for the each paired-sample in $SM^{(new)}$ to assign on the basis of the comparison between their membership of rows and columns in fuzzy samples and general samples. Therefore, it loses the appropriate gene shared patterns from their neighors for the reproduction of the next generations.

5. Conclusions and Future Works

The clustering ensemble has appeared as an outstanding method for improving clustering accuracy in the unsupervised classification. Apart from the correspondence problem in the unsupervised classification, other associated problems with the clustering ensemble are the diversity of clustering and consensus functions. Moreover, GAs are known methods with high ability to solve optimization problems like clustering ensemble. The standard GA-based clustering ensemble algorithms that applied the traditional crossover operators suffered the common problems containing the loss of population diversity, clustering invalidity, and context insensitivity. In response to the above-mentioned challenges, this study was devoted towards developing a clustering ensemble learning method based on the incremental GA-based algorithms to group unlabeled samples.

At first, according to the two main stages of the methodology, an architecture for the clustering ensemble based on the incremental GA-based algorithms was proposed that consists of two phases. In the first phase, random subspace method were applied to produce cluster partitions as population. In the second phase, several incremental GA-based clustering ensemble algorithms using different recombination operators such as single-point crossover, multi-point crossover and ensemble learning method were able to combine the cluster partitions for generating the new clustering solutions. An incremental GA-based clustering ensemble algorithm using the pattern ensemble learning method, termed as IGCE-PEL, has been developed that utilizes the evidence accumulation clustering method as consensus function.

In comparison to the clustering accuracy resulted by the incremental GA-based clustering ensemble algorithms using different recombination operators, experimental results have demonstrated that multi-point and single-point crossover as the traditional crossover operators usually were not able to reproduce the clustering solutions with high accuracy due to the loss of good patterns. Moreover, ensemble learning and pattern ensemble learning as the ensemble learning methods were always capable of reproducing the clustering solutions with high accuracy. Due to the fact that highest average clustering accuracy has been achieved by the pattern ensemble learning method on the twelve benchmark datasets, the pattern ensemble learning method has been regarded as best candidate for recombination operator as compared to other recombination opetrators used.

On the basis of different concepts and applied techniques, different aspects of future work can be considered. Firstly, since there were various generative mechanisms to produce cluster partitions as ensemble members, other generative mechanisms, for instance fuzzy C-Means clustering algorithm, can be applied in the first stage of the clustering ensemble architecture that may be more successful in the exploitation of diversity. Secondly, considering that adaptive GAs include varied parameters such as the population size, the crossover

Power System Technology
ISSN:1000-3673
Received: 06-11-2024 Revised: 15-12-2024 Accepted: 05-01-2025

probability, and the mutation probability, they may also be used to develop further superior GA-based clustering ensemble algorithm. Thirdly, other consensus functions may be applied in the clustering ensemble methods such as hypergraph partitioning and mixture model that can be considered as part of future work.

References

[1] Boongoen, T., and Iam-On, N., "Cluster ensembles: A survey of approaches with recent extensions and applications", *Computer Science Review*. 28, pp. 1-25, DOI: https://doi.org/10.1016/ j.cosrev.2018.01.003, (2018).

[2] Yang, Y., "Chapter 6 - Unsupervised Learning via an Iteratively Constructed Clustering Ensemble", *Temporal Data Mining Via Unsupervised Ensemble Learning*, pp. 75-92, DOI: https:// doi.org/10.1016/b978-0-12-811654-8.00006-3, (2017).

[3] Yang, Y., "Chapter 7 - Temporal Data Clustering via a Weighted Clustering Ensemble with Different Representations", *Temporal Data Mining Via Unsupervised Ensemble Learning*, pp. 93-122, DOI: https://doi.org/10.1016/b978-0-12-811654-8.00007-5, (2017).

[4] Sewell, M., "Ensemble learning", *Research Note*, 11(02), DOI: https://doi.org/10.1023/b:verc. 0000026724.82897.64, (2011).

[5] Wu, X., Ma, T., Cao, J., Tian Y., and Alabdulkarim, A., "A comparative study of clustering ensemble algorithms", *Computers & Electrical Engineering*, 68, pp. 603-615, DOI: https://doi.org/ 10.1016/j.compeleceng.2018.05.005, (2018).

[6] Azimi, J., and Mohammadi, M., "Clustering ensembles using genetic algorithm", *International Workshop on Computer Architecture for Machine Perception and Sensing*, IEEE, pp. 119-123, DOI: https://doi.org/10.1109/camp.2007.4350366, (2006).

[7] Hruschka, Eduardo Raul, Ricardo JGB Campello, and Alex A. Freitas, "A survey of evolutionary algorithms for clustering", *IEEE Transactions on systems, man, and cybernetics*, Part C (applications and reviews), 39(2), pp. 133-155, DOI: https://doi.org/10.1109/tsmcc.2008.2007252, (2009).

[8] Chatterjee, S., and Mukhopadhyay, A., "Clustering Ensemble: A Multi objective Genetic Algorithm based Approach", *Procedia Technology*, 10, pp. 443-449, DOI: https://doi.org/10.1016/j. protcy.2013.12.381, (2013).

Received: 06-11-2024 Revised: 15-12-2024 Accepted: 05-01-2025

[9] Hong, Y., and Kwong, S., "To combine steady-state genetic algorithm and ensemble learning for data clustering", *Pattern Recognition Letters*, 29(9), pp. 1416-1423, DOI: https://doi.org/10.1016/j. patrec.2008.02.017, (2008).

[10] Lu, Y., Lu, S., Fotouhi, F., Deng, Y., and Brown, S. J., "FGKA: a fast genetic k-means clustering algorithm", *Proceedings of the ACM symposium on Applied computing*, ACM, pp. 622-623, DOI: https://doi.org/10.1145/967900.968029, (2004).

[11] Yoon, H. S., Ahn, S. Y., Lee, S. H., Cho, S. B., and Kim, J., "A Novel Framework for Discovering Robust Cluster Results", *Discovery Science*, Springer, pp. 373-377, DOI: https://doi.org/ 10.1007/11893318_45, (2006).

[12] Yang, W., Zhang, Y., Wang, H., Deng, P., and Li, T., "Hybrid genetic model for clustering ensemble", *Knowledge-Based Systems*, 231, DOI: https://doi.org/10.1016/j.knosys.2021.107457, (2021).

[13] Ozyer, T., and Alhajj, R., "Parallel clustering of high dimensional data by integrating multi-objective genetic algorithm with divide and conquer", *Applied Intelligence*, 31(3), pp. 318-331, DOI: https://doi.org/10.1007/s10489-008-0129-8, (2009).

[14] Ghaemi, R., Sulaiman, N., Ibrahim, H., and Mustapha, N., "A review: accuracy optimization in clustering ensembles using genetic algorithms", *International Journal of Artificial Intelligence Review*, ISI, Springer, pp. 287-318, DOI: https://doi.org/10.1007/s10462-010-9195-5, (2011).

[15] Golalipour, K., Akbari, E., Hamidi, S., and Lee, M., "From clustering to clustering ensemble selection: A review", *Engineering Applications of Artificial Intelligence*, 104, DOI: https://doi.org/ 10.1016/j.engappai.2021.104388, (2021).

[16] Zhang, M., "Weighted clustering ensemble: A review", *Pattern Recognition*, 124, DOI: https:// doi.org/10.1016/j. patcog.2021.108428, (2022).

[17] Jugal K. Kalita, Dhruba K. Bhattacharyya, and Swarup R., "Book chapter: Ensemble learning, Elsevier", *Fundamentals of Data Science*, ISBN: 9780323917780, DOI: https://doi.org/10.1016/ b978-0-32-391778-0.00017-x, (2024).

[18] Ghaemi, R., Sulaiman, N., Ibrahim, H., and Mustapha, N., "A survey: clustering ensembles techniques", *Proceedings of the international conference on computer, electrical, and systems science, and engineering*, 38, pp. 644-653, DOI: https://doi.org/10.1109/itng.2010.88, (2009).

Received: 06-11-2024 Revised: 15-12-2024 Accepted: 05-01-2025

[19] Li, T., Rezaeipanah, A., and Tag El Din, E.M., "An ensemble agglomerative hierarchical clustering algorithm based on clusters clustering technique and the novel similarity measurement", *Journal of King Saud University - Computer and Information Sciences*, 34(6), Part B, pp. 3828-3842, DOI: https://doi.org/10.1016/j.jksuci.2022.04.010, (2022).

[20] Wong, W., and Tsuchiya, N., "Evidence accumulation clustering using combinations of features", *MethodsX*, 7, DOI: https://doi.org/10.1016/j.mex.2020.100916, (2020).

[21] Bai, L., Liang, J., and Cao, F., "A multiple k-means clustering ensemble algorithm to find nonlinearly separable clusters", *Information Fusion*. 61, pp. 36-47, DOI: https://doi.org/10.1016/j. inffus.2020.03.009, (2020).

[22] Ghaemi, R., Sulaiman, N., Ibrahim, H., and Mustapha, N., "A novel fuzzy C-means algorithm to generate diverse and desirable cluster solutions used by genetic-based clustering ensemble algorithms", *Memetic Computing Journal*, Springer, 4, pp. 49-71, DOI: https://doi.org/10.1007/ s12293-012-0073-3, (2012).

[23] Yang, Y., "Chapter 4 - Ensemble Learning", *Temporal Data Mining Via Unsupervised Ensemble Learning*, pp. 35-56, DOI: https://doi.org/10.1016/b978-0-12-811654-8.00004-x, (2017).

[24] Soufiane, K., and Tarek Khadir, M., "A multiple clustering combination approach based on iterative voting process", *Journal of King Saud University - Computer and Information Sciences*, 34(1), pp. 1370-1380, DOI: https://doi.org/10.1016/j.jksuci.2019.09.013, (2022).

[25] Kadhim, M.R., Zhou, G., and Tian, W., "A novel self-directed learning framework for cluster ensemble", *Journal of King Saud University - Computer and Information Sciences*, 34(10), Part A, pp. 7841-7855, DOI: https://doi.org/10.1016/j.jksuci.2022.07.003, (2022).

[26] Hong, Y., Kwong, S., Chang, Y., and Ren, Q., "Unsupervised feature selection using clustering ensembles and population based incremental learning algorithm", *Pattern Recognition*, 41(9), pp. 2742-2756, DOI: https://doi.org/10.1016/j.patcog.2008.03.007, (2008).

[27] Dietterichl, T. G., "Ensemble methods in machine learning", *Multiple classifier systems*, Springer, pp. 1-15, DOI: https://doi.org/10.1007/3-540-45014-9_1, (2000).

Received: 06-11-2024 *Revised: 15-12-2024* *Accepted: 05-01-2025*

[28] Dietterichl, T. G., "Ensemble learning", *Handbook of Brain Theory and Neural Networks*, The MIT Press, pp. 1-9, DOI: https://doi.org/10.7551/mitpress/3413.001.0001, (2002).

[29] Kang, K., Zhang, H. X., and Fan, Y., "A Novel Clusterer Ensemble Algorithm Based on Dynamic Cooperation", *Fifth International Conference on Fuzzy Systems and Knowledge Discovery*, IEEE, 1, pp. 32-35, DOI: https://doi.org/10.1109/fskd.2008.339, (2008).

[30] Zhang, L., Zhou, W., Wu, C., Huo, J., Zou, H., and Jiao, L., "Center matching scheme for k-means cluster ensembles", *Proceedings of Society for Optical Engineering*, 7496, pp. 14-39, DOI: https://doi.org/10.1117/12.832603, (2009).

[31] Yu, L., Cao, F., Zhao, X., Yang, X., and Liang, J., "Combining attribute content and label information for categorical data ensemble clustering", *Applied Mathematics and Computation*, 381, DOI: https://doi.org/10.1016/j.amc.2020.125280, (2020).

[32] Shi, H., Peng, Q., Xie, Z., and Wang, J., "A semi-supervised hierarchical ensemble clustering framework based on a novel similarity metric and stratified feature sampling", *Journal of King Saud University - Computer and Information Sciences*, 35(8), DOI: https://doi.org/10.1016/j.jksuci.2023. 101687, (2023).

[33] Kaufman, D., "Mutation invariant functions on cluster ensembles", *Journal of Pure and Applied Algebra*, 228(2) , DOI: https://doi.org/10.1016/j.jpaa.2023.107495, (2024).

[34] Aktaş, D., Banu Lokman, B., İnkaya, T., and Dejaegere, G., "Cluster ensemble selection and consensus clustering: A multi-objective optimization approach", *European Journal of Operational Research*, 34(3), pp. 1065-1077, DOI: https://doi.org/10.1016/j.ejor.2023.10.029, (2024).

[35] Shan, Y., Li, S., Li, F., Cui, Y., Chen, M., and He., X., "Fuzzy self-consistent clustering ensemble", *Applied Soft Computing*, 151, DOI: https://doi.org/10.1016/j.asoc.2023.111151, (2024).

[36] Golalipour, K., Akbari, E., and Motameni, H., "Cluster ensemble selection based on maximum quality-maximum diversity", *Engineering Applications of Artificial Intelligence*, 131, DOI: https:// doi.org/10.1016/j.engappai.2024.107873, (2024).

Received: 06-11-2024 *Revised: 15-12-2024* *Accepted: 05-01-2025*

[37] Chakraborty, J., Pradhan D.K., and Nandi, S., "A multiple k-means cluster ensemble framework for clustering citation trajectories", *Journal of Informetrics*, 18 (2), DOI: https://doi.org/10.1016/j.joi. 2024.101507, (2024).

[38] Chang L., and Shiwu Y., "A two-stage clustering ensemble algorithm applicable to risk assessment of railway signaling faults", *Expert Systems with Applications*, 249, Part A, DOI: https:// doi.org/10.1016/j.eswa.2024.123500, (2024).

[39] Lu, Y., Lu, S., Fotouhi, F., Deng, Y., and Brown, S. J., "Incremental genetic k-means algorithm and its application in gene expression data analysis", *BMC bioinformatics*, 5(1), pp. 172, DOI: https: //doi.org/10.1186/1471-2105-5-172, (2004).

[40] He, D., Wang, Z., Yang, B., and Zhou, C., "Genetic Algorithm with Ensemble Learning for Detecting Community Structure in Complex Networks", *Fourth International Conference on Computer Sciences and Convergence Information Technology*, IEEE, pp. 702-707, DOI: https://doi. org/10.1109/iccit.2009.189, (2009).

[41] Rogers, A., and Prügel-Bennett, A., "Modelling the dynamics of a steady state genetic algorithm", *Foundations of genetic algorithms*, 5, pp. 57-68, DOI: https://doi.org/10.1007/978-3-662 -04448-3_4, (1999).

[42] Dempster, A.P., N.M. Laird, and Rubin, D.B., "Maximum likelihood from incomplete data via the EM algorithm", *Journal of the Royal Statistical Society*, Series B (Methodological). 39(1), pp. 1-38, DOI: https://doi.org/10.1111/j.2517-6161.1977.tb01600.x, (1977).

[43] Fred, A. L. N., "Finding consistent clusters in data partitions", *Multiple Classifier Systems*, Springer, pp. 309-318, DOI: https://doi.org/10.1007/3-540-48219-9_31, (2001).

[44] Fred, A. L. N., and Jain, A. K., "Data clustering using evidence accumulation", *Pattern Recognition*, 4(40276), DOI: https://doi.org/10.1109/icpr.2002.1047450, (2002).

[45] Fred, A. L. N., and Jain, A. K., "Combining multiple clusterings using evidence accumulation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 835-850, DOI: https://doi.org/ 10.1109/tpami.2005.113, (2005).

[46] Kuo, R.J., Mei, C.H., Zulvia, F.E., and Tsai, C.Y., "An application of a metaheuristic algorithm-based clustering ensemble method to APP customer segmentation",

Received: 06-11-2024 Revised: 15-12-2024 Accepted: 05-01-2025

Neurocomputing, 205, pp. 116-129, DOI: https://doi.org/10.1016/j.neucom.2016.04.017, (2016).

[47] Zhu, S., Xu, L., and Goodman, E.D., "Evolutionary multi-objective automatic clustering enhanced with quality metrics and ensemble strategy", *Knowledge-Based Systems*, 188, DOI: https:// doi.org/10.1016/j.knosys.2019.105018, (2020).

[48] Hu, X., Wang, Z., Wang, Q., Chen, K., Han, Q., Bai, S., Du, J. and Chen, W., "Molecular classification reveals the diverse genetic and prognostic features of gastric cancer: A multi-omics consensus ensemble clustering", *Biomedicine & Pharmacotherapy*, 144, DOI: https://doi.org/10. 1016/j.biopha.2021.112222, (2021).

[49] Kordos, M., Blachnik, M., and Scherer, R., "Fuzzy clustering decomposition of genetic algorithm-based instance selection for regression problems", *Information Sciences*, 587, pp. 23-40, DOI: https://doi.org/10.1016/j.ins.2021.12.016, (2022).

[50] Zhong, Y., Wang, H., Yang, W., Wang, L., and Li, T., "Multi-objective genetic model for co-clustering ensemble", *Applied Soft Computing*, 135, DOI: https://doi.org/10.1016/j.asoc.2023. 110058, (2023).

[51] Ko, A.H.R., and Sabourin, R., "The Implication of Data Diversity for a Classifier-free Ensemble Selection in Random Subspaces", *19th International Conference on Pattern Recognition*, IEEE, DOI: https://doi.org/10.1109/icpr.2008.4761767, (2008).

[52] Ko, A.H.R., Sabourin, R., Britto Jr, A.S., and Oliveira, L.E.S., "A Classifier-free Ensemble Selection Method based on Data Diversity in Random Subspaces Technical Report", *arXiv*, Cornell University, pp. 1-24, DOI: https://doi.org/10.1109/icpr.2008.4761767, (2014).