Received: September 15, 2022. Revised: October 12, 2022.

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

An Adaptive Local Gravitation-based Optimized Weighted Consensus Clustering for Gene Expression Data Classification

Sangeetha Mani[1]* **Kousalya Rangaswamy[1]**

[1]*Department of Computer Science, Dr. N.G.P. Arts and Science College, Coimbatore, India*
* Corresponding author's Email: sangeethamphd234@gmail.com

Abstract: The appropriate categorization of tumors from a vast quantity of Gene Expression Data (GED) is one of the most difficult processes in clinical diagnosis. To combat this challenge, a Weighted Consensus of Lion Optimized K-means Ensemble with Peak Density Clustering (WECLO K-means-PDC) algorithm has been developed that calculates the Symmetric Neighborhood (SN) correlation among Data Points (DPs) using the lion optimization. An SN Graph (SNG) was created to select the number of Cluster Centroids (ClustCenter) at every iteration of clustering. But, it was not suitable if the dataset was sparse, as well as the constant density threshold may influence the distinguishing DPs within the cluster boundaries. In this paper, an Adaptive Local resultant Force neighborhood PDC for WECR K-means (WECLO K-means-ALFPDC) algorithm is proposed, which considers additional measures to determine the symmetric neighborhood correlation among the DPs when the dataset is sparse. The major goal is to consider the distinct variances between the sizes and orientations of the DPs nearer to the ClustCenters and edges. To find such variations, two novel local measures called Centrality (CR) and Coordination (CO) are introduced instead of SNG to select the ClustCenters and obtain more precise clustering for classifying cancer from genomic data. Finally, the test results show that the WECLO K-means-ALFPDC algorithm attains 88.7%, 89.1%, 88.42%, 88.38% and 89.04% accuracy on leukemia, lymphoma, prostate cancer, SRBCT and breast cancer databases, respectively compared to the WECLO K-means-PDC algorithm.

Keywords: Gene expression data, WECLO K-means, Peak density clustering, Symmetric neighborhood correlation, Centrality, Coordination.

1. Introduction

Medical diagnostics, including the categorization of malignancy into various disease groups that may appear almost identical in classical pathology conducted molecularly, are increasingly using GED. In this regard, machine-assisted imagery might support linking histopathologists' observations. The effectiveness of GED investigation backs them up for malignancies classification [1]. Nonetheless, studying GED provides a considerable problem because of the large number of chromosomes in the microarray collection.

The analysis gets progressively complicated as the volume of relevant data offered lowers. So, finding a small number of usable genes from a large number of genomic sequences to accurately describe the data becomes problematic. To solve this problem, gene selection procedures are often classified into one of two categories [2]. Before categorization, the chromosomal selection is performed using filtering approaches. In the wrapper approach, which is used to classify genes, the best genes in each chromosomal group are sought [3-5].

The t-test, Gini index and Wilcoxon rank-sum analyses are examples of filtering approaches or biomarker procedures and they are far more extensible than wrapper strategies. A few of the chosen genes can be deemed unnecessary because they don't contribute any further information to the subclass while using biomarker schemes. Different tumors and sub-tumors have a similar set of genes preferred over filtering procedures since they measure the amount of inter-correlation across distinct genes [6-8]. However, when dealing with large-scale datasets, such approaches entail a substantial computational overhead.

As a result, clustering algorithms have proven to be effective in investigating chromosomal processes, their regulation, parameters, and cell subdivisions. If 2 genes have similar activity characteristics (co-expressed chromosomes), then those genes can be clustered. This strategy results in a better understanding of the function of several chromosomes that were formerly identified [9-10]. To merge the findings of several grouping techniques, the Cluster Ensemble (CE) was recently evolved as a promising approach. It employs a consensus mechanism to ensure harmony across information chunks. Based on accuracy and consistency, it exploits all clusters in the composition [11]. Few researchers worked on Semi-Supervised CE (SSCE) that integrates SSC and CE [12-14]. SSCE frequently employs a low level of control in the primary step of CE, i.e. composite formation, by running several rounds of SSC techniques [15]. Nonetheless, in a dynamic learning scenario with shifting constrained control, it is neither computationally effective nor flexible.

To address this issue, a new algorithm known as the Weighted Consensus of Random K-means ensemble (WECR K-means) algorithm was designed [16], which considers the grouping intrinsic test rates and satisfaction level of a coupled rule. It was capable of dealing with non-spherical clusters. Conversely, if the database was very huge, the complexity of this algorithm becomes impractical. The condition applied to more attributes was a time-consuming procedure. As a result, while categorizing the massive quantity of genomic data, the grouping should be improved for Less Informative Composite Clusters (LICCs). So, the WECLO K-means-PDC algorithm was introduced [17] to remove the LICCs for the classification of GED. In this algorithm, a lion optimization algorithm was adopted rather than random subspace and sampling to achieve efficient clustering based on the fitness function, i.e. cluster validation metrics. Then, the SNG was created by the dynamic PDC coupled with K-means grouping for all DPs. The SNG was used to select the ClustCenters and update every group in all iterations of the K-means grouping without determining the cutoff value among DPs. According to this SNG, the outliers were identified as the DPs smaller than 2 adjacent. Also, each DP was assigned to an appropriate group using the breadth-first search on SNG and the LICCs were efficiently eliminated. On the other hand, the symmetric neighborhood correlation, i.e. SNG was formed only based on the distance and density. These two metrics were not satisfactory to form clusters while the dataset was sparse. Also, differentiating DPs within the boundary regions of clusters was difficult because of using a fixed threshold value of density.

So, this article develops a novel algorithm called the WECLO K-means-ALFPDC algorithm, which introduces additional metrics to handle the sparse dataset and categorizes the DPs as inner, inner edge, edge, or noisy points for enhancing cluster formation. The major goal is to consider the distinct variances between the sizes and orientations of the DPs nearer to the ClustCenters and edges. To find such variations, two novel local measures called CR and CO are introduced. The $CR \in [-1,1]$ is a parameter. If $CR > 0$, then the DP is marked as an inner point and if $CR < 0$, then the DP is an edge point. It is simpler to differentiate DPs in the ClustCenters and cluster boundaries. The CO is a measure of how well a DP fits in with its surroundings. If $CO > 0$, then the DPs have an approximately similar orientation to their surroundings and are marked as an edge point. Based on these measures, the ClustCenters are chosen at all iterations and more precise clusters are formed. Thus, the GED clustering is efficiently improved for enhancing their classification.

The following are the remaining portions of this article: Section 2 investigates the study on tumor classification from GED. Section 3 describes the WECLO K-means-ALFPDC algorithm and Section 4 demonstrates its performance. Section 5 summarizes the study and discusses future work.

2. Literature survey

Gclust, a parallel technique was developed [18] to group entire or partial gene data. In this model, a new multithreading mechanism and a rapid gene evaluation technique were adopted by the Sparse Suffix Arrays (SSAs) to speed up the grouping. Also, gene similarity across any pairs was determined according to their Maximal Exact Matches (MEMs). But, its scalability was less because the parallelism was not optimized.

A Distributed Density-based Hesitant Fuzzy Clustering (DDHFC) was introduced [19] using Apache spark to group the GED. In this method, a new weighted correlation metric was applied as a measurement for gene similarity analysis. Also, multi-valued inputs under HFC rules were integrated to increase the robustness to noise bias impacts. But, it needs to use a more informative GED to improve decision-making.

A new hybrid fuzzy means clustering and the majority vote was designed [20] to estimate the missing values in the GMD. But, the parameters were not optimized, which impacts the pre-processing stage. Also, it needs to consider the local and global strategy of missing values compared to a unified method.

Integrated grouping using non-Negative Matrix Factorization (nNMF) was developed [21] to group many varieties of interrelated databases that were tested on similar malignant instances. First, patient data were linked together using consensus matrices created by the nNMF on all varieties of data. A complete network pattern was generated by merging various networks and maximizing the correlation robustness. The resultant network data was also subjected to spectral grouping to estimate the groupings. But, the cost was high while using multiple categories of high-dimensional data.

A PCA and K-means clustering was suggested [22] to choose the most appropriate features and categorize microarray data of L1000 landmark genes. But, it needs to analyze the correlation between genetic and medicinal variables by considering the coding and non-coding genetic variants.

An Enhanced ANFIS (EANFIS) scheme was developed [23] to categorize the tumor genes. Initially, the input data was pre-processed by the Ensemble Kalman Filter (EnKF). Then, the genes having similar properties were grouped by an Adaptive Density-Based Spatial Clustering with Noise (ADBSCAN) method. But, it has high computation cost while increasing the number of samples.

An efficient technique was designed [24] to choose the most discriminatory sequences from high-dimensional GMD for malignant categorization. Initially, the affinity matrices were created for samples and genes, which define the correlation data between samples and genes, correspondingly. After that, the dual latent interpretation training was modeled using nNMF of the similarity matrices. The low-dimensional latent interpretation matrix of sample space was regarded as a pseudo-label matrix to assist data projection. The sample projection matrix was combined with the gene space latent interpretation matrix. Further, an alternated method was adopted to solve the final optimization issue. But, it has a high computational complexity while increasing the amount of data.

2.1 Research gap

Most of the existing algorithms find a relationship between the data with the neighborhood only based on distance and density. The cluster formation of these two metrics is not sufficient when the dataset is sparse. So, new metrics are required to find high-quality clusters.

3. Proposed methodology

In this section, the WECLO K-means-ALFPDC algorithm is explained briefly. Table 1 lists the notations used in this study.

Let the data matrix of size $a \times b$, $F = \{f_1, \dots, f_a\}$ be the group of a instances with f_x defining a b-dimensional vector. The CE exists for a consensus partition δ according to the collection of r partitions $(\delta_1, \dots, \delta_r)$ of the corpus F . The Co-association Matrix-based (CaM) consensus method is developed using this WECLO K-means and is used to fuse different groupings. This is according to the supposition that the effectiveness of the base partitions in the combination varies from single base partition to the next. Also, remember that entities with similar clusters still have different levels of efficacy. To reach the final consensus, they should have a variety of contributions.

Table 1. Lists of notations

Notations	Description
$a \times b$	Data matrix
F	GEM Corpus
δ	Consensus partition
r	Number of partitions
a	Number of instances
δ'	Ground truth groupings
k	Number of adjacent
p_1 and p_2	Two point masses
$\vec{F}_{p_1 p_2}$	Force between p_1 and p_2
$D_{p_1 p_2}$	Distance between p_1 and p_2
G	Gravitational constant
$\widehat{D}_{p_1 p_2}$	Orientation of the line, which links p_1 and p_2
x	Data
$\widehat{D}_{xy}$	Orientation information between x and its adjacent, as well as, the group of p_y
$\vec{F}_x$	Local resultant force
p_x	Mass of a DP f_x
CR_x	CR of the DP f_x
$\vec{D}_{yx}$	Displacement vector from y^{th} adjacent of f_x to it
CO_x	CO of f_x
$\vec{F}_y$	Force related to the adjacent of f_x
CR_{thres}	Centrality threshold
M	Initial momentum
θ	Threshold size of f_x
ψ	Force size threshold
Q	Queue

This WECLO K-means predicts the proximity between the base and ground truth groupings δ' via the direct and indirect evaluations of every cluster in base partitions. Then, based on proximity, distinct weights are given to each cluster. In this case, dynamic PDC and lion optimization K-means ensemble based on the 2 measures called CO and CR is used to efficiently select the ClustCenters and ideal base partitions.

Fig. 1 depicts the model of the WECLO K-means-ALFPDC algorithm for GED clustering. It encompasses the following tasks: (1) CE using K-means with lion optimization algorithm; (2) weight allocation for clusters using an adaptive PDC, which supports the selection of ClustCenters according to the CO and CR, as well as, performs modification on the real CaM; and (3) CaM partition based on the Cluster-based Similarity Partitioning (CSPA) to get the target consensus clustering. Initially, a lion optimization is applied to the feature vectors before every K-means clustering and then K-means is conducted on the new database [27]. A relatively maximum range of the parameter k and the number of clusters are chosen to get the neighborhood allocation of data so that many groups are formed in every clustering.

Thus, this algorithm includes 2 primary objectives: (i) to perform diverse base partitions, which provide multiple viewpoints on the real data, produce a very robust consensus and increase the possibility of finding the ground truth outcome; (ii) to do clustering in low-dimensional feature space for simplifying the dimensionality reduction.

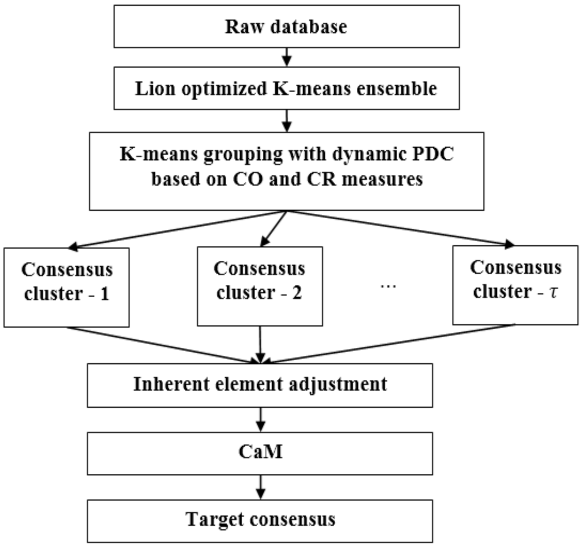
Raw database
Lion optimized K-means ensemble
K-means grouping with dynamic PDC based on CO and CR measures
Consensus cluster - 1
Consensus cluster - 2
...
Consensus cluster - τ
Inherent element adjustment
CaM
Target consensus

Figure. 1 Overview of WECLO K-means with ALFPDC algorithm

3.1 Dynamic configuration of cluster centroids based on local force of gravity in K-means clustering

A dynamic PDC is utilized during all K-means clustering iterations to choose the ClustCenters, preventing loss from biased input as the ClustCenters and ambiguity from automatically choosing the ClustCenters. The PDC approach is used with K-means grouping at all iterations to automatically choose the ClustCenters. The ClustCenters are determined by their distance from the next bigger density element and their substantial neighborhood density. However, in actual use, it includes complications and subjectivity. Observe that a cluster contains a large number of density peak components. Choosing the right number ClustCenters is difficult if the distribution of these items is uniform. Based on the input and neighborhood density distributions, a dynamic choice of ClustCenters is used to address this issue.

The statistical distribution is first measured by computing the kurtosis of the gap from the nearby bigger density factor of all data. The data element with a neighborhood density greater than the mean of the neighborhood densities of all the other data is then selected to serve as the ClustCenter. Thus, the conditions to select the ClustCenters are: (a) to measure the CO and CR for every instance for determining the neighborhood density and proximity of every instance and (c) group from the highest density and proximity according to the CO and CR.

The local force of gravity in GEM clustering defines the association between a DP and its proximate adjacent. According to the concept of gravity, the attraction force between 2 point masses (p_1 and p_2) is calculated by

$$\vec{F}_{p_1p_2} = G\frac{p_1p_2}{D^2_{p_1p_2}}\widehat{D}_{p_1p_2} \tag{1}$$

In Eq. (1), $\vec{F}_{p_1p_2}$ is the force between p_1 and p_2, $D_{p_1p_2}$ is the distance between p_1 and p_2, G defines the gravitational constant and the unit vector $\widehat{D}_{p_1p_2}$ stands for the orientation of the line, which links p_1 and p_2.

3.1.1. Local resultant force in K-means clustering

In a vicinity area, consider that the distances between the present data and its varied adjacent do not differ noticeably so Eq. (1) is simplified by Eq. (2).

$$\vec{F}_{p_1p_2} = Gp_1p_2\widehat{D}_{p_1p_2} \tag{2}$$

The resulting force on data x because of its k-NN is determined by

$$\vec{F}_x = \sum_{y=1}^{k} \vec{F}_{xy} = G p_x \sum_{y=1}^{k} p_y \, \widehat{D}_{xy} \tag{3}$$

In Eq. (3), the unit vector $\widehat{D}_{xy}$ summarizes the orientation information between x and its adjacent, as well as, the group of p_y values contributing the part of weighting variables in composing the forces in the vicinity. According to this perspective, data having greater masses have higher influence over their adjacent, whereas data with lesser masses have a greater sensitivity to the influence from their adjacent. So, a novel description of the local resultant force is considered that substitutes Newton's idea of gravitation in the clustering:

$$\vec{F}_x = \frac{1}{p_x} \sum_{y=1}^{k} \widehat{D}_{xy} \tag{4}$$

In Eq. (4), the mass p_x of a DP f_x is described by

$$p_x = 1 \Big/ \sum_{y=1}^{k} D_{xy} \tag{5}$$

Based on Eq. (5), the masses of these DPs will grow higher since they will be closer to their adjacent in larger density regions. But, the masses of the DPs get lesser in lower-density regions. So, the mass of a DP is termed an alternated method to determine the neighborhood density in clustering. This is also employed in the local resultant force in Eq. (4): in greater density regions, a DP is enclosed by adjacent DPs in a highly regular manner that provides a limited size for $\sum_{y=1}^{k} \widehat{D}_{xy}$ and a great p_x. In lower-density regions, a DP is normally enclosed by adjacent DPs in a regular manner that provides an imbalanced resulting force. The size of $\sum_{y=1}^{k} \widehat{D}_{xy}$ in low-density regions is normally greater and p_x is lesser. Based on Eq. (4), the resulting force $\vec{F}_x$ can contain a high size and a high dissimilar orientation toward the ClustCenter.

3.1.2. Correlation across local resultant forces

The fundamental concept of this presented grouping algorithm is that there is a major variance between the local resultant forces of the DPs nearer to the ClustCenters and those at the margin of the cluster. The local resultant force defines the correlation between all DPs and their adjacent.

To consider the advantage of the data contained in the local resultant force, two local clustering measures according to the local force of gravity called the CR and the CO are introduced. The CR of the DP f_x is described by

$$CR_x = \sum_{y=1}^{k} \frac{\cos(\vec{F}_y, \vec{D}_{yx})}{k} \tag{6}$$

In Eq. (6), $\vec{D}_{yx}$is the displacement vector from y^{th} adjacent of f_x to it and k denotes the number of adjacents. The DP has a CR value $CR_x > 0$ defines that most of the local resultant forces of its adjacent are spotted in Eqs. (7) and (8)

$$-1 \le cos(\vec{F}_y, \vec{D}_{yx}) \le 1 \tag{7}$$

$$-k \le \sum_{y=1}^{k} cos(\vec{F}_y, \vec{D}_{yx}) \le k \tag{8}$$

This algorithm possesses the CR property in Eq. (9)

$$-1 \le CR_x \le 1 \tag{9}$$

Similarly, the CO of f_x is described by

$$CO_x = \sum_{y=1}^{k} (\vec{F}_x \cdot \vec{F}_y) \tag{10}$$

In Eq. (10), $\vec{F}_x$ is the local resultant force of f_x and $\vec{F}_y$ is the force related to its adjacent. The CO defines the coherence between a certain DP and its adjacent. The DP having $CO_x > 0$ defines that its local resultant force has an approximately similar orientation to its adjacent and it is situated in the margin.

Thus, the variances between interior points and margin points for the mass, the size of the local resultant force, the CR and the CO are observed as:

- DPs with huge densities, limited sizes of local resultant forces, high CR and less CO values are recognized as interior points.
- Margin points contain the small masses, large sizes of the local resultant forces, small CR and high CO values.

By considering the variances between interior and margin DPs, more and LICCs are discovered. If the groupings have fewer interior DPs and more margin DPs, then those clusters are termed as LICCs, which are discarded.

In contrast, if the clusters have more interior DPs and fewer margin DPs, then those clusters are termed more informative composite clusters, which are considered to create the CaM. Moreover, the CSPA is used to divide the resulting consensus clustering.

Algorithm for WECLO K-means-ALFPDC

```
Input: GEM databases $(F)$, the number of adjacent $(k)$, the centrality threshold $CR_{thres}$ and the initial momentum $(M)$
Begin
for $(all\ data\ points\ f_x)$
    Determine its mass $p_x$ using Eq. (5);
    Determine its local resultant force $\vec{F}_x$ using Eq. (4);
    Determine $CR_x$ and $CO_x$ using Eq. (6) to Eq. (10), respectively;
    if $(CR_x < CR_{thres}$ && $CO_x \geq 0)$
        Identify $f_x$ as a margin point;
    elseif $(CR_x \geq 0)$
        Identify $f_x$ as an interior point;
    end if
end for
while $(not\ every\ interior\ point\ is\ clustered)$
    Discover such an unclustered interior point $f_x$ and include it in the new cluster $C_{f_x}$;
    Set the threshold size of $f_x$ as $\theta = M \cdot \psi$ ($\psi$ defines a force size threshold);
    Set a queue $Q = f_x$;
    while $(Q \neq \emptyset)$
        Arrange every $k-1$ adjacent of $f_x$ based on their distances and get a sorted set $\Pi$;
        for $\left(all\ points\ f_{\pi_y}\right)$ in $\Pi$;
            if $(\vartheta \leq \theta)$
                if $\left(f_{\pi_y}\ has\ been\ labeled\ as\ a\ margin\ point\right)$
                    Modify $\vartheta$ as $\vartheta = \vartheta + \vec{F}_{f_{\pi_y}}$;
                    Include $f_{\pi_y}$ into $C_{f_x}$;
                elseif $\left(\begin{matrix} f_{\pi_y}\ is\ labeled\ as\ an\ interior\ point \\ and\ f_{\pi_y} is\ not\ been\ grouped \end{matrix}\right)$
                    Include $f_{\pi_y}$ into $C_{f_x}$;
                    Dequeue $f_x$ from $Q$ and Enqueue $f_{\pi_y}$ into $Q$;
                elseif $\left(\begin{matrix} f_{\pi_y}\ is\ not\ labeled\ f_{\pi_y}\ is\ not \\ grouped \end{matrix}\right)$
                    Include $f_{\pi_y}$ into $C_{f_x}$;
                end if
            end if
        end for
    end while
end while
Cluster all data instances in the given database;
Find the less informative and more informative clusters;
Discard the less informative clusters and use more informative clusters to create CaM;
Apply CSPA to divide the final consensus grouping;
End
```

Thus, the abovementioned algorithm describes the presented clustering, wherein data elements are coupled as groups (clusters). Particularly, the interior points can couple to their adjacent and include such adjacent into their groups, when the edge points are considered as outliers. According to this, less informative and more informative clusters are created.

4. Results and discussion

In this part, the effectiveness of the WECLO K-means-ALFPDC algorithm is assessed by executing it in MATLAB 2017b with 5 distinct kinds of tumor-associated genomic corpora. As well, a comparative analysis is performed with the classical algorithms implemented by using the considered corpora, including WECR K-means [16], WECR K-means-PDC [17], DDHFC [19], nNMF [21] and EANFIS [23] in terms of various metrics. The details about 5 various corpora are given below.

1. ***Prostate Cancer:*** Records on 50 prostate tumors and 52 typical cases make up the 102 total DPs. There are 10509 genomes in all records [25].
2. ***SRBCT Data:*** There are 2308 genomes in each of the 83 chunks of information. It is a case of Burkett's Lymphoma (BL), the Ewing tumor family (EWS), Neuro Blastoma (NB), and Rhabdomyo Sarcoma (RMS) (RMS). For training, there are 63 samples, and 20 samples are chosen for testing. There are 8, 23, 12, and 20 DPs in the BL, EWS, NB, and RMS training sets. For the BL, EWS, NB, and RMS tests, the data set includes 3, 6, 6, 6 and 5 samples [26].
3. ***Leukemia:*** It holds 7129 genes in use over 72 models. It comprises 72 data, 25 data on Acute Myeloid Leukemia (AML) and 47 data on Acute Lymphoblastic Leukemia (ALL). A basis of the GEM values is obtained from 63 bone marrow data and 9 peripheral blood data [27].
4. ***Lymphoma:*** 24 Generic B-like and 23 files activated DLCLs are involved in this collection. 42 files of Diffuse Large B-cell lymphoma (DLBCL), 9 files of Follicular Lymphoma (FL), and 11 files of Chronic Lymphocytic Leukemia (CLL) are all included. The total number of DNA/RNAs in this dataset is 4026. In addition, the KNN assigns a few rates that are lacking [28].
5. ***Breast cancer:*** This is the Wisconsin breast cancer prognostic dataset from the UCI

Table 1. Gene database characteristics

Databases	#DNA/RNA	#Sample	#Tag
Leukemia	7129	72	2
Lymphoma	4026	62	3
Prostate cancer	10509	102	2
SRBCT	2308	83	4
Breast cancer	-	569	2

machine learning repository [29]. B and M samples are included in this section. It has 569 samples and 32 features, including 30 true input features. There are 357 cases of B and 212 examples of M in the 569 Details about these databases may be found in Table 1.

4.1 Accuracy

It is a measure of how many properly grouped DPs there are out of all the ones in the database.

$$Accuracy = \frac{1}{n}\sum_{i=1}^{n}\vartheta\big(r_i, map(S_i)\big) \qquad (11)$$

In Eq. (11), r_i is the original group tag and S_i is the index value acquired by grouping. If $u_1 = u_2$, then $\vartheta(u_1, u_2) = 1$; or else, $\vartheta(u_1, u_2) = 0$. Higher grouping efficacy is defined by a wider range of accuracy.

Fig. 2 demonstrates the accuracy achieved by various ensemble clustering algorithms implemented on 5 distinct databases. It addresses that the WECLO K-means-ALFPDC ensemble accomplishes a greater accuracy than all existing algorithms to group the vast quantity of GED. That is, in the case of grouping the SRBCT database, the accuracy of WECLO K-means-ALFPDC is 26.98% superior to the DDHFC, 20.57% superior to the nNMF, 13.75% superior to the EANFIS, 9.67% superior to the WECR K-means and 4.9% superior to the WECLO K-means-PDC algorithms.

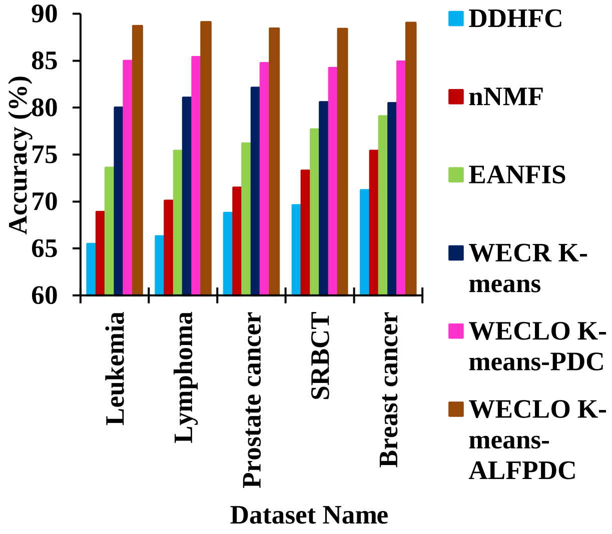

Figure. 2 Accuracy vs. databases

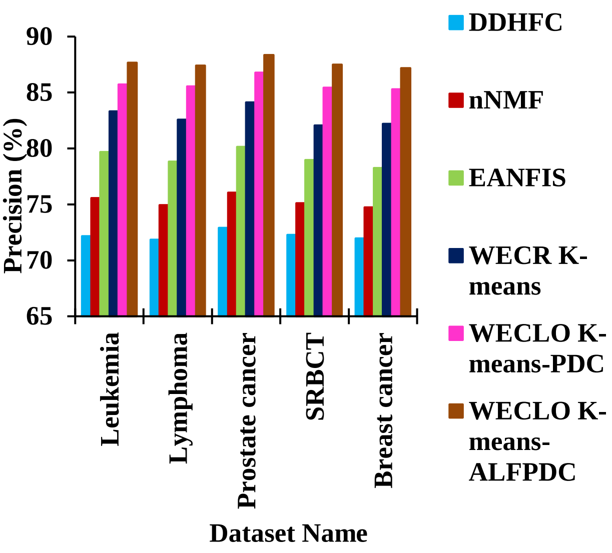

Figure. 3 Precision vs. databases

4.2 Precision

It is the ratio of accurately grouped data items at True Positive (TP) and False Positive (FP) rates.

Fig. 3 exhibits the precision achieved by the different ensemble grouping algorithms applied to the 5 distinct databases. It notices that the WECLO K-means-ALFPDC ensemble realizes a higher precision than the existing algorithms for clustering the large GED corpora. That is, in the case of grouping the leukemia database, the precision of WECLO K-means-ALFPDC is 21.52% increased than the DDHFC, 16.01% increased than the nNMF, 10.03% increased than the EANFIS, 5.22% increased than the WECR K-means and 2.29% increased than the WECLO K-means-PDC algorithms.

4.3 Recall

It is the rate of accurately grouped data items at TP and False Negative (FN) rates.

Fig. 4 portrays the recall achieved by various ensemble clustering algorithms tested on 5 distinct databases. It addresses that the WECLO K-means-ALFPDC ensemble algorithm realizes greater recall than the existing algorithms for clustering the large GED corpora. That is, in the case of clustering the breast cancer database, the recall of WECLO K-means-ALFPDC is 23.4% superior to the DDHFC,

Received: September 15, 2022. Revised: October 12, 2022.

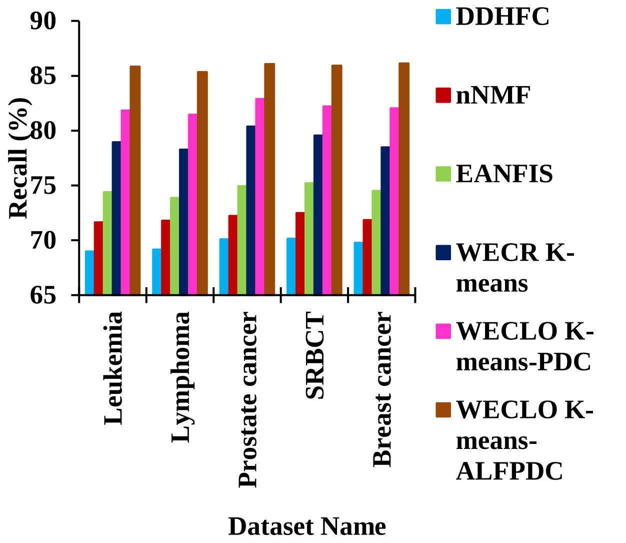

Figure. 4 Recall vs. databases

19.85% superior to the nNMF, 15.6% superior to the EANFIS, 9.74% superior to the WECR K-means and 4.96% superior to the WECLO K-means-PDC algorithms.

4.4 Root mean squared error (RMSE)

It measures the efficacy of the clustering algorithm as:

$$RMSE = \sqrt{\frac{\sum_{j=1,\dots,p}^{i=1,\dots k} \sum_{a=1}^{n_{ij}} (u_a - \bar{\mu}_{ij})^2}{\sum_{j=1,\dots,p}^{i=1,\dots,k} (n_{ij} - 1)}} \tag{12}$$

In Eq. (12), k is the amount of groups, p is the amount of independent attributes in the corpus, $\bar{\mu}_{ij}$ is the average of elements in attribute j and group i and n_{ij} is the amount of elements that are in attribute p and group k.

Fig. 5 depicts the RMSE achieved by various ensemble grouping algorithms tested on 5 distinct databases. It observes that the WECLO K-means-ALFPDC ensemble algorithm accomplishes a minimum RMSE compared to the existing grouping algorithms for the large-scale GED corpora. That is, in the case of clustering the prostate cancer database, the RMSE of WECLO K-means-ALFPDC is 23.55% less than the DDHFC, 22.61% less than the nNMF, 21.48% less than the EANFIS, 20% decreased than the WECR K-means and 6.08% decreased than the WECLO K-means-PDC algorithms.

4.5 Time cost

It determines the time needed for obtaining a final consensus using the different clustering algorithms.

Fig. 6 portrays the average time cost (sec) of the different ensemble grouping algorithms tested on 5 distinct databases. It observes that the WECLO K-means-ALFPDC ensemble algorithm accomplishes a minimum time cost compared to the existing grouping algorithms for the large-scale GED corpora. That is, in the case of clustering the SRBCT database, the average time cost of WECLO K-means-ALFPDC is 82.14% less than the DDHFC, 74.11% less than the nNMF, 67.61% less than the EANFIS, 38.22% decreased than the WECR K-means and 17.95% decreased than the WECLO K-means-PDC algorithms.

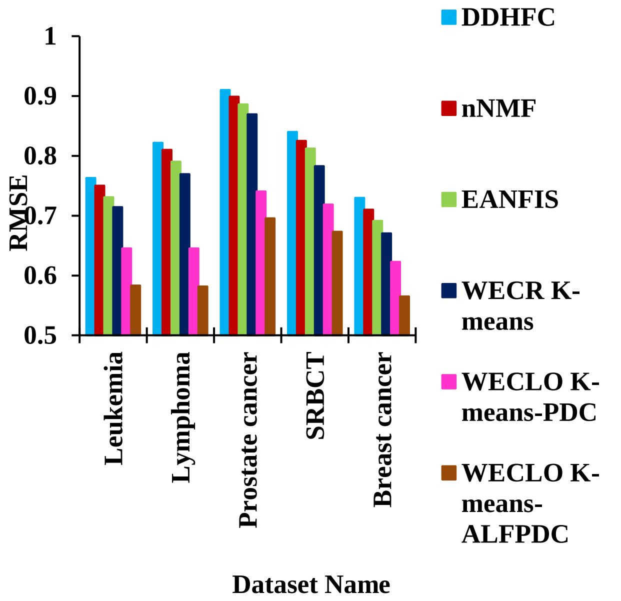

Figure.5 RMSE vs. databases

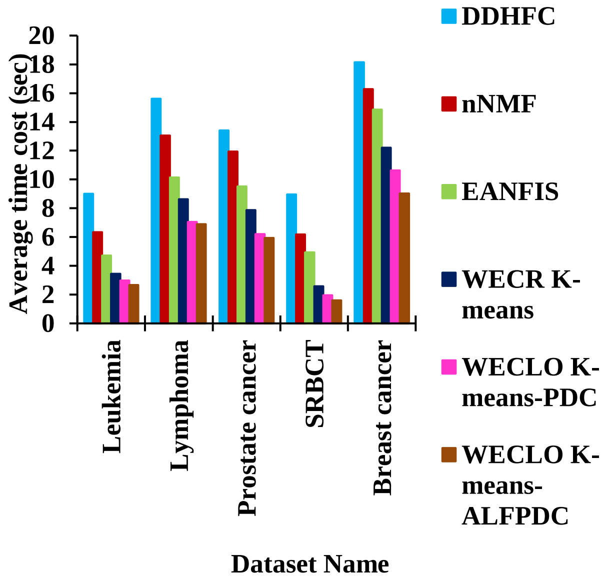

Figure.6 Average time cost vs. databases

5. Conclusion

In this study, the WECLO K-means-ALFPDC algorithm was designed for enhancing the efficiency of clustering the large-scale GED. It was used to adjust the clusters by considering the distinct variances between the sizes and orientations of the DPs nearer to the ClustCenters and edges in every iteration. The lion optimization scheme was applied that treats inter-and intra-group gaps as fitness values for efficient clustering. Also, CR and CO measures were determined for all DPs, which support choosing the ClustCenters and removing the LICCs with more margin points and fewer interior points. Thus, the more precise composite groupings were obtained for GED classification. At last, the experimental findings proved that the WECLO K-means-ALFPDC algorithm on leukemia, lymphoma, prostate cancer, SRBCT and breast cancer corpora has 88.7%, 89.1%, 88.42%, 88.38% and 89.04% accuracy, correspondingly than the DDHFC, nNMF, EANFIS, WECR K-means and WECLO K-means-PDC algorithms.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization, methodology, software, validation, Sangeetha; formal analysis, investigation, Kousalya; resources, data curation, writing—original draft preparation, Sangeetha; writing—review and editing, Sangeetha; visualization, supervision, Kousalya.

References

[1] O. M. Sigalova, A. Shaeiri, M. Forneris, E. E Furlong, and J. B. Zaugg, "Predictive Features of Gene Expression Variation Reveal Mechanistic Link with Differential Expression", *Molecular Systems Biology*, Vol. 16, No. 8, pp. 1-24, 2020.

[2] M. Dashtban and M. Balafar, "Gene Selection for Microarray Cancer Classification using a New Evolutionary Method Employing Artificial Intelligence Concepts", *Genomics*, Vol. 109, No. 2, pp. 91-107, 2017.

[3] H. A. Chowdhury, D. K. Bhattacharyya, and J. K. Kalita, "(Differential) Co-Expression Analysis of Gene Expression: A Survey of Best Practices", *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 17, No. 4, pp. 1154-1173, 2019.

[4] V. Tyagi and A. Mishra, "A Survey on Different Feature Selection Methods for Microarray Data Analysis", *International Journal of Computer Applications*, Vol. 67, No. 16, pp. 36-40, 2013.

[5] K. Tadist, S. Naja, N. S. Nikolov, F. Mrabti, and A. Zaha, "Feature Selection Methods and Genomic Big Data: A Systematic Review", *Journal of Big Data*, Vol. 6, No. 1, pp. 1-24, 2019.

[6] B. Sahu, S. Dehuri, and A. Jagadev, "A Study on the Relevance of Feature Selection Methods in Microarray Data", *The Open Bioinformatics Journal*, Vol. 11, No. 1, pp. 117-139, 2018.

[7] N. Almugren and H. Alshamlan, "A Survey on Hybrid Feature Selection Methods in Microarray Gene Expression Data for Cancer Classification", *IEEE Access*, Vol. 7, pp. 78533-78548, 2019.

[8] M. A. Hambali, T. O. Oladele, and K. S. Adewole, "Microarray Cancer Feature Selection: Review, Challenges and Research Directions", *International Journal of Cognitive Computing in Engineering*, Vol. 1, pp. 78-97, 2020.

[9] J. Oyelade, I. Isewon, F. Oladipupo, O. Aromolaran, E. Uwoghiren, F. F. Ameh, and E. Adebiyi, "Clustering Algorithms: Their Application to Gene Expression Data", *Bioinformatics and Biology Insights*, Vol. 10, pp. 237-253, 2016.

[10] H. W. Nies, Z. Zakari, M. S. Mohama, W. H. Chan, N. Zaki, R. O. Sinnott, and J. M. Corchado, "A Review of Computational Methods for Clustering Genes with Similar Biological Functions", *Processes*, Vol. 7, No. 9, pp. 1-18, 2019.

[11] Z. Yu, L. Li, J. Liu, J. Zhang, and G. Han, "Adaptive Noise Immune Cluster Ensemble using Affinity Propagation", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 27, No. 12, pp. 3176-3189, 2015.

[12] J. Wang, "Semi-Supervised Learning using Ensembles of Multiple 1D-Embedding-based Label Boosting", *International Journal of Wavelets, Multiresolution and Information Processing*, Vol. 14, No. 2, pp. 1-33, 2016.

[13] G. Casalino, G. Castellano, and C. Mencar, "Data Stream Classification by Dynamic Incremental Semi-Supervised Fuzzy Clustering", *International Journal on Artificial Intelligence Tools*, Vol. 28, No. 8, pp. 1-26, 2019.

[14] H. H. V. Engelen and H. H. Hoos, “A Survey on Semi-Supervised Learning”, *Machine Learning*, Vol. 109, No. 2, pp. 373-440, 2020.

[15] Z. Yu, Z. Kuang, J. Liu, H. Chen, J. Zhang, J. You, and G. Han, “Adaptive Ensembling of Semi-Supervised Clustering Solutions”, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 29, No. 8, pp. 1577-1590, 2017.

[16] Y. Lai, S. He, Z. Lin, F. Yang, Q, Zhou, and X. Zhou, “An Adaptive Robust Semi-Supervised Clustering Framework using Weighted Consensus of Random K-Means Ensemble”, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 33, No. 5, pp. 1877-1890, 2019.

[17] M. Sangeetha and R. Kousalya, “An Optimized Weighted Consensus Clustering with Removal of Less Informative Composite Clusters”, In: *Proc. of 9th International Conference on Computing for Sustainable Global Development*, pp. 392-399, 2022.

[18] R. Li, X. He, C. Dai, H. Zhu, X. Lang, W. Chen, and B. Niu, “Gclust: A Parallel Clustering Tool for Microbial Genomic Data”, *Genomics, Proteomics and Bioinformatics*, Vol. 17, No. 5, pp. 496-502, 2018.

[19] B. Hosseini and K. Kiani, “A Big Data Driven Distributed Density based Hesitant Fuzzy Clustering using Apache Spark with Application to Gene Expression Microarray”, *Engineering Applications of Artificial Intelligence*, Vol. 79, pp. 100-113, 2019.

[20] S. R. Kumaran, M. S. Othman, L. M. Yusuf, and A. Yunianta, “Estimation of Missing Values using Hybrid Fuzzy Clustering Mean and Majority Vote for Microarray Data”, *Procedia Computer Science*, Vol. 163, pp. 145-153, 2019.

[21] P. Chalise, Y. Ni, and B. L. Fridley, “Network-based Integrative Clustering of Multiple Types of Genomic Data using Non-Negative Matrix Factorization”, *Computers in Biology and Medicine*, Vol. 118, pp. 1-9, 2020.

[22] C. L. Clayman, S. M. Srinivasan, and R. S. Sangwan, “K-means Clustering and Principal Components Analysis of Microarray Data of L1000 Landmark Genes”, *Procedia Computer Science*, Vol. 168, pp. 97-104, 2020.

[23] P. Mishra and N. Bhoi, “Cancer Gene Recognition from Microarray Data with Manta Ray Based Enhanced ANFIS Technique”, *Biocybernetics and Biomedical Engineering*, Vol. 41, No. 3, pp. 916-932, 2021.

[24] X. Zheng and C. Zhang, “Gene Selection for Microarray Data Classification via Dual Latent Representation Learning”, *Neurocomputing*, Vol. 461, pp. 266-280, 2021.

[25] http://www.gems-system.org/

[26] http://www.biolab.si/supp/bi-cancer/projections /info/SRBCT.html

[27] http://cilab.ujn.edu.cn/datasets.html

[28] http://csse.szu.edu.cn/staff/zhuzx/Datasets.html

[29] http:/archive.ics.uci.edu/ml/index.php