IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 10, 2022, accepted April 4, 2022, date of publication April 13, 2022, date of current version April 22, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3167031

Double Weighted Ensemble Clustering for Cancer Subtypes Analysis

XIN ZHANG[1] **AND HUA HUO**[2]
[1]School of Information Engineering, Henan University of Science and Technology, Luoyang 471003, China
[2]Engineering Technology Research Center of Big Data and Computational Intelligence, Henan University of Science and Technology, Luoyang 471003, China

Corresponding author: Hua Huo (pacific_huo@126.com)

This work was supported by the National Natural Science Foundation of China under Grant 61672210.

ABSTRACT The era of big data provides the possibility of precision medicine. The most important idea we have for cancer is to divide and treat. Theoretically, each person's cancer should be different, so it is very necessary to make personalized treatment plans for different cancer patients. Subtype analysis of cancer can be viewed as a clustering problem, while ensemble clustering techniques are widely followed for their ability to combine multiple basic clusters into potentially better and more robust clusters. However, the reliability of the present ensemble clustering methods in cancer subtype analysis still needs to be improved. Therefore, we propose a double weighted ensemble clustering method (DWEC), which first derives the similarity matrix of each base cluster based on the local weighting method, and this process can be regarded as the first weighting based on clusters. Subsequently, the objective of finding the final partitions is regarded as an optimization problem, and the similarity matrix corresponding to each base cluster is weighted twice by the block coordinate descent algorithm to solve the optimal partitions result. The best experimental results were obtained in both labeled datasets and unlabeled cancer gene datasets, validating the superiority of the method. For cancer subtype analysis, although our proposed method did not show statistically significant differences in survival distributions of several subtypes in the subtype analysis of glioblastoma multiforme. However, it performed best in the results of the temporal test for all other four cancer gene data, and therefore, we conclude that our method is more effective for cancer subtype analysis compared with other methods.

INDEX TERMS Cancer subtypes analysis, ensemble clustering, double weighted ensemble clustering, similarity matrix, entropy.

I. INTRODUCTION

Cancer is the most complex disease faced by people in today's society, and more than 200 types of cancer have been found. At the same time, due to the dynamic changes of cancer genes, it is possible that genetic mutations, such as somatic mutations, copy number variations, altered gene expression profiles, and different epigenetic variations, are unique to each cancer. A variety of different molecular profiles can lead to a phenomenon that each cancer includes several subtypes, posing a formidable challenge to medical researchers. Each cancer with a different molecular structure requires a different treatment approach [1]. The heterogeneity of cancer is an important feature, which means that during the growth process of the tumor, after multiple divisions and proliferations, its daughter cells show molecular biology or genetic changes, so that the growth rate, invasive ability, and resistance of the tumor are affected. There are differences in drug sensitivity and prognosis [2]. It is an important scientific issue in oncology research to unearth the molecular subtypes inherent in cancer tissues and then understand the epigenetic regulation mechanism. In terms of incidence, the top three cancers are lung cancer, female breast cancer and colorectal cancer [3]. Data clustering is a very important method in the fields of data mining and machine learning. Its purpose is to divide a given dataset into clusters that each share common characteristics [4]. Therefore, the discovery of cancer subtypes using clustering algorithms has attracted a lot of attention. This solution can help clinicians develop precise treatments by combining methods that analyze the different molecular profiles between cancer patients and healthy subjects [5]. In the past decades of research on clustering algorithms, many kinds of methods have been proposed [6]–[17], but a major drawback is that these algorithms are good for data sets with

The associate editor coordinating the review of this manuscript and approving it for publication was M. Shamim Kaiser.

specific structures and cannot be applied to all data sets and are not universally applicable. Therefore, in view of the above defects, cluster ensemble was proposed and quickly became a hot research topic. In ensemble clustering, each input cluster is called a base cluster, and the final clustering result is called a consensus cluster.

We tried to use ensemble clustering to discover cancer subtypes, and by ensemble the results of different base clustering in the same data set, we obtained a more stable result that was better than other clusters. We can view this problem as an ensemble clustering problem. A large number of integrated clustering algorithms have been proposed in the past period [18]–[29]. The evidence accumulation clustering algorithm proposed by Fred and Jain is based on the co-association matrix. First, the connection matrix is obtained by whether two data objects belong to the same class in the same base clustering result, and then the connection matrix generated by each base cluster is combined through a voting mechanism to obtain The co-association matrix is finally used as the input of hierarchical clustering to obtain the ensemble clustering result [30]. However, the quality of base clusters plays a crucial role in the consistency process, and consistent results can be compromised by low-quality base clusters. In recent years, research results on the weight of base cluster members or selection measurement strategies have emerged one after another. For example: Li *et al.* proposed to jointly learn the data partition weights and the final consensus clustering connection matrix under the Bregman divergence framework, proposed a weighted clustering integration scheme, which weighted data vectors of different dimensions differently to obtain data clustering [5]. Huang *et al.* proposed the Normalized Group Consistency Index (NCAI) to assess the quality of base class clusters in an unsupervised manner, thereby weighting base class clusters according to their clustering effectiveness [31]. However, these methods are developed based on an implicit assumption that all clusters in the same base cluster have the same reliability. They usually treat each base cluster as an individual and assign a global weight to each base cluster, regardless of the diversity of the clusters within it. However, due to the noise and inherent complexity of real datasets, different clusters within the same cluster may have different reliability. It is necessary to respect the local diversity of the set and deal with the reliability of different clusters. Recently, D. Huang *et al.* proposed a new ensemble-driven clustering effectiveness measure and proposed a locally weighted co-correlation matrix to summarize the ensemble of different clusters The calculation of inter-entropy makes it easier for them to get rid of the influence of low-quality base cluster [28]. The method in the paper [28] uses a local weighting strategy based on set-driven cluster validity to refine the co-association matrix, and proposes the concept of locally weighted co-association matrix. The locally weighted co-association matrix can be regarded as a consensus function for cluster weighting, which is obtained by the local weighted average of the connection matrix of each base cluster. However, there is no theoretical guarantee of optimality for such a simple averaging method.

To address the above issues, this paper proposes a novel co-clustering framework for determining the optimal clustering of cancer datasets to assist in the analysis of cancer subtypes. We refer to the local weighting method in [28] to integrate the entropy and validity of clusters into a local weighting scheme to improve consistency performance. A cluster can be viewed as a local area within the corresponding basic cluster. The entropy of each cluster is estimated based on the entropy criterion for the cluster labels in the entire set. In particular, given a cluster, investigate its uncertainty by considering how objects within that cluster are grouped in multiple base clusters. On the basis of cluster uncertainty estimation, the reliability of clustering is measured by an ensemble-driven clustering index (ECI). After obtaining the locally weighted similarity matrix of each base cluster, the process of integrating the connection matrix of the base cluster into the final result is regarded as an optimization problem, and a new consensus function is proposed to construct the final cluster. Figure 1 is the flow chart of the proposed algorithm.

The main contributions of our method are summarized as follows:

Our method not only integrates the uncertainty and validity of clusters into a local weighting scheme, but also fully considers the uncertainty of clusters in the same base clustering. In addition, a new consensus function is proposed to theoretically support the optimality of the final clustering result.

Multiple experiments are conducted on a large number of real data and cancer datasets, and the results demonstrate the superiority of the proposed ensemble clustering method in terms of clustering quality and efficiency.

The rest of this article is organized as follows. Related work will be presented in Section 2. The formulation of the ensemble clustering problem will be given in Section 3. The ensemble clustering method proposed in this paper will be introduced in Section 4. Experimental results are reported in Section 5. Section 6 summarizes the full text.

II. RELATED WORKS

In recent years, with the rapid development of whole-genome sequencing and bioinformatics technology, cluster analysis of gene expression profiles has become an important research topic in the diagnosis of cancer subtypes, which helps to provide more precise medical treatment for cancer patients. For example, Ronglai Shen *et al.* developed an integrated clustering of joint latent variable models called iCluster, which integrates flexible modeling of associations between different data types and variance covariance structure in data types in a single framework, while reducing the dimensionality of the data set, using an expectation maximization algorithm Likelihood inference is performed to finally identify cancer subtypes characterized by DNA copy number variations and gene expression [32]. Wang *et al.* developed Similarity

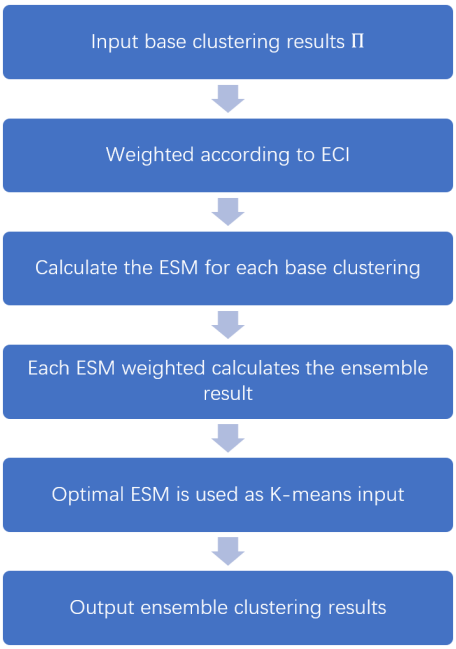

FIGURE 1. **Flow diagram of the proposed approach.**

Network Fusion (snF) to identify cancer by constructing a network of samples (such as patients) for each available data type, and then efficiently fusing these samples into a network that represents the full spectrum of the underlying data subtype [33]. Li *et al.* proposed Bregmannian Consensus Clustering (BCC), which generalizes the loss between the consensus clustering result and all input clusters from the traditional Euclidean distance to the general Bregman loss, and to the weighted and semi-supervised [5]. Xu *et al.* used yeast dataset, human serum dataset and Arabidopsis dataset to construct a minimum spanning tree (MST) expressing multi-dimensional gene expression profiles, and proposed a clustering algorithm based on MST Cluster these genetic data [34]. Yu *et al.* proposed a random double-clustering clustering ensemble framework (RDCCE) for tumor clustering based on gene expression data. RDCCE uses a randomly selected clustering algorithm in the ensemble to generate a representative set of features, and then assign the samples to the corresponding clusters according to the grouping results [25]. Lock and Dunson *et al.* proposed an integrated statistical model that clusters objects individually for each data source. Using an extensible Bayesian framework, both consensus clustering and source-specific clustering were estimated and eventually used in the subtyping of breast cancer tumor samples [35]. Under the current scale of big data, the dimension of data objects is usually very high, and gene selection is also a method for clustering. Tang *et al.* integrated feature extraction and feature selection into a unified framework and designed an unsupervised linear feature selection projection (FSP) for suppressing the effects of noise while making FSP robust to noise [36]. Tang *et al.* proposed a multi-view unsupervised feature selection (MV-UFS) model to preserve diversity and consensus learning through cross-view local structure, abbreviated as CvLP-DCL, which utilizes the shared and discriminative information between different views to Each view is projected into the label space, and finally discriminative features can be selected from different views [37].

In order to improve the robustness and stability of clustering methods, researchers have begun to focus on ensemble clustering, and there are many ensemble clustering methods. Pair co-occurrence-based methods typically construct co-association(CA) matrix by considering the number of occurrences of two objects in the same cluster across multiple base clusters. Using the CA matrix as the similarity matrix, the traditional clustering method can be used to construct the final clustering result [21], [28], [30] and [38]. Fred *et al.* first proposed the concept of CA matrix and proposed the Evidence Accumulation Clustering (EAC) method. The idea of evidence accumulation clustering is to combine the results of multiple clusters into a single data partition, and each cluster The results are treated as an independent data organization evidence and consistent data partitions are extracted from the merged evidence [30]. Wang *et al.* extended the EAC method with the construction of the correlation matrix considering the cluster size of the original clusters and proposed a probabilistic accumulation method [21]. Huang *et al.* proposed an ensemble clustering method based on ensemble-driven clustering uncertainty estimation and a local weighting strategy. The labels of the clusters in the entire set were considered through the entropy criterion, and the uncertainty of each cluster was estimated for weighting. The local diversity in the ensemble further proposes two new consensus functions [28]. Lourenço *et al.* proposed a consensus clustering method based on the EAC paradigm, which is not limited to clear partitions and takes full advantage of the nature of the covariance matrix to determine the probabilistic assignment of data points to clusters by minimizing the Bregman scatter between the observed co-association frequencies and the corresponding co-occurrence probabilities expressed as unknown assignment functions [38]. The graph partitioning-based approach solves the integration clustering problem by constructing a graph model to reflect the integration information. Strehl *et al.* formalized the clustering ensemble problem as a combinatorial optimization problem based on mutual information sharing, and proposed three graph partition-based ensemble clustering algorithms, CSPA, HGPA and MCLA [39]. The median division-based approach formulates the integrated clustering problem as an optimization problem whose goal is to find a clustering result by maximizing the similarity between this cluster and multiple base clusters. Huang *et al.* introduced the concept of hyperobjects, which are compact and adaptive representations of integrated data that greatly facilitate computation. The ensemble clustering problem is transformed into a binary linear programming problem by means of a probabilistic formulation. The constrained objective function is represented as a factor graph, and the maximum product belief propagation is used to

generate solutions that are insensitive to initialization and converge to the neighborhood maximum [40].

However, the above studies still have significant limitations in practical applications, and the clustering analysis for cancer gene expression profiles still cannot achieve the desired results. To this end, we propose a new ensemble clustering framework for determining the optimal clusters of cancer datasets by fully combining the ideas based on pairwise co-occurrence methods and median-based partitioning methods.

III. PRELIMINARY

A. ENSEMBLE CLUSTERING

Suppose there is a dataset $X = \{x_1, x_2, \ldots, x_N\}$ consisting of N data objects, where x_i represents the ith data object. he dataset X is clustered M times to obtain M partitions, where each partition contains a certain number of clusters. Formally, we represent the set Π of M-based clusters as follows:

$$\Pi = \left\{\pi^1, \pi^2, \ldots, \pi^M\right\} \tag{1}$$

where

$$\pi^m = \{c_1^m, c_2^m, \ldots, c_{n^m}^m\}, \tag{2}$$

where Π^m represents the mth base cluster in Π, c_i^m represents the ith cluster in π^m, and n^m represents a total of n^m clusters in π^m. Each base cluster is a collection of multiple samples, and different clusters in the same base cluster do not intersect with each other. The following conditions must be met here: $\forall \pi^j \in \Pi, \cup_{i=1}^{n^m} c_i^j = X, c_i^j \cap c_k^j = \emptyset \; s.t. \; i \neq k$. Assuming $Cls^m(x_i) = c_i^m$, it means that the data x_i belongs to the ith cluster in the mth base cluster. For convenience, all clusters in the base cluster set Π are denoted as:

$$C = \{c_1, c_2, \ldots, c_{n_c}\} \tag{3}$$

where c_i represents the ith cluster, and n_c represents a total of n_c clusters in the base cluster set Π, i.e., $n_c = n^1 + n^2 + \ldots + n^M$.

Perform M times of clustering on the dataset X to obtain M partitions $\Pi = \left\{\pi^1, \pi^2, \ldots, \pi^M\right\}$, and each partition can get its connection matrix, ie, $CM = \left\{CM^1, CM^2, \ldots, CM^M\right\}$. The connection matrix is defined as follows.

Definition 1 (Connection Matrix): The connection matrix CM^m that partition π^m is an N × N symmetric square matrix, which reflects whether the two data objects in the division are grouped into the same cluster. Therefore, CM^m can be used to represent the partitions π^m, where the (u, v)th term is expressed as follows:

$$CM_{uv}^m = \begin{cases} 1, & if \; Cls^m(x_u) = Cls^m(x_v) \\ 0, & otherwise \end{cases} \tag{4}$$

B. INFORMATION ENTROPY

In information theory, entropy is a measure of uncertainty associated with random variables. Joint entropy is a measure of uncertainty associated with a set of random variables.

Definition 2 (Joint Entropy): For a pair of discrete random variables (X, Y), the joint entropy H(X, Y) is defined as:

$$H(X, Y) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log_2 p(x, y) \tag{5}$$

where p(x, y) is the joint probability of (x, y). H(X, Y) = H(X) + H(Y) if and only if two random variables X and Y are independent of each other. Therefore, given n independent random variables $X_1, X_2, \ldots X_n$, then,

$$H(X_1, X_2, \ldots, X_n) = H(x_1) + H(x_2) + \ldots + H(x_n) \tag{6}$$

C. EUCLIDEAN METRIC

The Euclidean metric is a commonly used definition of distance, referring to the natural length of two vectors in an M-dimensional space.

Definition 3 (Euclidean Metric) The true distance between two points in M-dimensional space, the distance ED based on the euclidean metric is defined as:

$$\begin{aligned} ED(X, Y) &= \frac{1}{2}\|X - Y\|^2 \\ &= \frac{1}{2}((x_1 - y_1)^2 + \ldots + (x_M - y_M)^2) \end{aligned} \tag{7}$$

when X and Y are N × N-dimensional matrices, then,

$$\begin{aligned} ED(X, Y) &= \frac{1}{2}\|X - Y\|^2 \\ &= \frac{1}{2}\sum_{u \in N} \sum_{v \in N} (x_{uv} - y_{uv})^2 \end{aligned} \tag{8}$$

IV. DOUBLE WEIGHTED ENSEMBLE CLUSTERING

This paper proposes a double weighted ensemble clustering method based on local weighting [28]. In this section, we describe each step of the method in detail.

A. LOCAL WEIGHTING METHOD

Due to the unsupervised nature of clustering algorithms, it is difficult to know in advance which similarity measure is correct and reasonable. Different clustering algorithms have their own scope of application. Due to the difference in similarity, the clustering results are also different. Therefore, how to measure the similarity between clusters is the key to obtain reasonable clustering results. In order to evaluate the reliability of each cluster, the cluster uncertainty estimation method based on entropy criterion estimates the uncertainty of the cluster by considering the cluster labels in the whole set, and then proposes the concept of ECI to evaluate the clustering uncertainty and reliability.

As introduced in Information entropy, entropy is a measure of uncertainty associated with random variables. Each cluster is a set of data objects. Given two clusters $C_i, C_j \in C$ and C_i, C_j do not belong to the same base cluster, when there are more overlapping data objects in C_i, C_j the value of $H(C_i, C_j)$ is smaller. By analyzing the clustering of C_i and C, the entropy of C_i for the base cluster set Π can be calculated.

Definition 4: Given a set Π, the entropy of C_i for the base cluster set Π is defined as follows:

$$H^{\Pi}(c_i) = -\sum_{m=1}^{M}\sum_{j=1}^{n^m} p(c_i, c_j^m)\log_2 p(c_i, c_j^m) \quad (9)$$

where

$$p(c_i, c_j^m) = \frac{\left|c_i \cap c_j^m\right|}{|c_i|} \quad (10)$$

where M represents the number of base clusters, n^m represents the number of clusters in the partition π^m, and c_j^m represents the jth cluster in the mth partition. $\cap$ denotes the coincident elements in the two clusters, and $|c_i|$ denotes the number of elements in the cluster C_i.

In summary, $p(c_i, c_j^m) \in [0, 1]$ for any i, j and m, so we have $H^{\Pi}(c_i) \in [0, +\infty)$. The entropy of C_i for the base cluster set Π can reflect how objects in C_i are clustered in other base clusters in Π. If the objects in C_i belong to the same cluster in each base cluster, it can be seen that all base clusters agree to assign the objects in C_i to the same cluster, then the entropy of C_i about Π reaches the minimum value, namely 0. When the entropy of C_i with respect to Π is larger, the objects in C_i are less likely to be in the same cluster.

After obtaining the entropy of each cluster in the cluster set, we consider the uncertainty of the cluster relative to the set through the concept of ECI, and add weights to the data objects within each cluster.

Definition 5: Given a cluster set Π of M base clusters, the ECI of cluster C_i is defined as follows:

$$ECI(c_i) = e^{-\frac{H^{\Pi}(c_i)}{M}} \quad (11)$$

According to Definition 5, since $H^{\Pi}(C_i) \in [0, +\infty)$, for any $C_i \in C$, $ECI(C_i) \in (0, 1]$. Obviously, the entropy of C_i for the base cluster set Π is more The smaller the value, the larger the ECI value. When the entropy of the cluster reaches the minimum value, that is, $H^{\Pi}(C_i) = 0$, its ECI will reach the maximum value, that is, $\mathrm{ECI}(C_i) = 1$. When the entropy of the cluster tends to infinity, the ECI of this cluster tends to 0.

According to Definition 1, the connectivity matrix reflects whether two data objects in the partition are grouped into the same cluster. Combining the concept of ECI, we propose the E-similarity matrix(ESM) theory, which reflects the possibility of two data objects being grouped into the same cluster in the partition.

Definition 6: (E-Similarity Matrix): The E-similarity matrix is essentially a symmetric matrix. Given a partition $\pi^m = \{c_1^m, c_2^m, \ldots, c_{n^m}^m\}$, calculate its ESM The way is as follows:

$$w_i^m = ECI(c_i^m) \quad (12)$$

$$ESM_{uv}^m = \begin{cases} w_i^m, & if\ Cls^m(x_u) = Cls^m(x_v) = c_i^m \\ i, & if\ u = v \\ 0, & otherwise \end{cases} \quad (13)$$

where c_i^m represents the ith cluster in the mth partition, w_i^m represents the ECI value of the cluster c_i^m, and $ESM_u v^m$ represents that the data objects x_u and x_v are clustered to the same in the mth base clustering possibility of clusters.

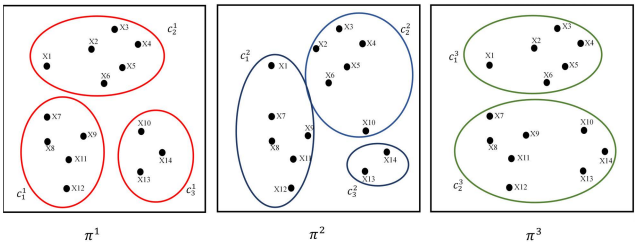

FIGURE 2. Partitions of π^1, π^2, π^3 base clustering.

We provide an example in Figure 2 and Table 1 to show the computation of local weights with respect to three basis clusters. The data set $X = \{x_1, x_2, \ldots, x_{14}\}$ has a total of 14 data objects, which are divided into three partitions π^1, π^2, π^3 after three clustering. The base cluster set Π has a total of 8 clusters. According to formula (9) (11), the entropy and ECI of the 8 clusters can be calculated. The results are shown in Table 1. As can be seen from Table 1, the entropy values of clusters c_1^1 and c_3^2 are the smallest, which means that the certainty of this cluster is the largest. The value of $H^{\Pi}(c_2^3)$ is the largest, indicating that the certainty of the c_2^3 cluster is the smallest, that is, the aggregate set has the smallest support for the cluster to appear in the final clustering result. According to Definition 6, formula (14) is calculated from the ECI in Table 1, which represents the ESM that partitions π^1, that is, the possibility that 14 data objects are clustered into the same cluster in the first base clustering. When any two data objects in the data set $X = \{x_1, x_2, \ldots, x_{14}\}$ are grouped into the same cluster by the base clustering, then the items corresponding to the two data objects in the ESM are assigned a value, which is the ECI of this cluster. Figure 3 shows the ESM corresponding to partitions π^1.

x_1	1	0.81	0.81	0.81	0.81	0.81	0	0	0	0	0	0	0	0
x_2	0.81	1	0.81	0.81	0.81	0.81	0	0	0	0	0	0	0	0
x_3	0.81	0.81	1	0.81	0.81	0.81	0	0	0	0	0	0	0	0
x_4	0.81	0.81	0.81	1	0.81	0.81	0	0	0	0	0	0	0	0
x_5	0.81	0.81	0.81	0.81	1	0.81	0	0	0	0	0	0	0	0
x_6	0.81	0.81	0.81	0.81	0.81	1	0	0	0	0	0	0	0	0
x_7	0	0	0	0	0	0	1	1	1	0	1	1	0	0
x_8	0	0	0	0	0	0	1	1	1	0	1	1	0	0
x_9	0	0	0	0	0	0	1	1	1	0	1	1	0	0
x_{10}	0	0	0	0	0	0	0	0	0	1	0	0	0.74	0.74
x_{11}	0	0	0	0	0	0	1	1	1	0	1	1	0	0
x_{12}	0	0	0	0	0	0	1	1	1	0	1	1	0	0
x_{13}	0	0	0	0	0	0	0	0	0	0.74	0	0	1	0.74
x_{14}	0	0	0	0	0	0	0	0	0	0.74	0	0	0.74	1

FIGURE 3. ESM of partitions π^1 according to Definition 6.

B. EUCLIDEAN METRIC BASED ESM WEIGHTED ENSEMBLE FUNCTION

The ensemble process refers to the process of finding the optimal one among the partitions generated by each base cluster.

TABLE 1. Entropy and ECI calculations for clusters in the example in Figure 2.

Base Clustering	Cluster	Entropy	ECI
π^1	c_1^1	$H^{\Pi}(c_1^1) = 0.00$	$ECI(c_1^1) = 1.00$
	c_2^1	$H^{\Pi}(c_2^1) = 0.65$	$ECI(c_2^1) = 0.81$
	c_3^1	$H^{\Pi}(c_3^1) = 0.92$	$ECI(c_3^1) = 0.74$
π^2	c_1^2	$H^{\Pi}(c_1^2) = 1.30$	$ECI(c_1^2) = 0.65$
	c_2^2	$H^{\Pi}(c_2^2) = 1.30$	$ECI(c_2^2) = 0.65$
	c_3^2	$H^{\Pi}(c_3^2) = 0.00$	$ECI(c_3^2) = 1.00$
π^3	c_1^3	$H^{\Pi}(c_1^3) = 0.65$	$ECI(c_1^3) = 0.81$
	c_2^3	$H^{\Pi}(c_2^3) = 2.25$	$ECI(c_2^3) = 0.47$

In this study, we calculated the corresponding ESM^i from the partitions π^m generated by each base clustering. Therefore, there also exists a similarity matrix ESM, corresponding to the final optimal consensus partition π. In this sense, finding an optimal ESM is the key to obtain good clustering results. In general, different base clusters also have different importance to the final consensus. Therefore, we treat this objective as an optimization problem. According to Definition 3, this optimization problem is described as follows:

$$\min_{ESM,w} w_i ED(ESM, ESM^i)$$
$$s.t.\ ESM = ESM^T, \quad ESM > 0,\ w_i \geq 0,\ \sum_{i=1}^{M} w_i = 1 \tag{14}$$

where ESM is a non-negative symmetric matrix. $ED(ESM, ESM^i) = ED(ESM, \{ESM^1, ESM^2, \ldots, ESM^M\})$ represents the sum of the Euclidean metric of ESM^i corresponding to ESM and each base cluster. w_i represents the contribution of each base cluster in the consensus process.

We use a block coordinate descent algorithm to minimize the above problem. When we fix one variable, optimization over another variable can be viewed as a convex problem with a unique solution. In order to avoid solving the result w_i only take 1 and 0, we add a regularization term to formula (15).

Definition 7: Calculate the optimal ESM, defined as follows:

$$\min_{ESM,w} w_i ED(ESM, ESM^i) + \lambda \|W\|^2$$
$$s.t.\ ESM = ESM^T, \quad ESM > 0,\ w_i \geq 0,\ \sum_{i=1}^{M} w_i = 1 \tag{15}$$

where λ is the regularization coefficient. When λ approaches 0, w_i only takes 1 and 0. When λ approaches 1, the value of w_i is $1/M$, which is the average of all partitions.

According to Equation (8) and Equation (15), we describe the problem as:

$$J(ESM, w) = \frac{1}{2} \sum_{i=1}^{M} w_i \sum_{u \in N} \sum_{v \in N} (ESM_{uv} - ESM_{uv}^i)^2 + \lambda \|w\|^2 \tag{16}$$

By fixing w such that $\frac{\partial J(ESM,w)}{\partial ESM} = 0$, where 0 is an N × N-dimensional matrix. We can get:

$$ESM_{uv} = \frac{1}{M} \sum_{i=1}^{M} w_i ESM_{uv}^i \tag{17}$$

Similarly, by fixing the ESM so that $\frac{\partial J(ESM,w)}{\partial w} = 0$, the problem is transformed into a linear programming problem. Then, we can prove that formula (16) converges, $J(ESM, w) \geq 0$ for any ESM and w. By fixing $w = w^t$, the minimization of $J(ESM, w)$ is convex, ESM^{t+1} is the optimal solution, and $J(ESM^t, w^t) \geq J(ESM^{t+1}, w^t)$. Similarly, by fixing $ESM = ESM^{t+1}$, we have $J(ESM^{t+1}, w^t) \geq J(ESM^{t+1}, w^{t+1})$. Therefore, we get a monotonically decreasing sequence $J(ESM^0, w^0) \geq J(ESM^1, w^0) \geq J(ESM^1, w^1) \geq \ldots \geq 0$. indicating that formula (17) converges. After finding the optimal solution ESM through optimization, we use the K-means algorithm to cluster the ESM to get the final data object labels. Among them, the input of the K-means algorithm is a vector composed of the similarity between each data object and all data objects, that is, each column of the ESM.

The Double weighted ensemble clustering algorithm specifically described as follows:

Input:

- $\Pi = \{\pi^1, \pi^2, \ldots, \pi^M\}$// M base clustering results
- $C = \{c_1, c_2, \ldots, c_{n_c}\}$// n_c clusters of Π
- k//number of clusters
- λ//regularization coefficient
- ϵ// precision

Output: labels//the final ensemble clustering result

Step1: Compute the entropy of the clusters in C as Definition 4.

Step2: Compute the ECI measures of the clusters in C as Definition 5.

Step3: Compute the ESM set of the partitions in Π as Definition 6.

Step4: Optimization ESM in ESM set as Definition 7.

- Initialize $w^t = [\frac{1}{M}, \frac{1}{M}, \ldots, \frac{1}{M}], t = 0, \Delta = +\infty$
- While $\Delta > \epsilon$ do:
- $t = t + 1$
- find the minimum of $J(ESM^t, w^t)$ by fixing w^t
- find the minimum of $J(ESM^t, w^t)$ by fixing ESM^t
- Compute $\Delta = \left|J(ESM^t, w^t) - J(ESM^{t-1}, w^{t-1})\right|$
- End while
- Output ESM

Step5: The final result is obtained by clustering ESM through k-means algorithm.

V. EXPERIMENTS

In this section, we conduct experiments on 10 real datasets in the UCI database and five cancer datasets from TCGA to illustrate the advantages of the doubly weighted ensemble clustering method compared to the state-of-the-art ensemble clustering method. This paper selects new and some classic clustering algorithms in recent years and compares them with the double weighted ensemble clustering

algorithm (DWEC). Local Weighted Evidence Accumulation Algorithm (LWEA) and Local Weighted Graph Partition Algorithm (LWGP) [28], Double granularity weighted ensemble clustering (DGWEC) [41], Kullback-Leibler Distance Weighted Bregman Consensus Clustering (KLWBCC) and Exponential Distance weighted Bregman consensus clustering (eWBCC) [5] were selected as the newly proposed ensemble clustering comparison algorithms in recent years; Evidence Accumulation Clustering (EAC) [30], Cluster-Based Similarity Partitioning Algorithm(CSPA)and Hyper-Graph-Partitioning Algorithm(HGPA) [39]were selected as the classic ensemble clustering comparison algorithm; In addition, k-means [8] and Spectral Clustering(SC) clustering algorithm [7] were used as basic comparison algorithms. All parameters involved in the algorithm are set according to the parameters established in the corresponding literature experiments, and the performance of the used comparison algorithm is for reference only. For all ensemble clustering algorithms, we use 50 partitions generated by K-means clustering algorithm as base clusters, and all evaluation metrics are averaged 20 times.

A. EXPERIMENTAL COMPARISON OF UCI DATASETS

In our experiments, 10 datasets from the UCI database were used, namely, Balance, Breast, Glass, Heart, Ionosphere, Iris, Sonar, Vehicle, Wine and Zoo. All datasets are available through the UCI official website http://archive.ics.uci.edu/ml/index.php. The details of the dataset are shown in Table 2.

Mutual information (MI) is a symmetric measure that quantifies the statistical information shared between two distributions, which can be seen as the amount of information contained in one random variable about another random variable, or the amount of information a random variable has Reduced uncertainty by knowing another random variable. As the name suggests, normalized mutual information (NMI) is to put mutual information between [0, 1], which is widely used to evaluate the quality of clustering. Suppose there are two random variables (X, Y), MI and NMI are defined as follows:

$$MI(X, Y) = \sum_x \sum_y p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \tag{18}$$

where the joint distribution of random variables (X, Y) is $p(x, y)$, and the marginal distributions are $p(x), p(y)$ respectively. In essence, the mutual information $MI(X, Y)$ is the joint distribution $p(x, y)$ and the relative entropy between the marginal distribution product $p(x)p(y)$.

$$NMI(X, Y) = \frac{2MI(X, Y)}{H(X) + H(Y)} \tag{19}$$

where $H(X)$ is the definition of entropy in Equation 5.

The application of normalized mutual information in this paper, the reference [28] is defined as follows:

$$NMI(\pi', \pi^G) = \frac{\sum_{i=1}^{n'} \sum_{j=1}^{n^G} \log \frac{n_{ij} n}{n_{i'} n_j^G}}{\sqrt{\sum_{i=1}^{n'} n_{i'} \log \frac{n_{j'}}{n} \sum_{i=1}^{n^G} n_j^G \log \frac{n_j^G}{n}}} \tag{20}$$

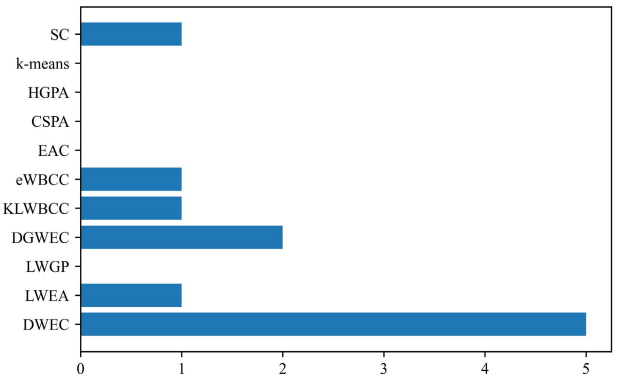

FIGURE 4. Number of times ranked first.

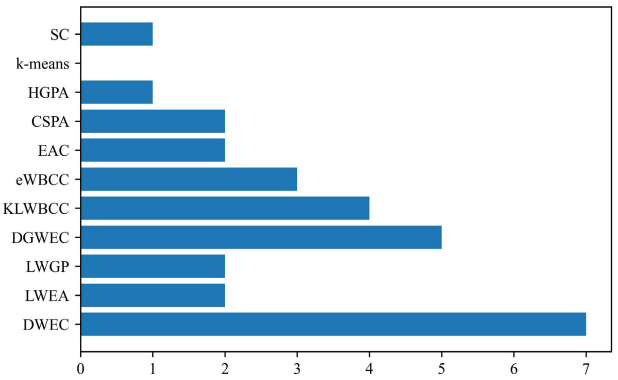

FIGURE 5. Number of times ranked in the top three.

where π' is the clustering result of the experiment, π^G is the clustering result of the truth, n is the number of data objects, n' is the number of clusters in π', and n^G is the cluster in π^G, n_i' is the number of data objects in the ith cluster in π', n_j^G is the number of data objects in the jth cluster in π^G, n_{ij} is the ith cluster in π' and the number of data objects in common in the jth cluster in π^G.

Table 3 reports the NMI scores of different ensemble clustering methods based on the 50th K-means algorithm as base clustering. Deepening font in each line is the maximum. As shown in the table, the DWEC method obtains the best value among the NMI scores of each ensemble clustering method for the five datasets of Balance, Glass, Heart, Vehicle and Zoo. In addition, Figure 4 and Figure 5 respectively show the number of times that each method obtains the first and the top three in the NMI scores of the 10 datasets. Our proposed DWEC method achieves 5 firsts and 7 firsts, which is the best among many existing methods. It shows that the DWEC method has higher accuracy and robustness.

B. EXPERIMENTAL COMPARISON OF THE TCGA CANCER DATASET

To make our research more meaningful, we now take the real cancer gene dataset as the research object. We selected

TABLE 2. Experimental data set description from UCI.

Dataset number	Dataset name	No. of samples	No. of features	No. of classes
1	Balance	625	4	3
2	Breast	106	9	2
3	Glass	214	9	6
4	Heart	270	13	2
5	Ionosphere	351	34	2
6	Iris	150	4	3
7	Sonar	208	60	2
8	Vehicle	846	18	3
9	Wine	178	13	3
10	Zoo	101	16	7

TABLE 3. Comparison of NMI values between DWEC algorithm and other algorithms on 10 real UCI datasets.

Dataset	Method										
	DWEC	LWEA	LWGP	DGWEC	KLWBCC	eWBCC	EAC	CSPA	HGPA	k-means	SC
Balance	**0.1432**	0.1234	0.1350	0.1230	0.1302	0.1125	0.1386	0.1179	0.1296	0.1125	0.1136
Breast	0.0834	0.0420	0.0477	0.0345	0.0645	**0.0853**	0.0420	0.0433	0.0416	0.0410	0.0635
Glass	**0.4276**	0.3061	0.3005	0.3193	0.3785	0.2988	0.3000	0.3017	0.3022	0.2862	0.2534
Heart	**0.1440**	0.0607	0.0583	0.0525	0.0795	0.0986	0.0432	0.1195	0.1233	0.0420	0.0211
Ionosphere	0.1341	0.1412	0.1566	**0.1712**	0.1496	0.1325	0.1453	0.1526	0.1497	0.1312	0.0952
Iris	0.7565	0.7156	0.7115	0.7364	0.7264	0.6891	0.7325	0.7178	0.7256	0.7169	**0.7966**
Sonar	0.0052	0.0078	0.0075	**0.0080**	**0.0080**	0.0065	0.0070	0.0053	0.0066	0.0068	0.0032
Vehicle	**0.1055**	0.0712	0.0756	0.0760	0.1022	0.1035	0.0653	0.0633	0.0791	0.0149	0.0652
Wine	0.4258	**0.6405**	0.5990	0.6087	0.5791	0.6210	0.4993	0.3274	0.5674	0.4287	0.4567
Zoo	**0.7499**	0.5624	0.6012	0.6061	0.5497	0.5914	0.6066	0.5371	0.4294	0.5327	0.6784

five cancers with high global incidence: Glioblastoma multiforme (GBM), Breast invasive carcinoma (BIC), Kidney renal clear cell carcinoma (KRCCC), Lung squamous cell carcinoma (LSCC) and Colon adenocarcinoma (COAD) from The Cancer Genome Atlas (TCGA). TCGA is a project overseen by the National Cancer Institute and the National Human Genome Research Institute to apply high-throughput genome analysis techniques to help people have a better understanding of cancer. Complete understanding, thereby improving the ability to prevent, diagnose and treat cancer. TCGA data requires complex preprocessing, and fortunately, the dataset provided by the article [33] can meet our research conditions. Among them, each type of cancer provides three types of gene expression (mRNA, miRNAs and methylation): mRNA is short for messenger RNA, and RNA expression measured by RNA sequencing is transcribed from DNA; MicroRNAs (miRNAs) are small endogenous non-coding RNA molecules, the amount of RNA expression detected by microRNA sequencing; methylation refers to the degree of DNA methylation, as measured by methylation chips. The data preprocessing process is described in detail in the paper [33]. All datasets are available at http://compbio.cs.toronto.edu/SNF/SNF/Software.html.

Glioblastoma multiforme dataset: This dataset contains 215 samples with dimensions of 12042, 534 and 1305 for mRNA, miRNAs and methylation, respectively.

Breast invasive carcinoma dataset: This dataset contains 105 samples with dimensions of 17814, 354 and 23094 for mRNA, miRNAs and methylation, respectively.

Kidney renal clear cell carcinoma dataset: This dataset contains 122 samples with dimensions of 17899, 329 and 24960 for mRNA, miRNAs and methylation, respectively.

Lung squamous cell carcinoma dataset: This dataset contains 106 samples with dimensions of 12042, 532 and 23074 for mRNA, miRNAs and methylation, respectively.

Colon adenocarcinoma dataset: This dataset contains 92 samples with dimensions of 17814, 312 and 23088 for mRNA, miRNAs and methylation, respectively.

Different conclusions can be drawn due to the use of different types of gene expression. For example, using miRNAs and methylation it is possible to profile different cancer subtypes. Therefore, this study combined the gene expression of three different types of five cancers, respectively, to obtain five datasets. The details of the dataset are shown in Table 4.

TABLE 4. Experimental data set description from TCGA.

Dataset number	Dataset name	No. of samples	No. of features
1	GBM	215	13881
2	BIC	105	41262
3	KRCCC	122	43188
4	LSCC	106	35468
5	COAD	92	41214

The difference between the TCGA dataset and the UCI dataset is that the former has no labels, so we cannot use the normalized mutual information in Section 5.1 to evaluate the quality of the clusters. Here we use $-log_2 p$ in logrank-test

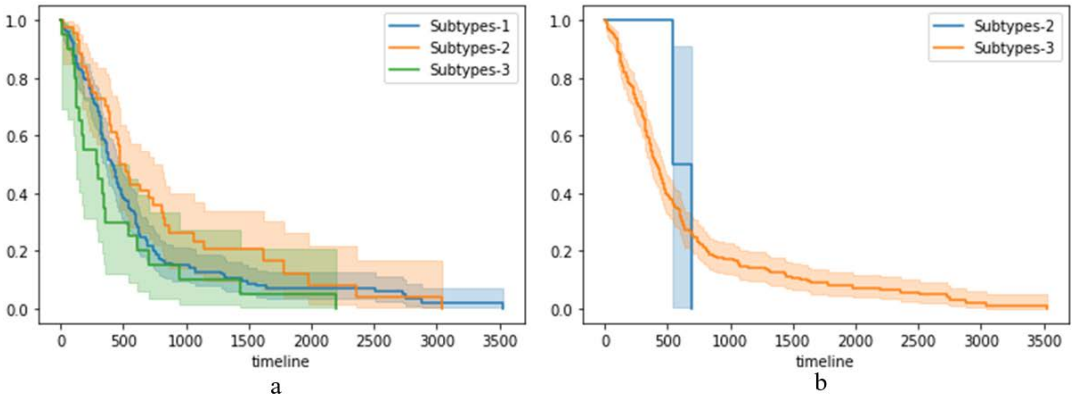

FIGURE 6. Survival curves of DWEC method and Spectral Clustering (SC) clustering algorithm for GBM dataset clustering results. (a) DWEC method (b) Spectral Clustering clustering algorithm.

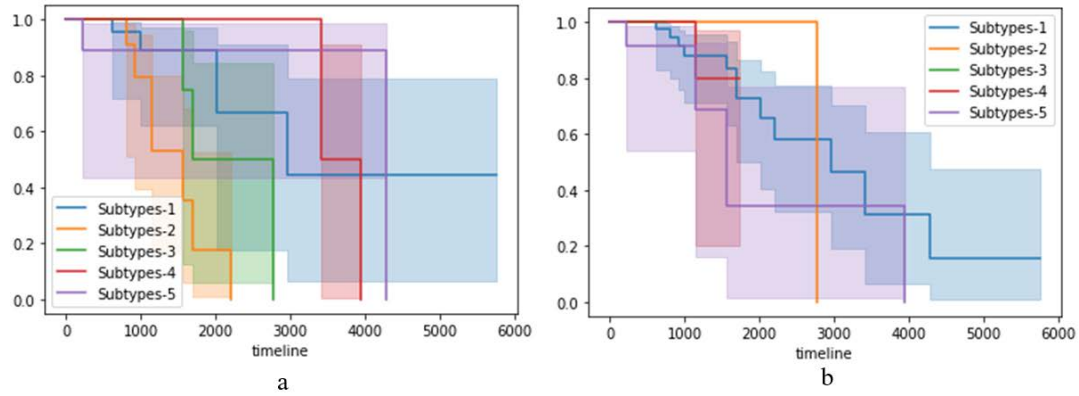

FIGURE 7. Survival curves of DWEC method and Spectral Clustering clustering algorithm for BIC dataset clustering results. (a) DWEC method (b) Spectral Clustering clustering algorithm.

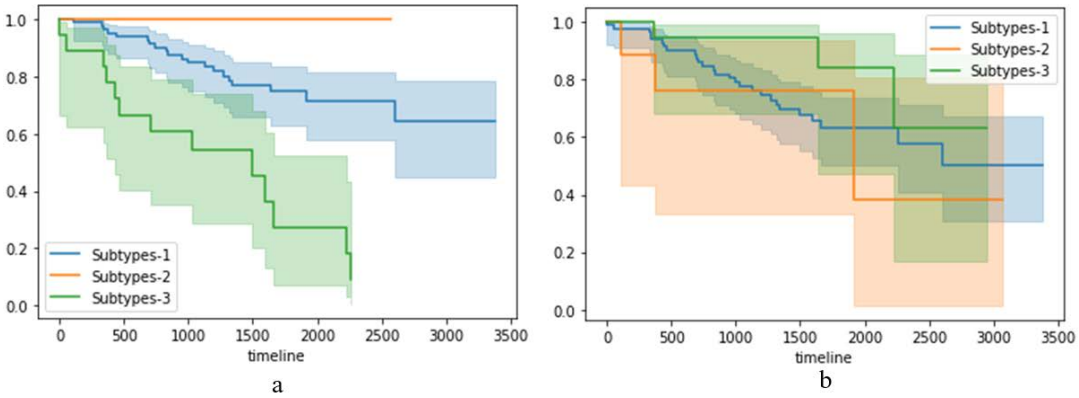

FIGURE 8. Survival curves of DWEC method and Spectral Clustering clustering algorithm for KRCCC dataset clustering results. (a) DWEC method (b) Spectral Clustering clustering algorithm.

as the evaluation criterion for clustering results of different methods to evaluate the significance of differences in survival information among different subtypes [42]. The application of log-rank test in this study is to compare the survival curves between multiple groups to study whether the difference in survival distribution among multiple groups is statistically significant. Suppose there are i survival distributions. H_0 is defined as the null hypothesis that all survival distributions are the same. In this study, we take $p \geq 0.05$ as H_0, and the difference of i survival distributions is not statistically

TABLE 5. Comparison of $-log_2p$ values between DWEC algorithm and other algorithms in log-rank test on 5 TCGA cancer datasets.

Dataset	Method										
	DWEC	LWEA	LWGP	DGWEC	KLWBCC	eWBCC	EAC	CSPA	HGPA	k-means	SC
GBM	4.15	4.20	**4.32**	4.31	4.27	1.66	2.97	4.26	1.69	2.14	0.22
BIC	**9.53**	7.11	8.64	9.39	3.46	4.23	4.33	2.98	5.76	3.51	0.76
KRCCC	**8.02**	5.79	5.41	4.66	6.81	7.11	5.24	3.19	1.55	4.85	1.91
LSCC	**10.57**	8.73	6.31	7.25	5.48	8.12	5.98	4.65	3.32	5.73	1.35
COAD	**5.94**	5.45	4.82	5.16	5.53	5.16	5.20	4.84	4.79	4.68	0.64

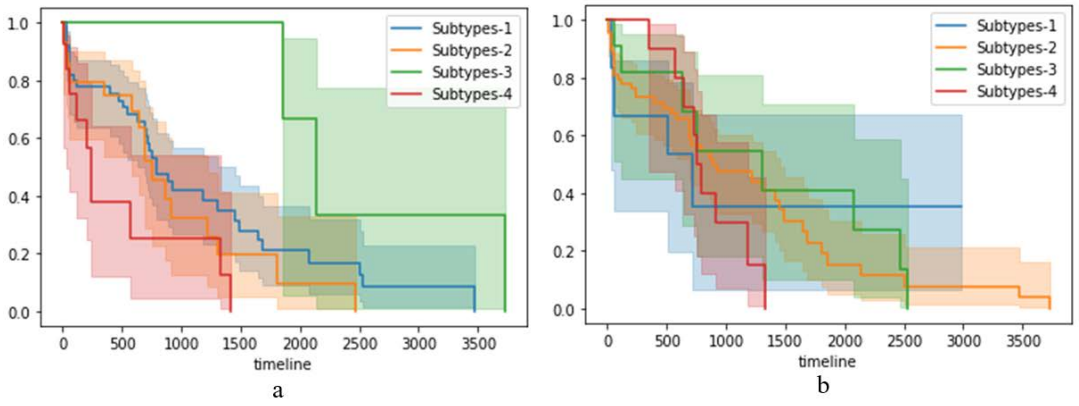

FIGURE 9. Survival curves of DWEC method and Spectral Clustering clustering algorithm for LSCC dataset clustering results. (a) DWEC method (b) Spectral Clustering clustering algorithm.

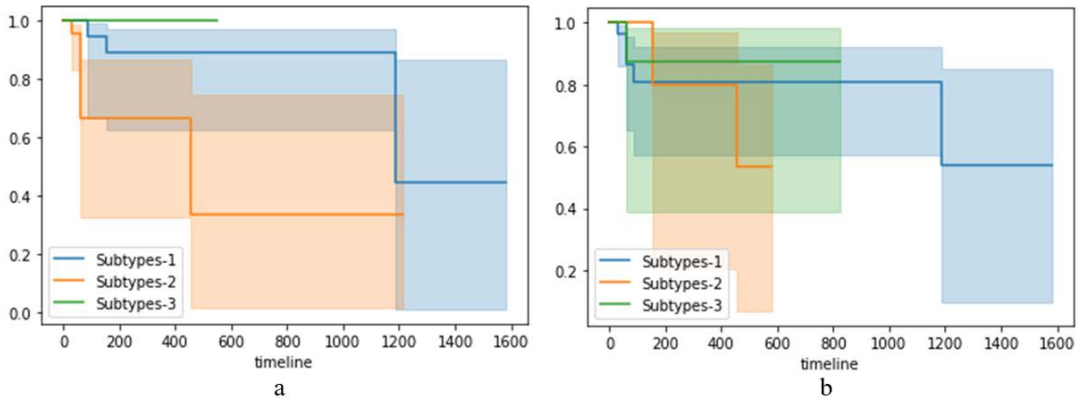

FIGURE 10. Survival curves of DWEC method and Spectral Clustering clustering algorithm for COAD dataset clustering results. (a) DWEC method (b) Spectral Clustering clustering algorithm.

significant under the condition that H_0 is established, that is, when $-log_2p \geq 4.32$, H_0 is rejected. $-log_2p$ the larger the value, the better the clustering effect.

In the study [33], by reviewing numerous literatures, the experiment divided glioblastoma multiforme into 3 subtypes, breast invasive carcinoma into 5 subtypes, kidney renal clear cell carcinoma is divided into 3 subtypes, lung squamous cell carcinoma is divided into 4 subtypes, and colon adenocarcinoma is divided into 3 subtypes. This study will use this conclusion as the number of clusters for different datasets.

Table 5 reports the values of $-log_2p$ in the log-rank test of DWEC and ten contrasting algorithms on the TCGA cancer dataset. Deepening font in each line is the maximum. As shown in Table 5, for the time series detection of GBM, all methods except the LWGP method satisfy the H_0 hypothesis, indicating that there is no significant difference in the survival distribution among the groups of the clustering results. For the BIC, KRCCC, LSSS and COAD datasets, the DWEC method proposed by us is superior to other methods, indicating that the clustering results of the DWEC method have greater differences in survival distribution among the groups, and the results are more reliable. On the whole, the DWEC method has the best clustering effect, and the SC clustering algorithm has the worst effect. Figure 6 - Figure 10 show the survival curves of each group by the DWEC method and the SC algorithm for GBM, BIC, KRCCC, LSSS and COAD datasets, respectively. Survival curves are used to describe the survival status of several groups of patients. The horizontal

axis of the survival curve is the observation time, and the vertical axis is generally the survival rate. Each point on the curve represents the patient's survival rate at that time point. In general, the greater the distance between the curves, the greater the difference in the prognosis of patients in each group, and the easier it is to make statistical differences. As shown in the figure, the DWEC method has significant advantages over the SC clustering algorithm.

VI. CONCLUSION

This paper proposes a double weighted ensemble clustering algorithm. First, according to the local weighting method, the similarity matrix of each base cluster is obtained. The problem is then transformed into a convex optimization problem, and the optimal similarity matrix is obtained by a block coordinate descent algorithm. Finally, the K-Means algorithm is used on the basis of the similarity matrix to obtain the final partitions result.

Due to the high latitude characteristics of cancer gene data, higher requirements are placed on clustering accuracy. The DWEC method is weighted twice and aims to improve the clustering accuracy. We conduct extensive experiments to demonstrate the large superiority of DWEC algorithm in terms of clustering quality. First, a large number of experiments are conducted on the labeled UCI dataset. By comparing the NMI values of various algorithms, the results show that our method performs best compared with existing ensemble clustering methods. On the clinical side, these algorithms are used for clustering of cancer genes to analyze cancer subtypes. Through experiments on five common cancers in TCGA, and comparing the time series detection results of each clustering algorithm, the effectiveness of our method is verified.

In the next step, we plan to apply our consensus clustering framework to large-scale data from the World Health Organization, such as analyzing potential links between public health expenditures and life expectancy in countries, potential factors affecting suicide rates in countries, etc.

REFERENCES

[1] K. Tomczak, P. Czerwinska, and M. Wiznerowicz, "The cancer genome atlas (TCGA): An immeasurable source of knowledge," *Contemp. Oncol.*, vol. 19, no. 1A, p. A68, 2015.

[2] C. E. Meacham and S. J. Morrison, "Tumour heterogeneity and cancer cell plasticity," *Nature*, vol. 501, no. 7467, p. 327, 2013.

[3] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA, Cancer J. Clin.*, vol. 68, no. 6, pp. 394–424, 2018.

[4] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, 2010.

[5] J. Li, L. Xie, Y. Xie, and F. Wang, "Bregmannian consensus clustering for cancer subtypes analysis," *Comput. Methods Programs Biomed.*, vol. 189, Jun. 2020, Art. no. 105337.

[6] R. Nock and F. Nielsen, "On weighting clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 8, pp. 1223–1235, Aug. 2006.

[7] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, Feb. 2007.

[8] J. A. Hartigan and M. A. Wong, "A K-means clustering algorithm," *Appl. Stat.*, vol. 28, no. 1, pp. 100–108, Jan. 1979.

[9] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. KDD*, vol. 96. Palo Alto, CA, USA: AAAI Press, 1996, pp. 226–231.

[10] L. Wang, C. Leckie, R. Kotagiri, and J. Bezdek, "Approximate pairwise clustering for large data sets via sampling plus extension," *Pattern Recognit.*, vol. 44, no. 2, pp. 222–235, 2011.

[11] C.-D. Wang, J.-H. Lai, and J.-Y. Zhu, "Graph-based multiprototype competitive learning and its applications," *IEEE Trans. Syst., Man, Cybern., C, Appl. Rev.*, vol. 42, no. 6, pp. 934–946, Nov. 2012.

[12] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy *c*-means clustering algorithm," *Comput. Geosci.*, vol. 10, nos. 2–3, pp. 191–203, 1984.

[13] N. R. Pal, K. Pal, J. M. Keller, and J. C. Bezdek, "A possibilistic fuzzy c-means clustering algorithm," *IEEE Trans. Fuzzy Syst.*, vol. 13, no. 4, pp. 517–530, Aug. 2005.

[14] C.-D. Wang, J.-H. Lai, D. Huang, and W.-S. Zheng, "SVStream: A support vector-based algorithm for clustering data streams," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 6, pp. 1410–1424, Jun. 2013.

[15] H. Wang, T. Li, T. Li, and Y. Yang, "Constraint neighborhood projections for semi-supervised clustering," *IEEE Trans. Cybern.*, vol. 44, no. 5, pp. 636–643, May 2014.

[16] B.-K. Bao, W. Min, T. Li, and C. Xu, "Joint local and global consistency on interdocument and interword relationships for co-clustering," *IEEE Trans. Cybern.*, vol. 45, no. 1, pp. 15–28, Jan. 2015.

[17] Y. Yang, Z. Ma, Y. Yang, F. Nie, and H. T. Shen, "Multitask spectral clustering by exploring intertask correlation," *IEEE Trans. Cybern.*, vol. 45, no. 5, pp. 1083–1094, May 2015.

[18] A. Strehl and J. Ghosh, "Cluster ensembles—A knowledge reuse framework for combining partitionings," *J. Mach. Learn. Res.*, vol. 3, pp. 583–617, Dec. 2002.

[19] C. Gao, W. Pedrycz, and D. Miao, "Rough subspace-based clustering ensemble for categorical data," *Soft Comput.*, vol. 17, no. 9, pp. 1643–1658, Sep. 2013.

[20] P. Lingras, M. Chen, and D. Miao, "Qualitative and quantitative combinations of crisp and rough clustering schemes using dominance relations," *Int. J. Approx. Reasoning*, vol. 55, no. 1, pp. 238–258, Jan. 2014.

[21] X. Wang, C. Yang, and J. Zhou, "Clustering aggregation by probability accumulation," *Pattern Recognit.*, vol. 42, no. 5, pp. 668–675, May 2009.

[22] T. Wang, "CA-Tree: A hierarchical structure for efficient and scalable coassociation-based cluster ensembles," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 41, no. 3, pp. 686–698, Jun. 2011.

[23] L. Franek and X. Jiang, "Ensemble clustering by means of clustering embedding in vector spaces," *Pattern Recognit.*, vol. 47, no. 2, pp. 833–842, Feb. 2014.

[24] J. Wu, H. Liu, H. Xiong, J. Cao, and J. Chen, "K-means-based consensus clustering: A unified view," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 1, pp. 155–169, Jan. 2015.

[25] Z. Yu, L. Li, J. Liu, J. Zhang, and G. Han, "Adaptive noise immune cluster ensemble using affinity propagation," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 12, pp. 3176–3189, Dec. 2015.

[26] P. Rathore, J. C. Bezdek, S. M. Erfani, S. Rajasegarar, and M. Palaniswami, "Ensemble fuzzy clustering using cumulative aggregation on random projections," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 3, pp. 1510–1524, Jun. 2018.

[27] D. Huang, C.-D. Wang, J.-S. Wu, J.-H. Lai, and C.-K. Kwoh, "Ultra-scalable spectral clustering and ensemble clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 6, pp. 1212–1226, Jun. 2020.

[28] D. Huang, C.-D. Wang, and J.-H. Lai, "Locally weighted ensemble clustering," *IEEE Trans. Cybern.*, vol. 48, no. 5, pp. 1460–1473, May 2018.

[29] D. Huang, C.-D. Wang, H. Peng, J. Lai, and C.-K. Kwoh, "Enhanced ensemble clustering via fast propagation of cluster-wise similarities," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 51, no. 1, pp. 508–520, Jan. 2021.

[30] A. L. N. Fred and A. K. Jain, "Combining multiple clusterings using evidence accumulation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 6, pp. 835–850, Jun. 2005.

[31] D. Huang, J.-H. Lai, and C.-D. Wang, "Combining multiple clusterings via crowd agreement estimation and multi-granularity link analysis," *Neurocomputing*, vol. 170, pp. 240–250, Dec. 2015.

[32] S. Ronglai, A. B. Olshensup, and M. Ladanyisup, "Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis," *Bioinformatics*, vol. 25, no. 22, pp. 2906–2912, Nov. 2009.

[33] B. Wang, A. M. Mezlini, and F. Demir, "Similarity network fusion for aggregating data types on a genomic scale," *Nature Methods*, vol. 11, no. 3, pp. 333–337, Jan. 2014.
[34] Y. Xu, V. Olman, and D. Xu, "Clustering gene expression data using a graph-theoretic approach: An application of minimum spanning trees," *Bioinformatics*, vol. 18, no. 4, pp. 536–545, 2002.
[35] E. F. Lock and D. B. Dunson, "Bayesian consensus clustering," *Bioinformatics*, vol. 29, pp. 2610–2616, Oct. 2013.
[36] C. Tang, X. Liu, X. Zhu, J. Xiong, M. Li, J. Xia, X. Wang, and L. Wang, "Feature selective projection with low-rank embedding and dual Laplacian regularization," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 9, pp. 1747–1760, Sep. 2020.
[37] C. Tang, X. Zheng, X. Liu, W. Zhang, J. Zhang, J. Xiong, and L. Wang, "Cross-view locality preserved diversity and consensus learning for multi-view unsupervised feature selection," *IEEE Trans. Knowl. Data Eng.*, early access, Jan. 1, 2021, doi: 10.1109/TKDE.2020.3048678.
[38] A. Lourenço, S. R. Bulò, N. Rebagliati, A. L. N. Fred, M. A. T. Figueiredo, and M. Pelillo, "Probabilistic consensus clustering using evidence accumulation," *Mach. Learn.*, vol. 98, nos. 1–2, pp. 331–357, Jan. 2015.
[39] A. Strehl and J. Ghosh, "Cluster ensembles—A knowledge reuse framework for combining multiple partitions," *J. Mach. Learn. Res.*, vol. 3, no. 3, pp. 583–617, 2002.
[40] D. Huang, J. Lai, and C.-D. Wang, "Ensemble clustering using factor graph," *Pattern Recognit.*, vol. 50, pp. 131–142, Feb. 2016.
[41] L. Xu and S. Ding, "Dual-granularity weighted ensemble clustering," *Knowl.-Based Syst.*, vol. 225, Aug. 2021, Art. no. 107124.
[42] S. Lemeshow, S. May, and D. W. Hosmer, Jr., *Applied Survival Analysis: Regression Modeling of Time-to-Event Data*. Hoboken, NJ, USA: Wiley, 2011.

XIN ZHANG received the B.S. degree from the School of Software, Henan University of Science and Technology, where he is currently pursuing the master's degree with the College of Information Science and Engineering. His main research interests include machine learning algorithms, data mining technology research, and medical big data.

HUA HUO received the Ph.D. degree from Xi'an Jiaotong University, Xi'an, China, in 2005. He is currently a Professor at the Henan University of Science and Technology, Luoyang. Prior to that, he was a Postdoctoral Fellow at the State Key Laboratory of Intelligent Technology and Systems, Tsinghua University, from 2006 to 2008, and Ireland Dublin University Trinity College, from 2009 to 2010. He has published 70 academic papers. His research interests include pattern recognition and intelligent systems, artificial intelligence big data, data mining, visual information processing, and intelligent information processing. He is a Senior Member of the Chinese Artificial Intelligence Society, a member of the Chinese Computer Society, the Director of the Henan Computer Society, a Scientific and Technological Leader of the Henan Education Department, and an Excellent Expert at Luoyang city.