Journal of Cancer Research and Clinical Oncology (2024) 150:3
https://doi.org/10.1007/s00432-023-05559-4

RESEARCH

A semi-supervised ensemble clustering algorithm for discovering relationships between different diseases by extracting cell-to-cell biological communications

Xiuchao Shi[1] · Chunxiao Yue[2] · Meiping Quan[1] · Yalin Li[1] · Hiba Nashwan Sam[3]

Received: 14 July 2023 / Accepted: 1 November 2023 / Published online: 2 January 2024

Abstract

Introduction In recent decades, many theories have been proposed about the cause of hereditary diseases such as cancer. However, most studies state genetic and environmental factors as the most important parameters. It has been shown that gene expression data are valuable information about hereditary diseases and their analysis can identify the relationships between these diseases.

Objective Identification of damaged genes from various diseases can be done through the discovery of cell-to-cell biological communications. Also, extraction of intercellular communications can identify relationships between different diseases. For example, gene disorders that cause damage to the same cells in both breast and blood cancers. Hence, the purpose is to discover cell-to-cell biological communications in gene expression data.

Methodology The identification of cell-to-cell biological communications for various cancer diseases has been widely performed by clustering algorithms. However, this field remains open due to the abundance of unprocessed gene expression data. Accordingly, this paper focuses on the development of a semi-supervised ensemble clustering algorithm that can discover relationships between different diseases through the extraction of cell-to-cell biological communications. The proposed clustering framework includes a stratified feature sampling mechanism and a novel similarity metric to deal with high-dimensional data and improve the diversity of primary partitions.

Results The performance of the proposed clustering algorithm is verified with several datasets from the UCI machine learning repository and then applied to the FANTOM5 dataset to extract cell-to-cell biological communications. The used version of this dataset contains 108 cells and 86,427 promoters from 702 samples. The strength of communication between two similar cells from different diseases indicates the relationship of those diseases. Here, the strength of communication is determined by promoter, so we found the highest cell-to-cell biological communication between "basophils" and "ciliary. epithelial.cells" with 62,809 promoters.

Conclusion The maximum cell-to-cell biological similarity in each cluster can be used to detect the relationship between different diseases such as cancer.

Keywords Cell-to-cell communication · Ensemble clustering · Semi-supervised clustering · Gene expression · FANTOM5 dataset

✉ Xiuchao Shi
shixiuchao@yeah.net

Chunxiao Yue
414037459@qq.com

Meiping Quan
327311912@qq.com

Yalin Li
lyal1222@126.com

Hiba Nashwan Sam
nashwansamhiba23@gmail.com

[1] College of Environment and Life Sciences, Weinan Normal University, Weinan 714099, Shaanxi, China

[2] Weinan Junior Middle School, Weinan 714000, Shaanxi, China

[3] Department of Radiology and Sonar Techniques, Al-Noor University College, Nineveh, Iraq

Introduction

The human body consists of hundreds of types of cells (Kayal et al. 2019; Peng et al. 2022). These cells are directly or indirectly dependent on each other and have the ability to communicate and influence each other. Therefore, an effective mechanism is needed to find the relationship between this astronomical number of cells. Finding these communications will help identify relationships between different diseases. The nucleus of each cell has the coded instructions necessary to direct the cell's activities and make the necessary proteins. A whole group of these instructions is called a genome (Sivadas et al. 2022). Human genome is the genetic set and genes inside the nucleus of human cells (de Souza et al. 2016). There are millions of genes on each of the chromosomes, each of which has a specific role in the cell. Let the gene expression associated with a cell be represented by promoters (Shahraki et al. 2023).

To date, many diagnostic models have been presented for different diseases such as cancer. Each model uses different tools based on a specific dataset for prediction work (Zhang et al. 2022a). In recent years, datasets have been created that include a wide range of diseases. Datasets based on gene expression such as FANTOM5 include 1836 different samples from 201,803 regions of different genes that simultaneously cover several diseases (Rezaeipanah and Ahmadi 2022). Each sample contains the information of one patient from one cell or tissue. Here, sampling has been done in the form of gene expression, which shows how many times a gene has produced itself (de Souza et al. 2016).

In general, damaged cells from the body due to a disease can also be observed for other diseases (Li et al. 2023). If the promoters of a cell from two or more diseases are high enough, then it can be said that these diseases have a similar effect on this cell. Since there is information related to cells/tissues for each person, this dataset can be used to detect the communications between cells and tissues in the expression of different genes (Forouzandeh et al. 2023). In general, the analysis of gene expression information in order to identify intercellular communications requires mapping the problem to a clustering problem. Clustering algorithms can find relationships between different diseases by finding the most similar damaged cells.

Clustering algorithms are one of the most important techniques in data mining, machine learning and pattern recognition and are known as an effective method in data visualization and analysis (Rezaeipanah et al. 2021). These algorithms have wide applications in image processing, image segmentation, document analysis, market research, etc. Data clustering is data analysis without any prior information to assign each sample of the dataset to a group as a cluster (Zhang et al. 2022b; Zhao et al. 2023a). Each clustering algorithm seeks to create groups of data with maximum similarity between samples in the same clusters and minimum similarity between samples in different clusters. These algorithms are known as unsupervised learning methods, because the class labels are not available in the data analysis process (Cao et al. 2022; Tang et al. 2023).

In general, the types of clustering algorithms include hierarchical and partitional (Zhang et al. 2018). Hierarchical algorithms use a similarity metric for the clustering task. In each step of these algorithms, the data are divided into two categories to finally create a tree structure as a dendrogram (Wang et al. 2022). Dendrogram is a tree-structured graph that visualizes the result of a clustering algorithm at different levels of partitions (Forouzandeh et al. 2023). Meanwhile, partitional algorithms directly put data into multiple clusters based on distance or similarity. Hard and soft are common types of partitional clustering algorithms (Cheng et al. 2023). In hard clustering, a sample belongs to only one cluster; while in soft clustering, the degree of belonging of a sample to each cluster is determined by a number between 0 and 1.

In many real-world applications, the number of features in a dataset is too large for clustering. In most cases, there are a large number of unrelated features for clustering (Hou et al. 2020). Also, some features may be less important than others. Therefore, applying clustering with a subset of features can lead to an increase in the quality of the final partition. Meanwhile, not all clustering algorithms perform best for all data (Mojarad et al. 2021). Ensemble clustering is very popular to improve the performance of individual clustering algorithms. In an ensemble clustering algorithm, several individual clustering algorithms are combined to cover each other's weaknesses (Zhang et al. 2018). According to this, it is expected that the use of ensemble clustering algorithms will perform better than individual clustering algorithms.

Combining individual clustering algorithms with fixed weight is a common approach in ensemble technique. However, using fixed weights in the whole clustering process leads to a decrease in efficiency. In recent years, approaches based on adaptive weights during the clustering process have been developed to solve this shortcoming (Hou et al. 2020). In general, the use of traditional clustering algorithms does not perform well in dealing with high-dimensional data due to the correlation of features, noise, and dispersion.

On the other hand, applying the information of paired constraints can increase the effectiveness of individual clustering algorithms (Wang et al. 2020; Zhang et al. 2022c). This information includes must-link and cannot-link constraints. The must-link constraint indicates that a pair of samples belong to the same cluster, and the cannot-link constraint indicates that a pair of samples belongs to two

different clusters. Since effective clustering is challenging due to the lack of prior knowledge, using the constraints information as limited prior knowledge can improve the clustering performance. The use of constraint information in the clustering process has led to the emergence of clustering with semi-supervised learning (Mojarad et al. 2021; Bridges and Miller-Jensen 2022).

This paper proposes a semi-supervised ensemble clustering framework to discover relationships between diseases based on the extraction of cell-to-cell biological communications. The proposed semi-supervised framework uses prior knowledge in both parts of the ensemble, including the creation of primary partitions and the consensus function. Also, we present a stratified feature sampling mechanism to deal with high-dimensional data, which can reduce the risk of not selecting features to create primary partitions. In addition, the proposed clustering framework uses a new similarity metric developed based on the information of all primary partitions. Our method has medical applications for the treatment and prevention of cancer. In fact, we are looking to identify cells that may be destroyed in the same way in two different cancers.

The main contribution of this study is as follows:

- A clustering framework is proposed by joining "semi-supervised learning" and "ensemble technique", which is configured based on stratified feature sampling mechanism and a novel similarity metric
- Identification of cells with the highest promoters in order to discover relationships between different diseases on the FANTOM5 dataset
- Validation of the effectiveness of the proposed clustering framework on a wide range of UCI datasets

The remainder of this paper is organized as follows: The related work is summarized in "Related works". The fundamental concepts related to the problem are given in "Background". "Proposed clustering framework" explains the proposed clustering framework. The effectiveness of the proposed framework is discussed through numerical simulations in "Experiments". Finally, the paper ends with conclusions in "Conclusions".

Related works

Identification of intercellular communication from gene expression data with clustering algorithms is very common (Mojarad et al. 2021). Clustering is one of the data analysis techniques and so far, various solutions have been provided for it (Tan et al. 2022; Chang et al. 2022). For example, k-means, density-based spatial clustering of applications with noise (DBSCAN), multi-view spectral clustering, non-negative matrix factorization-based clustering, unsupervised deep embedding clustering, mean shift clustering, hierarchical clustering, etc. (Zhang et al. 2020; Lei et al. 2022). Compared to partitional clustering algorithms, many efforts have been reported for the improvements of hierarchical clustering algorithms in the last few decades.

Compared to classification, prior knowledge such as class labels is not available for clustering. Some studies use limited prior knowledge as constraint information in clustering (Hou et al. 2020). Zhang et al. (2018) used the pairwise constraints information to improve clustering performance and obtained some successes. Other semi-supervised clustering algorithms include constraints k-means, Constraint-based DBSCAN (C-DBSCAN), Pairwise Constrained k-means (PCKmeans), semi-supervised deep fuzzy c-mean clustering, semi-supervised denpeak clustering with pairwise constraints, semi-supervised deep embedded clustering, exhaustive and efficient constraint propagation, and semi-supervised maximum margin clustering (Mojarad et al. 2021).

Prades et al. (2020) proposed an agglomerative clustering approach to detect the number of endmembers in hyperspectral images. The authors follow this hypothesis in clustering that there is a cluster for each material different from image. Here, an approach based on principal component analysis applied to the centered image is used to reduce the dimensions. With reducing the dimensions of the image, the authors use a k-means algorithm to create primary clusters. Here, symmetric Kullback–Leibler (SKL) divergence is used as the distance calculation metric. SKL, also known as relative entropy, is a statistical metric from information theory to quantify the difference. This study uses principal component analysis to calculate the density of clusters. After that, a model-based agglomerative clustering approach is applied to provide a hierarchy of partitions. Eventually, the final partition of the hierarchy is determined by a validation algorithm. The number of clusters in the results of this model is considered as the number of materials.

Rezaeipanah and Ahmadi (2022) introduced multi-stage weights adjustment in the multi-layer perceptron (MWAMLP) for breast cancer detection. MWAMLP is an ensemble approach that uses three homogeneous multi-layer perceptron (MLP) neural networks for the classification task. The consensus function used in MWAMLP is developed based on the meta-classifier technique. The accuracy of this method on the WBCD dataset is 98.76% on average.

Mojarad et al. (2021) used an ensemble clustering algorithm to model inherited disease behavior (ECIDB). Here, cell-to-cell and tissue-to-tissue communications are extracted from the FANTOM5 dataset to identify cells with the highest disruption in each disease pair. The proposed algorithm uses the graph topological structure to represent the FANTOM5 dataset and uses an innovative

similarity metric to calculate the cell-to-cell similarity matrix. An ensemble clustering is then applied to identify primary intercellular or intertissue communications. Finally, a friend recommender-based system considering clustering information and topological similarities is used to identify related cells.

Sangeetha and Prakash (2021) proposed using deep learning to improve breast cancer disease prediction. A stacked sparse auto encoder network (SSAE) is constructed to learn features effectively. The network consists of a softmax classifier and several sparse autoencoders. In addition to adjusting the parameters of the algorithm, deep learning models are required. The parameters of the stacked sparse autoencoder can, therefore, be adjusted using particle swarm optimization (PSO). Regarding feature learning and classification, the PSO improves the performance of the SSAE.

Kayal et al. (2019) conducted a study to provide a new advanced classification method using a deep neural network (DNN) to predict the survival of patients with hepatic cancer. In the proposed method, the authors selected 15 risk factors out of 49 risk factors which are significantly responsible for hepatocellular carcinoma and then applied their method. According to the results, the proposed method is more accurate than other methods.

Sivadas et al. (2022) attempted to investigate the impact of racial information and natural factors on the incidence and progression of cancer by employing a multi-omics data fusion breast cancer survival cycle marker detection prediction model. The primary objective of this research is to enhance the prediction of breast cancer survival cycles through the development of a multi-omics fusion prediction model based on ensemble learning. This model incorporates clinical data, transcriptomics data, and methylomics data derived from The Cancer Genome Atlas (TCGA) datasets. The experimental results demonstrate that the three-omics fusion approach (with an accuracy rate of 97.43%) outperforms single-omics experiments and other race-based multi-omics and single-omics experiments in the context of the three-omics experiments, considering racial disparities. This research offers technical support for the classification of breast cancer patient survival cycle predictions and introduces novel concepts for the study of breast cancer survival prognostics.

Talatian Azad et al. (2021) proposed an intelligent ensemble classification method based on multi-layer perceptron (IEC-MLP) for breast cancer detection. IEC-MLP uses genetic algorithm for feature selection and parameter settings of MLP neural network. Here, MLP is developed based on an ensemble classification approach with three classifiers. This method detects breast cancer with high accuracy on the WBCD dataset, where the average accuracy is reported to be 98.74%.

Background

In this section, some basic concepts about the research method are explained. These concepts include system model, hierarchical clustering, semi-supervised clustering, ensemble clustering, and feature sampling.

System model

An individual clustering algorithm is denoted by π. Ensemble clustering consists of several individual clustering algorithms. We assume that $\Pi = \{\pi_1, \pi_2, \ldots, \pi_k, \ldots, \pi_Z\}$ is the set of Z individual clustering algorithms, where π_k represents the k-th clustering algorithm. Each $\pi_k \in \Pi$ can be applied to a dataset. We assume that $X = \{x_1, x_2, \ldots, x_i, \ldots, x_N\}$ is a dataset with N samples, where $x_i = \langle x_{i,1}, x_{i,2}, \ldots, x_{i,j}, \ldots, x_{i,M}\rangle$ represents the i-th sample with M features.

Applying each π_k to X results in a partition with multiple clusters. We assume that $p_k = \left[c_{k,1}, c_{k,2}, \ldots, c_{k,l}, \ldots, c_{k,|p_k|}\right]$ is the partition obtained by applying π_k on X with $|p_k|$ clusters. Here, $c_{k,l}$ represents the l-th cluster of the k-th partition. Considering ensemble clustering, applying set Π on X results in $P = \{p_1, p_2, \ldots, p_Z\}$. We assume that $p_* = \Gamma\langle p_1, p_2, \ldots, p_Z\rangle$ is the final partition obtained by consensus of set P. Here, Γ represents a consensus function such as majority vote. Let $p_* = \left[c_1, c_2, \ldots, c_K\right]$ be the details of the final partition, where K represents the total number of clusters.

Hierarchical clustering

Clustering is an unsupervised learning mechanism for grouping data, where samples belonging to each group have the highest similarity to each other and samples from different groups have the lowest similarity to each other. Partitional clustering and hierarchical clustering are two common types of clustering (Rostami et al. 2023). Partitional clustering clusters samples based on an objective function, where each sample belongs to only one cluster and the total number of clusters is known in advance. The k-means is one of the most common partitional clustering algorithms that performs clustering with the objective of minimizing the average distance to the center of each cluster (Torabi et al. 2022; Cao et al. 2023a). Meanwhile, hierarchical clustering can show a hierarchy of samples by dendrogram.

There are two general types of hierarchical clustering: (1) Divisive hierarchical clustering (DHC) or top-down approach where all samples belong to the same cluster at first. After that, each cluster is divided into smaller clusters so that finally each sample has its own cluster. (2) Agglomerative hierarchical clustering (AHC) or bottom-up approach

where each sample represents a cluster at first. After that, each pair of clusters with the highest similarity are merged until finally all samples belong to the same cluster (Farahbakhsh et al. 2021). As shown in Fig. 1, the final result for both DHC and AHC is in the form of a dendrogram. Each level in the dendrogram represents a partition as the result of clustering.

Linkage-based metrics are one of the most common AHC methods, which are defined by inter-cluster distance metrics (Rostami et al. 2023). Single linkage, average linkage, centroid linkage, and complete linkage are examples of linkage-based AHC clustering. A summary of these methods is presented in Table 1. In this table, $x \in c_i$ represents sample x from cluster c_i, $|c_i|$ indicates the number of cluster members c_i and $d_{x,y}$ indicates the distance between x and y based on a distance measure such as Euclidean (Sivadas et al. 2022). Basically, the difference between these methods is in the distance calculation metric.

Semi-supervised clustering

In unsupervised clustering, the learning algorithm has no knowledge about the labels of the samples. However, semi-supervised clustering can use prior knowledge such as labels of samples for clustering (Wang et al. 2023; Yue et al. 2023). Usually, the prior knowledge used by semi-supervised learning is known as constraint information (Sangeetha and Prakash 2021). Constraint information can include dependencies between samples or an additional set of labeled samples. Pairwise constraints information is the most common prior knowledge used for semi-supervised learning. Pairwise constraints include pairs of samples that are labeled as belonging to the same or different clusters. Therefore, the quality of the partition created by semi-supervised clustering should be improved compared to unsupervised clustering, because semi-supervised clustering uses prior knowledge.

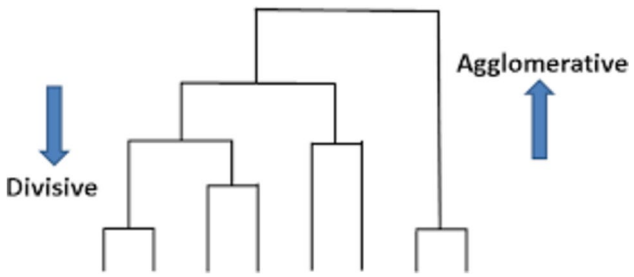

Fig. 1 An example of hierarchical clustering

Basically, the constraint information can be based on metrics, clusters, and samples (Rostami et al. 2023). Metric-based constraint information allows the use of different distance/similarity measures in the learning process. Cluster-based constraint information provides the possibility of using cluster characteristics such as shape, size, and diameter. Also, sample-based constraint information includes must-link and cannot-link parameters (Jannesari et al. 2023). Here, must-link indicates the possibility of assigning two samples to one cluster, while cannot-link indicates the impossibility of assigning two samples to one cluster. Selecting the most potential sample for semi-supervised clustering is an important challenge for using information constraints (Shahidinejad et al. 2021). Since the labels of samples are not available in clustering, dense groups should be identified in order to find samples that definitely belong to the same cluster.

According to the above, semi-supervised clustering simultaneously uses both labeled and unlabeled samples, as shown in Fig. 2. Typically, semi-supervised clustering is configured based on a small number of labeled samples compared to a large number of unlabeled samples. Constraint-based semi-supervised clustering and distance-based semi-supervised clustering are two common categories of semi-supervised clustering (Hayashi et al. 2018). The former uses constraint information to support the algorithm and improve clustering, while the latter includes adaptive distance metrics to extract constraints in supervised learning.

Ensemble clustering

It has been proven that no individual clustering method can provide the best performance for all contexts (Sivadas et al. 2022). Since each individual clustering method has its own

Table 1 AHC clustering methods based on linkage

Linkage method	Distance function	Description
Single linkage	$\min_{x\in c_i, y\in c_j} d_{x,y}$	This method measures the distance between two clusters by considering the closest members between them
Average linkage	$\frac{1}{\lvert c_i\rvert\times\lvert c_j\rvert}\sum_{x\in c_i, y\in c_j} d_{x,y}$	This method measures the distance between two clusters by considering the average distance between all their members
Centroid linkage	$\max_{x\in c_i, y\in c_j} d_{x,y}$	This method measures the distance between two clusters by considering the distance between their mean center vectors
Complete linkage	$d\left(\frac{\sum_{x\in c_i} x)}{\lvert c_i\rvert}\right).\left(\frac{\sum_{y\in c_j} y)}{\lvert c_j\rvert}\right)$	This method measures the distance between two clusters by considering their farthest neighbor

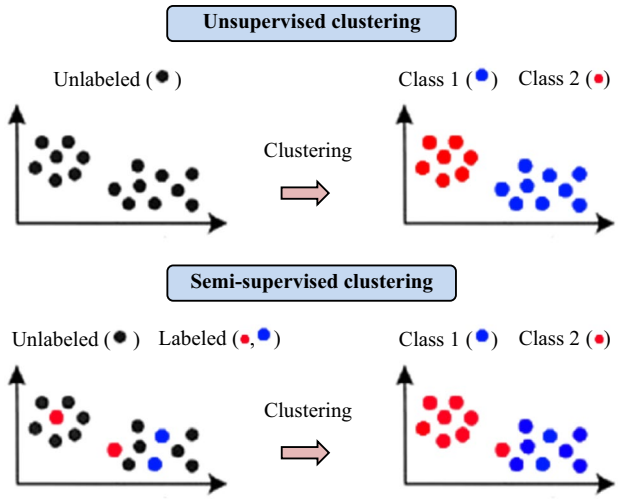

Fig. 2 Example of clustering with unsupervised and semi-supervised learning

advantages and disadvantages, combining several methods can provide more stable, scalable and accurate results compared to each of the individual methods. Ensemble clustering-based methods combine the results of several clustering methods to avoid the disadvantages of each of them and enable effective clustering for more datasets. As shown in Fig. 3, ensemble clustering consists of a number of individual homogeneous or heterogeneous clustering algorithms. These algorithms are considered as members of ensemble clustering. Selecting suitable members that can achieve quality and diversity in the final consensus is an important challenge in ensemble clustering.

Each individual clustering algorithm π_k is applied as a weak method on the dataset and outputs a partition p_k. The partitions created in this step are merged by a consensus function Γ to create the final partition p_*. Although all partitions can participate in the consensus process, a subset of primary partitions or part of their associated clusters can be candidates for the consensus function. This is a major challenge to address in ensemble clustering. Therefore, ensemble clustering has two main phases: creating primary partitions and merging them by a consensus function (Forouzandeh et al. 2023). The consensus function is an important issue in ensemble clustering, for which various methods have been introduced so far. The most common consensus functions include simple voting, iterative voting, weighted similarity, mixture model, correlation matrix, meta-clustering, etc.

In various studies, the superiority of semi-supervised clustering algorithms against unsupervised clustering has been proven (Sangeetha and Prakash 2021). Meanwhile, ensemble clustering provides better performance than individual clustering. With this motivation, we focus on SSEC-based approaches. The use of constraint information in SSEC is a hot research topic in machine learning. Here, prior knowledge such as pairwise constraints and labels of samples are incorporated into ensemble clustering in order to improve efficiency. Most of the existing SSEC approaches use constraint information only to create primary partitions, while the use of this information is ignored in the consensus function (Rezaeipanah and Ahmadi 2022). Figure 4 shows a schematic framework of SSEC-based approaches considering prior knowledge.

Feature sampling

Today, the number of large-scale datasets has increased significantly due to the growth of data collection devices (Zhao et al. 2023b; Cao et al. 2023b). Machine learning algorithms for effective analysis of these datasets face serious challenges. Meanwhile, clustering algorithms face issues such as feature correlation, noise, sparseness, and computational complexity when processing big data, which may lead to their failure. Reducing the dimensions of the data by selecting a subset of the original features is one of the most common solutions to address this problem (Rezaeipanah and Ahmadi 2022).

Techniques based on randomization such as random projection (Rostami et al. 2023) and random feature sampling (Sangeetha and Prakash 2021) are among the most common methods for selecting the subset of suitable features. However, randomization-based techniques do not consider correlations between features and cannot select effective features

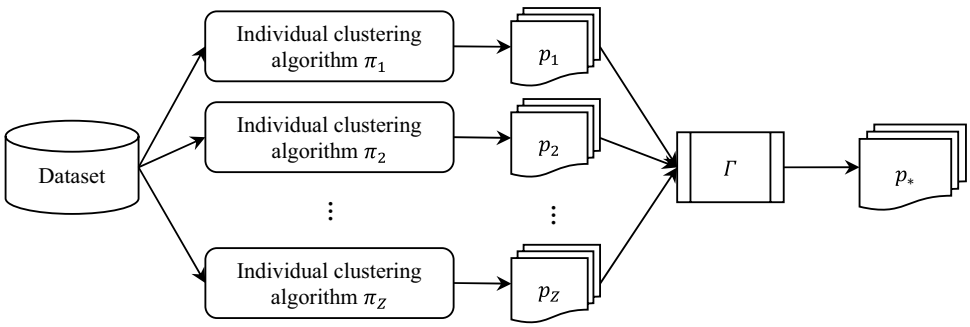

Fig. 3 Ensemble clustering architecture

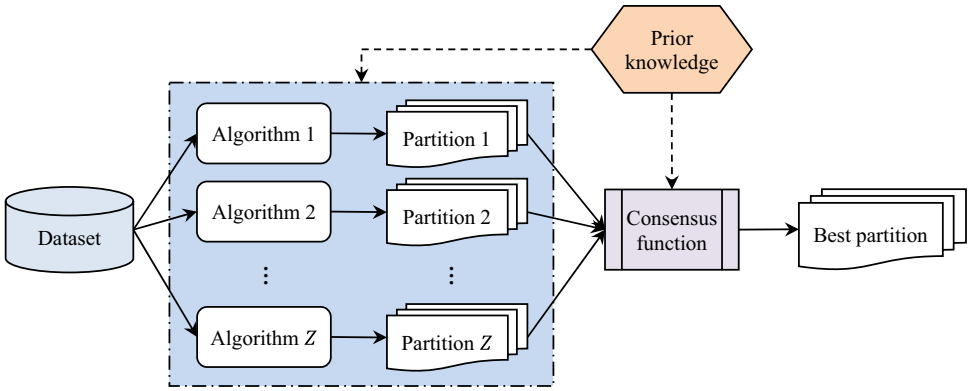

Fig. 4 SSEC framework considering prior knowledge

for clustering. Stratified feature sampling mechanism was introduced by Jing et al. (2015) to address this issue. This mechanism uses the k-means algorithm to cluster features into a specified number of groups. After that, a number of features are randomly selected from each cluster with the same proportion to obtain several subsets of features. The ensemble clustering architecture considering feature sampling is shown in Fig. 5.

Proposed clustering framework

The proposed clustering framework has four general phases. In the first phase, stratified feature sampling mechanism is applied. This mechanism clusters the features of the dataset using the K-means algorithm to create an independent subset of features for each individual clustering algorithm. Here, feature selection probabilities are adjusted with the aim of reducing the risk of not selecting some features for the clustering task. The second phase is related to the generation of primary partitions by Z individual clustering algorithms. We use AHC-based algorithms for the clustering task, where each algorithm creates its own partition based on a subset of specified features. The output partition in each AHC-based algorithm is determined from the dendrogram by Bayesian PAC theory (Abddallah and Yousef 2018).

The third phase consists of presenting a new similarity metric that uses a wide range of information to calculate the similarity between each sample pair, cluster pair and meta-cluster pair. The consensus function is configured in the fourth phase. Since not all primary clusters and not all primary partitions have the same strength, we develop a weighting policy in which the merits of clusters and the strength of partitions are considered to contribute to the final consensus. Finally, the meta-clustering technique is applied as a consensus function to create the final partition. We configure each AHC-based clustering algorithm with semi-supervised learning and use the information of pairwise constraints to improve the clustering performance in both parts of creating primary partitions and the consensus function. An overview of the proposed clustering framework is shown in Fig. 6.

The proposed algorithm for large-scale data clustering uses the stratified feature sampling mechanism. In this mechanism, each $\pi_k \in \Pi$ performs clustering based on a subset of the main features. Let π_k form an primary partition based on $\mathcal{s}_k$, where $\mathcal{s}_k \in \mathcal{S}$ represents the subset of the k-th selected feature. The mechanism of stratified feature sampling can provide the most suitable set $\mathcal{S}$ for ensemble clustering. Here, the features of the dataset X are clustered by K-means, and then a number of features are sampled from each cluster to form $\mathcal{s}_k$. This process is applied to all $\mathcal{s}_k \in \mathcal{S}, \forall k = 1, 2, \ldots, Z$.

To reduce the risk of not selecting some features, we calculate the probability of selecting the features by considering the sampling history. Let $\mathcal{r}_j$ refer to the sampling rate of the jth feature from the dataset X. Here, the sampling rate for selecting the first subset is the same for all features, for example, $\mathcal{r}_j = 1/M$. The sampling rate is updated to select

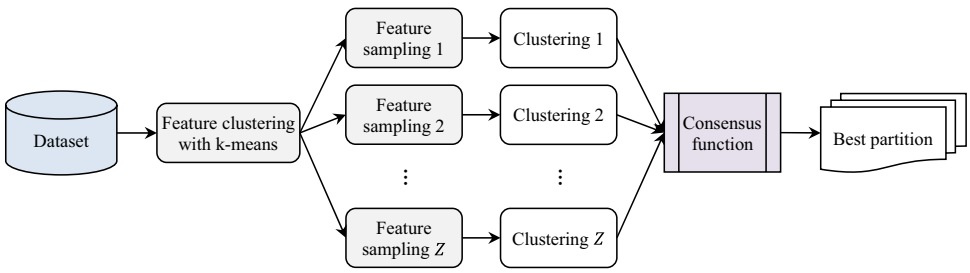

Fig. 5 Ensemble clustering architecture considering feature sampling

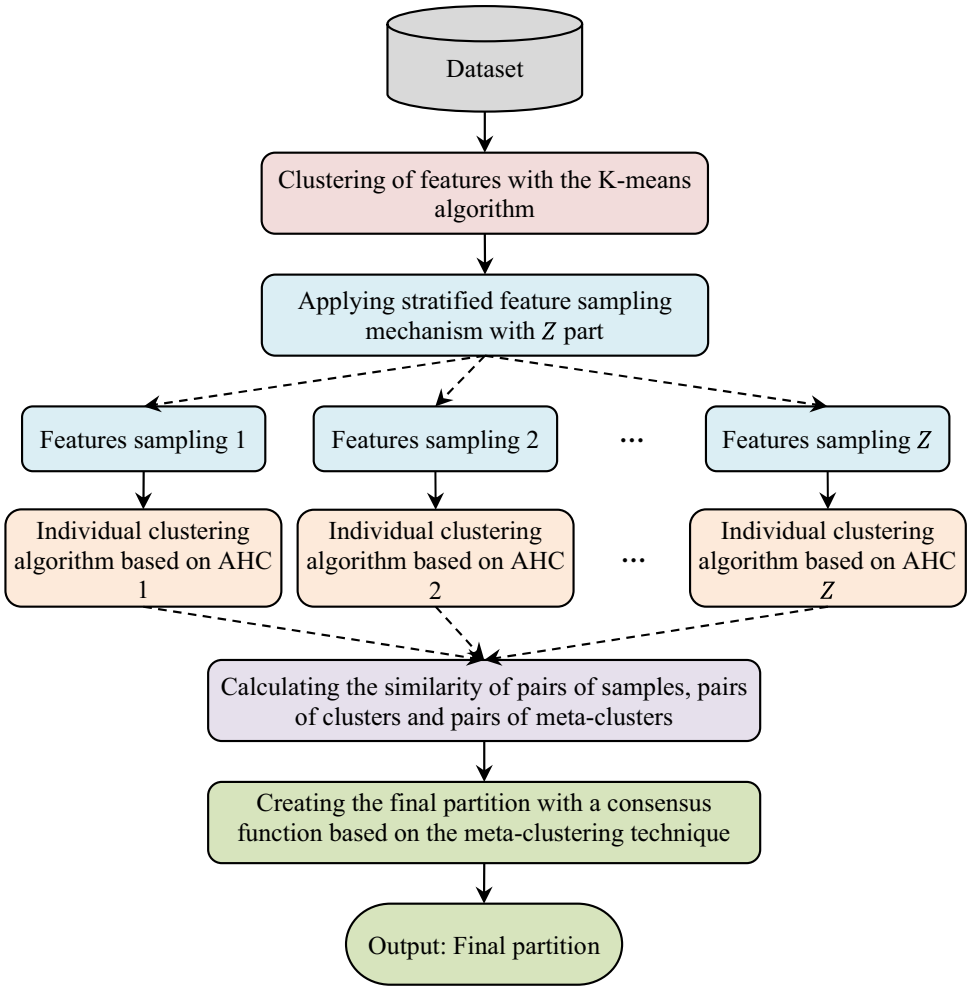

Fig. 6 An overview of the proposed clustering framework

the second subset, where the sampling rate of unselected features is halved. This process is repeated for other subsets to reduce the risk of not selecting features.

Let $s_{i,j} \in S$ be the similarity between samples x_i and x_j. We use a new similarity metric considering a wide range of information to calculate the similarity matrix S. The Eq. (1) defines the similarity for $s_{i,j}$.

$$s_{i,j} = \frac{1}{Z} \times \sum_{p_k \in P} \left[\frac{1}{|p_k|} \times \sum_{c_{k,l} \in p_k} \begin{cases} \frac{M_{c_{k,l}} + W_{p_k}}{d_{i,j}} \times \beta^{|c_{k,l}|} & (x_i, x_j) \in c_{k,l} \\ \frac{1}{d_{i,j}} \times \beta^{|c_{k,l}|} & \text{otherwise} \end{cases} \right], \tag{1}$$

where Z is the total number of partitions, p_k is the detail of the kth partition, P is the set of all partitions, $|p_k|$ is the number of clusters in p_k, $c_{k,l}$ is the detail of the lth cluster in p_k, $|c_{k,l}|$ is the number of samples of $c_{k,l}$, $d_{i,j}$ is the Euclidean distance between x_i and x_j, $M_{c_{k,l}}$ is the merit associated with $c_{k,l}$, W_{p_k} is the strength/weight associated with p_k, and β is a damping factor to reduce the effect of large cluster sizes.

In addition to the similarity between each pair of samples, we calculate the similarity between each pair of clusters and each pair of meta-clusters. Let each meta-cluster be a set of several clusters. Equation (2) formulates the similarity between two clusters $c_{k,1}$ and $c_{k,2}$ as $\mathrm{CS}_{c_{k,1},c_{k,2}}$. Also, Eq. (3) formulates the similarity between two meta-clusters $\psi_1 = \left\{ c_{1,1}, c_{1,2}, \ldots, c_{1,u}, \ldots, c_{1,|\psi_1|} \right\}$ and $\psi_2 = \left\{ c_{2,1}, c_{2,2}, \ldots, c_{2,v}, \ldots, c_{2,|\psi_2|} \right\}$ as $\mathrm{MS}_{\psi_1,\psi_2}$.

$$\mathrm{CS}_{c_{k,1},c_{k,2}} = \frac{\sum_{i=1}^{|c_{k,1}|} \sum_{j=1}^{|c_{k,2}|} s_{i,j}}{|c_{k,1}|.|c_{k,2}|}, \tag{2}$$

$$\mathrm{MS}_{\psi_1,\psi_2} = \frac{\sum_{u=1}^{|\psi_1|} \sum_{v=1}^{|\psi_2|} CS_{c_{1,u},c_{2,v}}}{|\psi_1| \times |\psi_2|}. \tag{3}$$

Finally, we use a consensus function based on the meta-clustering technique to create the final partition. According to this technique, candidate clusters are considered from all partitions in a set and then re-clustered by average linkage to create meta-clusters. Here, the number of meta-clusters represents the number of final clusters. Eventually, the final partition is created by assigning each sample of the dataset

X to a meta-cluster with the highest similarity. In this paper, candidate clusters are selected to participate in the final consensus based on the merit of the primary clusters and the strength of the primary partitions. In extensive studies, normalized mutual information (NMI) is used to evaluate the partition generated from a clustering algorithm (Rezaeipanah and Ahmadi 2022). NMI can calculate the similarity between two partitions such as p_u and p_v by Eq. (4).

$$\mathrm{NMI}(p_u, p_v) = \frac{2\sum_{i=1}^{|p_u|}\sum_{j=1}^{|p_v|} N_{ij}\log\left(\frac{N.N_{ij}}{N_{i1}.N_{2j}}\right)}{\sum_{i=1}^{|p_u|} N_{i1}\log\left(\frac{N_{i1}}{N}\right) + \sum_{j=1}^{|p_v|} N_{2j}\log\left(\frac{N_{2j}}{N}\right)}, \tag{4}$$

where N_{ij} is the number of identical samples in $c_{u,i} \in p_u$ and $c_{v,j} \in p_v$ and N_{iu} is the number of samples in $c_{u,i}$.

If p_v is assumed as the reference partition, then $\mathrm{NMI}(p_u, p_v)$ represents the strength of the partition p_u. Let the strength of partition p_u be formulated as the weight of partition p_u by W_{p_u}. In addition to robustness, we use the merit of the clusters to determine the candidate clusters in the final consensus. Law et al. (2004) developed the NMI criterion and used it to calculate the merit of clusters. The authors converted a cluster into a partition in order to use NMI for evaluation work. Let $\bar{c}_k$ be a cluster with all samples not in c_k. c_k is considered a positive cluster if at least half of its samples are found in another cluster. According to these definitions, the cluster c_k is considered as a partition $\hat{p}_k = \{c_k, \bar{c}_k\}$ with the union of all positive clusters. With converting c_k to $\hat{p}_k$, cluster merit of c_k is formulated by Eq. (5). According to the aforementioned concepts, each $c_{k,l} \in p_k$ with a predefined threshold can participate in the final consensus. The Eq. (6) defines the condition for $c_{k,l}$ to be a candidate for participating in the final consensus.

$$M_{c_k} = \mathrm{NMI}(p_0, \hat{p}_k), \tag{5}$$

$$\left(\xi \times W_{p_k} + (1-\xi) \times M_{c_{k,l}}\right) \geq \theta, \tag{6}$$

where p_0 is defined as the reference partition. Also, ξ is the influence coefficient of the cluster level with respect to the partition level and θ is a threshold for determining the consensus candidates.

Each $\pi_k \in \Pi$ is an individual clustering algorithm based on AHC such as average linkage. Here, all $\pi_k \in \Pi$ are configured using average linkage and based on semi-supervised learning. Also, the algorithm used in the consensus function is applied using average linkage and based on semi-supervised learning. Let $d_{i,j}$ be the distance between samples x_i and x_j. We use the information of pairwise constraints such as must-link and cannot-link to define $d_{i,j}$ in semi-supervised learning. If the sample pair (x_i, x_j) is covered by the must-link, then it belongs to the set Δ_M. Meanwhile, if the sample pair (x_i, x_j) is covered by cannot-link, then it belongs to the set Δ_C. Let all members of sets Δ_M and Δ_C have symmetry and transitivity properties. The symmetry property is formulated by Eq. (7) and the transitivity property is formulated by Eq. (8). Considering semi-supervised learning in the average linkage algorithm, $d_{i,j}$ is formulated by pairwise constraints information with Eq. (9).

$$\begin{cases} (x_i, x_j) \in \Delta_M \rightarrow (x_j, x_i) \in \Delta_M \\ (x_i, x_j) \in \Delta_C \rightarrow (x_j, x_i) \in \Delta_C, \end{cases} \tag{7}$$

$$\begin{cases} (x_i, x_k)\&(x_k, x_j) \in \Delta_M \rightarrow (x_i, x_j) \in \Delta_M \\ (x_i, x_k)\&(x_k, x_j) \in \Delta_C \rightarrow (x_i, x_j) \in \Delta_C, \end{cases} \tag{8}$$

$$d_{i,j} = \begin{cases} 0 & (x_i, x_j) \in \Delta_M \\ \infty & (x_i, x_j) \in \Delta_C. \end{cases} \tag{9}$$

Experiments

We validate the performance of the proposed framework with several numerical experiments considering the UCI dataset and then use it to extract intercellular communication on the FANTOM5 dataset. The proposed clustering algorithm has been implemented using the MATLAB 2021a simulator on a personal computer with Intel® Core™ i7 Processor up to 3.4.00 GHz and 16 GB DDR3 Memory.

Datasets

The evaluations are based on 10 datasets from the UCI machine learning repository, as shown in Table 2. We use a mean replacement policy when dealing with missing values. All datasets used have class labels, which are used as reference partitions in clustering. Since the proposed clustering framework is based on semi-supervised learning, we consider 5% of the supervised samples as the constraint information.

In addition, we use the FANTOM5 dataset to analyze gene expression data and extract intercellular communication. FANTOM5 was compiled in collaboration with the University of Sydney, Australia. In addition to cell information, this dataset also contains tissue information, which is not considered in the current study. Details of this dataset are available at http://fantom.gsc.riken.jp/5. The full version of the FANTOM5 dataset contains 1836 samples per column, where each sample contains information related to a cell or tissue from a single patient. For each sample, 201,802 promoters from different regions of a gene from a specific cell are available. With filtering data related to tissues, we

Table 2 Details of the datasets used in the simulations

Dataset	Number of samples	Number of features	Number of classes
Iris	150	4	3
Titanic	24	2	2
Brain	42	5597	5
Laryngeal1	213	16	2
Colon	62	2000	2
ORL32	400	1024	40
Pendigits	10,992	16	10
Banana	5300	2	2
Digits	5620	63	10
Splice	2991	60	3

found 108 unique cells. Here, there are 702 examples related to cells.

Meanwhile, the rows in this dataset represent the numbers of promoters, which are identified using "entrezgene_id". Some promoter values are not specified and specifically have the value "NA". Unavailable promoter information is removed. After that, 86,428 promoters are available for each sample. The columns related to cells are taken from different samples of the human body and there may be several samples of the same cell. In general, the first 7 columns related to the promoter information have been sampled and the 8th columns are samples. In addition, the ID of each sample includes details such as disease type, time point, cell name and patient ID. For example, the ID of a sample from the FANTOM5 dataset is: "239SLAM rinderpest infection, 00hr, biol_rep1.CNhs14406.13541-145H4". Here, "SLAM" represents a family of cell surface receptors and other coding are related to the patient. An overview of the FANTOM5 dataset for cells is shown in Table 3.

Evaluation metrics

A partition generated by a clustering algorithm is ideal if it has a maximum inter-cluster distance and a minimum intra-cluster distance. We use criteria such as NMI, Adjusted Rand Index (ARI) and silhouette coefficient to evaluate the clustering results (Talatian Azad et al. 2021). NMI is a common criterion for evaluating the performance of clustering algorithms that can measure the similarity between two independent partitions. NMI is defined according to Eq. (4). ARI is another criterion for evaluating the performance of clustering algorithms. ARI uses the Rand Index (RI) to calculate the similarity between two independent partitions. ARI can calculate the similarity between two partitions such as p_u and p_v by Eq. (10).

$$\mathrm{ARI}(p_u,p_v)=\frac{\sum_{i=1}^{|p_u|}\sum_{j=1}^{|p_v|}\binom{N_{ij}}{2}-\frac{\left[\sum_{i=1}^{|p_u|}\binom{N_{iu}}{2}\sum_{j=1}^{|p_v|}\binom{N_{vj}}{2}\right]}{\binom{N}{2}}}{\frac{1}{2}\left[\sum_{i=1}^{|p_u|}\binom{N_{iu}}{2}+\sum_{j=1}^{|p_v|}\binom{N_{vj}}{2}\right]-\frac{\left[\sum_{i=1}^{|p_u|}\binom{N_{iu}}{2}\sum_{j=1}^{|p_v|}\binom{N_{vj}}{2}\right]}{\binom{N}{2}}}. \tag{10}$$

The silhouette coefficient is an internal index to calculate the performance of clustering algorithms, which performs the evaluation process based on density and separation characteristics. In silhouette, the validity of a partition is calculated based on the combination of intra-cluster and inter-cluster similarity for each pair of independent clusters. The obtained value of the silhouette coefficient is between -1 and $+1$, and a silhouette with a value of $+1$ represents an ideal clustering. The silhouette coefficient for $x_i \in c_l$ from the p_k partition is calculated by Eq. (11).

$$Sil_i=\frac{a_i-b_i}{\max(a_i,b_i)}, \tag{11}$$

where a_i and b_i are calculated by Eqs. (12) and (13), respectively.

$$a_i=\frac{1}{|c_l|}\sum_{x_j\in X|x_j\in c_l} d_{i,j}, \tag{12}$$

$$b_i=\min_{c_q\in p_k|c_q\neq c_l}\left(\frac{1}{|c_q|}\sum_{x_j\in X|x_j\in c_q} d_{i,j}\right). \tag{13}$$

Table 3 Overview of the FANTOM5 dataset

Gene region	Cell 1		…	Cell 2		…	Cell *i*	…	Cell 108		
	Sample 1	Sample 2	…	Sample *j*	Sample *j* + 1	…	…	…	…	…	Sample 702
1	Promoter	Promoter	…	Promoter	Promoter	…	Promoter	…	…	…	Promoter
2	Promoter	Promoter	…	Promoter	Promoter	…	Promoter	…	…	…	Promoter
3	Promoter	Promoter	…	Promoter	Promoter	…	Promoter	…	…	…	Promoter
…	…	…	…	…	…	…	…	…	…	…	…
86,428	Promoter	Promoter	…	Promoter	Promoter	…	Promoter	…	…	…	Promoter

Analysis of results

The proposed clustering algorithm is compared with several equivalent algorithms such as MWAMLP (Rezaeipanah and Ahmadi 2022), ECIDB (Mojarad et al. 2021), SSAE (Sangeetha and Prakash 2021), and TCGA (Sivadas et al. 2022). Before the comparisons, we prove that the proposed clustering algorithm using the average linkage algorithm provides the best performance in both the creation of primary partitions and the consensus function. We compare the average linkage algorithm with other AHC-based algorithms such as single linkage, centroid linkage and complete linkage. Table 4 shows the results of this comparison. The results of this comparison are presented based on accuracy and the best results are bolded. Also, each row presents the results associated with a dataset, while the last row is the average of the results. The results clearly prove the superiority of the average linkage algorithm and its use in the proposed clustering framework.

The comparison of the proposed algorithm based on NMI and ARI criteria compared to MWAMLP, ECIDB, SSAE and TCGA is presented in Tables 5 and 6, respectively. The best results of these tables are highlighted in bold. The proposed algorithm performs better than all existing algorithms in many datasets. However, the simulation results show that ECIDB produces quite competitive results with the proposed algorithm. Among the 10 existing ECIDB datasets, the proposed algorithm outperforms the proposed algorithm considering the NMI criterion in the Iris and Colon datasets. Also, ECIDB performs best considering the ARI criterion on the Titanic, Banana and Splice datasets. On average, in the NMI criterion, the proposed algorithm is 8.8%, 1.7%, 12.9%, and 16.5% superior compared to MWAMLP, ECIDB, SSAE, and TCGA, respectively. This superiority for the ARI criterion is 4.6%, 1.8%, 11.5%, and 8.1%, respectively.

Although the proposed clustering algorithm performs better in terms of accuracy, NMI and ARI compared to equivalent algorithms, runtime analysis is also important. High-complexity clustering algorithms are not capable of processing large-scale datasets. The proposed clustering algorithm is equipped with a stratified feature sampling mechanism to deal with big data. This mechanism leads to the reduction of computational complexity and it is expected that the runtime in the proposed algorithm is lower than other algorithms. Figure 7 shows the runtime results of

Table 4 Comparison of average linkage algorithm compared to other AHC-based algorithms

Dataset	Single linkage	Average linkage	Centroid linkage	Complete linkage
Iris	0.8706 ± 0.017	0.9073 ± 0.014	**0.9105 ± 0.018**	0.8940 ± 0.019
Titanic	0.7718 ± 0.009	**0.7990 ± 0.022**	0.7985 ± 0.023	0.7820 ± 0.029
Brain	0.5215 ± 0.025	**0.5274 ± 0.011**	0.4846 ± 0.010	0.5169 ± 0.008
Laryngeal1	0.9101 ± 0.023	**0.9184 ± 0.012**	0.9183 ± 0.017	0.9166 ± 0.006
Colon	0.7370 ± 0.018	0.7375 ± 0.018	0.7313 ± 0.015	**0.7460 ± 0.017**
ORL32	0.5952 ± 0.021	**0.6067 ± 0.006**	0.6034 ± 0.018	0.5842 ± 0.018
Pendigits	0.4775 ± 0.016	**0.4893 ± 0.036**	0.4889 ± 0.019	0.4687 ± 0.009
Banana	0.7419 ± 0.027	**0.7743 ± 0.011**	0.7714 ± 0.019	0.7608 ± 0.011
Digits	0.8413 ± 0.011	0.8550 ± 0.012	**0.8580 ± 0.029**	0.8382 ± 0.028
Splice	0.5907 ± 0.009	**0.5958 ± 0.027**	0.5953 ± 0.013	0.5952 ± 0.015
Average	0.7058	**0.7211**	0.7160	0.7103

Table 5 Comparison of different algorithms in terms of NMI criterion

Dataset	MWAMLP	ECIDB	SSAE	TCGA	Proposed algorithm
Iris	0.8860 ± 0.014	**0.8952 ± 0.017**	0.8706 ± 0.023	0.8737 ± 0.014	0.8851 ± 0.026
Titanic	0.2781 ± 0.020	0.3141 ± 0.015	0.2709 ± 0.023	0.2442 ± 0.012	**0.3315 ± 0.026**
Brain	0.4819 ± 0.028	0.4989 ± 0.018	0.3729 ± 0.010	0.2235 ± 0.023	**0.5322 ± 0.029**
Laryngeal1	0.3991 ± 0.015	0.4195 ± 0.027	0.4166 ± 0.028	0.3290 ± 0.019	**0.4464 ± 0.027**
Colon	0.5115 ± 0.025	**0.7873 ± 0.023**	0.6825 ± 0.027	0.6616 ± 0.012	0.7329 ± 0.008
ORL32	**0.6023 ± 0.014**	0.5688 ± 0.009	0.5177 ± 0.010	0.5088 ± 0.023	0.5549 ± 0.026
Pendigits	0.6699 ± 0.012	0.6997 ± 0.025	0.5813 ± 0.012	0.7507 ± 0.034	**0.7185 ± 0.028**
Banana	0.7379 ± 0.027	0.7355 ± 0.013	0.6977 ± 0.020	0.7237 ± 0.027	**0.7770 ± 0.017**
Digits	0.8381 ± 0.012	0.8610 ± 0.010	0.7444 ± 0.021	0.7786 ± 0.029	**0.8949 ± 0.028**
Splice	0.3748 ± 0.016	0.4101 ± 0.018	0.4187 ± 0.024	0.3098 ± 0.009	**0.4234 ± 0.025**
Average	0.5780	0.6190	0.5573	0.5404	**0.6297**

Table 6 Comparison of different algorithms in terms of ARI criterion

Dataset	MWAMLP	ECIDB	SSAE	TCGA	Proposed algorithm
Iris	**0.7852 ± 0.027**	0.7799 ± 0.024	0.7733 ± 0.017	0.7722 ± 0.016	0.7841 ± 0.010
Titanic	0.4464 ± 0.016	**0.4872 ± 0.022**	0.3103 ± 0.022	0.4377 ± 0.026	0.4870 ± 0.023
Brain	0.3185 ± 0.022	0.3325 ± 0.011	0.3160 ± 0.012	0.3185 ± 0.012	**0.3499 ± 0.027**
Laryngeal1	0.4217 ± 0.018	0.4224 ± 0.024	0.4158 ± 0.028	0.4103 ± 0.029	**0.4592 ± 0.027**
Colon	0.7757 ± 0.019	0.7624 ± 0.010	0.7493 ± 0.008	0.7592 ± 0.025	**0.7722 ± 0.012**
ORL32	0.5175 ± 0.018	0.5525 ± 0.021	0.5026 ± 0.021	0.5087 ± 0.023	**0.5601 ± 0.010**
Pendigits	0.5866 ± 0.027	0.6244 ± 0.024	0.5776 ± 0.008	0.5746 ± 0.019	**0.6479 ± 0.027**
Banana	0.7357 ± 0.011	**0.7561 ± 0.028**	0.7157 ± 0.012	0.7239 ± 0.017	**0.7546 ± 0.022**
Digits	0.8429 ± 0.028	0.8572 ± 0.036	0.7386 ± 0.023	0.7681 ± 0.025	**0.8728 ± 0.011**
Splice	0.4290 ± 0.012	**0.4591 ± 0.018**	0.4072 ± 0.024	0.4055 ± 0.028	**0.4528 ± 0.018**
Average	0.5869	0.6034	0.5506	0.5679	**0.6141**

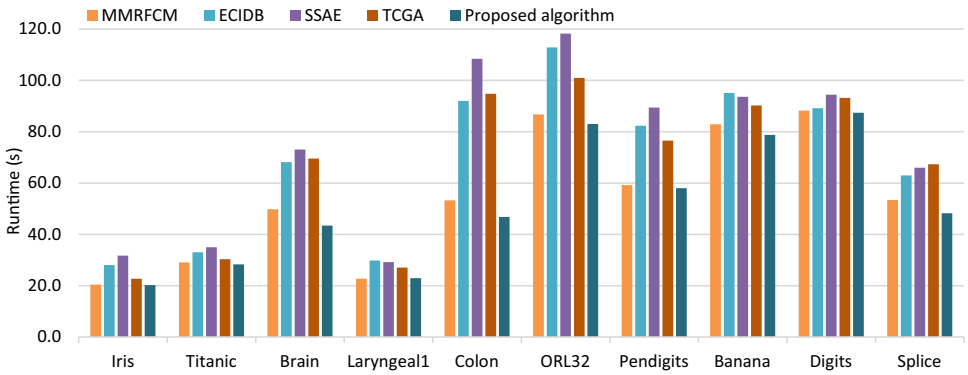

Fig. 7 Comparison of different algorithms in terms of running time

different clustering algorithms. The results clearly show that our algorithm has lower runtime in all datasets. On average, the proposed clustering algorithm provides 6.1%, 34.6%, 43.5%, and 30.8% less runtime compared to MWAMLP, ECIDB, SSAE, and TCGA algorithms, respectively.

We proved that the proposed clustering framework has ideal performance for clustering real-world datasets. Hence, we apply it to clustering the FANTOM5 dataset and extracting cell-to-cell biological communications. The FANTOM5 dataset is multifaceted, where multiple samples from the same cell with multiple patients are available. Also, there are different samples of the same cell in different diseases. Therefore, each cell may be related to other cells through various diseases. The concept of communication in FANTOM5 is expressed with promoters. A high value of a promoter indicates the reproduction or disruption of a part of gene expression related to a cell. The activation threshold of promoters has a significant effect on the discovery of intercellular communication. Here, we cluster with different thresholds from 500 to 4000 samples of the FANTOM5 dataset and report the results in terms of the silhouette coefficient. We compare the presented results with ECIDB (Mojarad et al. 2021), as this algorithm was also applied to the FANTOM5 dataset. The results of this comparison are presented in Table 7. The results show the superiority of the proposed algorithm in most thresholds. Meanwhile, the best results are obtained for the silhouette factor with a threshold of 1000. Here, the proposed algorithm with a silhouette coefficient of 0.952 and 19 clusters of samples related to cells have been clustered. These results were obtained for ECIDB with silhouette coefficient equal to 0.809 and 20 clusters.

We analyzed the clustering of the FANTOM5 dataset with different thresholds. A suitable threshold is equal to 1000, considering it leads to the identification of strong communications between cells. In each cluster, the pair of cells with the strongest correlation may indicate a relationship between different diseases. We extracted pairs of cells from different clusters with the highest correlation, whose samples belong to different diseases. Table 8 shows some of the strongest cell-to-cell communications, along with disease names and genes sampled. It shows the hereditary behavior between which diseases, based on which genes and in which cells.

Conclusions

Gene expression data contain important information of various diseases. The gene expression data of some diseases may be similar. Indeed, some cells in different

Table 7 Comparison of FANTOM5 dataset clustering results in terms of silhouette coefficient with different thresholds

Algorithms	Component	Thresholds							
		500	1000	1500	2000	2500	3000	3500	4000
ECIDB algorithm	Number of clusters	17	20	20	13	11	13	9	12
	Silhouette coefficient	0.694	0.809	0.652	0.805	0.75	0.486	0.432	0.282
Proposed algorithm	Number of clusters	16	19	16	15	12	12	17	15
	Silhouette coefficient	0.762	0.952	0.649	0.727	0.683	0.694	0.594	0.393

Table 8 Number of the strongest cell-to-cell communications identified

Cell name	Disease name	Gene name	Disease name	Cell name
hes.gfp.embryonic.stem.cells	Cancer	ABLIM1	Inflammation/monocytosis	cd14.cd16..monocytes.2
hes.gfp.embryonic.stem.cells	Cancer	ABLIM1	Aortic aneurysm	fibroblast…aortic.advantitial donor2..cytoplasmic.fraction
hes.gfp.embryonic.stem.cells	Cancer	ABLIM1	Inflammation/ Monocytosis	cd14.cd16..monocytes.1
Basophils	Blood coagulation	ABLIM1	Urinary tract infections	ciliary.epithelial.cells
cd14..monocytes…treatedwith. lipoplysaccharide	Alzheimer	ABLIM1	Urinary tract infections	ciliary.epithelial.cells

diseases may contain similar gene expression data. Therefore, discovering the relationships between diseases through the extraction of cell-to-cell biological communications is challenging and can change our understanding of how diseases such as cancer develop. The communication between two cells occurs when the number of promoters is significantly expressed in a number of cells. It is obvious that designing a method to discover cell-to-cell biological communications and identify the real communication between diseases is important for the medical society. A clustering algorithm based on semi-supervised learning and ensemble technique was proposed in the paper to identify intercellular communication. This framework is equipped with a stratified feature sampling mechanism to deal with high-dimensional data. Also, in this framework, a new similarity metric is developed that uses a wide range of primary partition information to estimate similarity. Our proposed framework uses the constraints information in both the phases of creating the primary partitions and the consensus function. The performance of the proposed framework has been validated through clustering of the UCI dataset. Therefore, the proposed framework for extracting intercellular communication was successfully applied to the FANTOM5 dataset. The results of the simulations show that the most promoters between cancer and diseases such as inflammation, monocytosis and aortic aneurysm occur on the "ABLIM1" gene.

Author contributions All authors contributed to the design and implementation of the research, to the analysis of the results and to the writing of the manuscript.

Funding This work was supported by the talent project of Weinan Normal University. No.18ZRRC07.

Data availability Data sharing not applicable to this manuscript as no datasets were generated or analyzed during the current study.

Code availability There is no free code for this study.

Declarations

Conflict of interest The authors declare no competing interests.

Ethical approval Not applicable.

References

Abddallah L, Yousef M (2018) Ensemble clustering based dimensional reduction. In: Database and expert systems applications: DEXA 2018 international workshops, BDMICS, BIOKDD, and TIR, Regensburg, Germany, September 3–6, 2018, Proceedings 29. Springer International Publishing, pp 115–125

Bridges K, Miller-Jensen K (2022) Mapping and validation of scRNA-Seq-derived cell-cell communication networks in the tumor microenvironment. Front Immunol 13:885267

Cao C, Wang J, Kwok D, Cui F, Zhang Z, Zhao D et al (2022) webT-WAS: a resource for disease candidate susceptibility genes identified by transcriptome-wide association study. Nucleic Acids Res 50(D1):D1123–D1130

Cao Z, Niu B, Zong G, Xu N (2023a) Small-gain technique-based adaptive output constrained control design of switched networked nonlinear systems via event-triggered communications. Nonlinear Anal Hybrid Syst 47:101299

Cao Y, Xu N, Wang H, Zhao X, Ahmad AM (2023b) Neural networks-based adaptive tracking control for full-state constrained switched nonlinear systems with periodic disturbances and actuator saturation. Int J Syst Sci 54(14):2689–2704

Chang Y, Niu B, Wang H, Zhang L, Ahmad AM, Alassafi MO (2022) Adaptive tracking control for nonlinear system in pure-feedback form with prescribed performance and unknown hysteresis. IMA J Math Control Inf 39(3):892–911

Cheng F, Liang H, Niu B, Zhao N, Zhao X (2023) Adaptive neural self-triggered bipartite secure control for nonlinear MASs subject to DoS attacks. Inf Sci 631:256–270

de Souza PS, Faccion RS, Bernardo PS, Maia RC (2016) Membrane microparticles: shedding new light into cancer cell communication. J Cancer Res Clin Oncol 142:1395–1406

Farahbakhsh F, Shahidinejad A, Ghobaei-Arani M (2021) Multiuser context-aware computation offloading in mobile edge computing based on Bayesian learning automata. Trans Emerg Telecommun Technol 32(1):e4127

Forouzandeh S, Berahmand K, Sheikhpour R, Li Y (2023) A new method for recommendation based on embedding spectral clustering in heterogeneous networks (RESCHet). Expert Syst Appl 231:120699

Hayashi T, Ozaki H, Sasagawa Y, Umeda M, Danno H, Nikaido I (2018) Single-cell full-length total RNA sequencing uncovers dynamics of recursive splicing and enhancer RNAs. Nat Commun 9(1):619

Hou R, Denisenko E, Ong HT, Ramilowski JA, Forrest AR (2020) Predicting cell-to-cell communication networks using NATMI. Nat Commun 11(1):5011

Jannesari V, Keshvari M, Berahmand K (2023) A novel nonnegative matrix factorization-based model for attributed graph clustering by incorporating complementary information. Expert Syst Appl 242:122799

Jing L, Tian K, Huang JZ (2015) Stratified feature sampling method for ensemble clustering of high dimensional data. Pattern Recogn 48(11):3688–3702

Kayal CK, Bagchi S, Dhar D, Maitra T, Chatterjee S (2019) Hepatocellular carcinoma survival prediction using deep neural network. In: Proceedings of international ethical hacking conference 2018: eHaCON 2018, Kolkata, India. Springer, Singapore, pp 349–358

Law MH, Topchy AP, Jain AK (2004) Multiobjective data clustering. In: Proceedings of the 2004 IEEE computer society conference on computer vision and pattern recognition, 2004. CVPR 2004, vol 2. IEEE, pp II-II

Lei X, Li Z, Zhong Y, Li S, Chen J, Ke Y et al (2022) Gli1 promotes epithelial–mesenchymal transition and metastasis of non-small cell lung carcinoma by regulating snail transcriptional activity and stability. Acta Pharm Sin B 12(10):3877–3890

Li X, Chen X, Rezaeipanah A (2023) Automatic breast cancer diagnosis based on hybrid dimensionality reduction technique and ensemble classification. J Cancer Res Clin Oncol 149:7609–7627

Mojarad M, Sarhangnia F, Rezaeipanah A, Parvin H, Nejatian S (2021) Modeling hereditary disease behavior using an innovative similarity criterion and ensemble clustering. Curr Bioinform 16(5):749–764

Peng L, Wang F, Wang Z, Tan J, Huang L, Tian X et al (2022) Cell–cell communication inference and analysis in the tumour microenvironments from single-cell transcriptomics: data resources and computational strategies. Brief Bioinform 23(4):bbac234

Prades J, Safont G, Salazar A, Vergara L (2020) Estimation of the number of endmembers in hyperspectral images using agglomerative clustering. Remote Sens 12(21):3585

Rezaeipanah A, Ahmadi G (2022) Breast cancer diagnosis using multi-stage weight adjustment in the MLP neural network. Comput J 65(4):788–804

Rezaeipanah A, Syah R, Wulandari S, Arbansyah A (2021) Design of ensemble classifier model based on MLP neural network for breast cancer diagnosis. Intel Artif 24(67):147–156

Rostami M, Oussalah M, Berahmand K, Farrahi V (2023) Community detection algorithms in healthcare applications: a systematic review. IEEE Access 11:30247–30272

Sangeetha K, Prakash S (2021) An early breast cancer detection system using stacked auto encoder deep neural network with particle swarm optimization based classification method. J Med Imaging Health Inform 11(12):2897–2906

Shahidinejad A, Ghobaei-Arani M, Masdari M (2021) Resource provisioning using workload clustering in cloud computing environment: a hybrid approach. Clust Comput 24(1):319–342

Shahraki K, Boroumand PG, Lotfi H, Radnia F, Shahriari H, Sargazi S et al (2023) An update in the applications of exosomes in cancer theranostics: from research to clinical trials. J Cancer Res Clin Oncol 149:8087–8116

Sivadas A, Kok VC, Ng KL (2022) Multi-omics analyses provide novel biological insights to distinguish lobular ductal types of invasive breast cancers. Breast Cancer Res Treat 193(2):361–379

Talatian Azad S, Ahmadi G, Rezaeipanah A (2021) An intelligent ensemble classification method based on multi-layer perceptron neural network and evolutionary algorithms for breast cancer diagnosis. J Exp Theor Artif Intell 34(6):949–969

Tan J, Liu L, Li F, Chen Z, Chen GY, Fang F et al (2022) Screening of endocrine disrupting potential of surface waters via an affinity-based biosensor in a rural community in the Yellow River Basin, China. Environ Sci Technol 56(20):14350–14360

Tang F, Wang H, Chang XH, Zhang L, Alharbi KH (2023) Dynamic event-triggered control for discrete-time nonlinear Markov jump systems using policy iteration-based adaptive dynamic programming. Nonlinear Anal Hybrid Syst 49:101338

Torabi E, Ghobaei-Arani M, Shahidinejad A (2022) Data replica placement approaches in fog computing: a review. Clust Comput 25(5):3561–3589

Wang J, Jiang X, Zhao L, Zuo S, Chen X, Zhang L et al (2020) Lineage reprogramming of fibroblasts into induced cardiac progenitor cells by CRISPR/Cas9-based transcriptional activators. Acta Pharm Sin B 10(2):313–326

Wang H, Sha Y, Wang D, Nazari H (2022) A gene expression clustering method to extraction of cell-to-cell biological communication. Intel Artif 25(69):1–12

Wang T, Zhang L, Xu N, Alharbi KH (2023) Adaptive critic learning for approximate optimal event-triggered tracking control of nonlinear systems with prescribed performances. Int J Control. https://doi.org/10.1080/00207179.2023.2250880

Yue S, Niu B, Wang H, Zhang L, Ahmad AM (2023) Hierarchical sliding mode-based adaptive fuzzy control for uncertain switched under-actuated nonlinear systems with input saturation and dead-zone. Robot Intell Autom 43(5):523–536

Zhang D, Jiao L, Bai X, Wang S, Hou B (2018) A robust semi-supervised SVM via ensemble learning. Appl Soft Comput 65:632–643

Zhang L, Deng S, Zhang Y, Peng Q, Li H, Wang P et al (2020) Homotypic targeting delivery of siRNA with artificial cancer cells. Adv Healthc Mater 9(9):1900772

Zhang H, Zhao X, Wang H, Zong G, Xu N (2022a) Hierarchical sliding-mode surface-based adaptive actor–critic optimal control for switched nonlinear systems with unknown perturbation. IEEE Trans Neural Netw Learn Syst. https://doi.org/10.1109/TNNLS.2022.3183991

Zhang H, Zhao X, Zhang L, Niu B, Zong G, Xu N (2022b) Observer-based adaptive fuzzy hierarchical sliding mode control of uncertain under-actuated switched nonlinear systems with input quantization. Int J Robust Nonlinear Control 32(14):8163–8185

Zhang H, Zou Q, Ju Y, Song C, Chen D (2022c) Distance-based support vector machine to predict DNA N6-methyladenine modification. Curr Bioinform 17(5):473–482

Zhao Y, Niu B, Zong G, Xu N, Ahmad AM (2023a) Event-triggered optimal decentralized control for stochastic interconnected nonlinear systems via adaptive dynamic programming. Neurocomputing 539:126163

Zhao H, Wang H, Xu N, Zhao X, Sharaf S (2023b) Fuzzy approximation-based optimal consensus control for nonlinear multiagent systems via adaptive dynamic programming. Neurocomputing 553:126529