BMC Bioinformatics

RESEARCH Open Access

PWSC: a novel clustering method based on polynomial weight-adjusted sparse clustering for sparse biomedical data and its application in cancer subtyping

Xiaomeng Zhang[1], Hongtao Zhang[2], Zhihao Wang[2], Xiaofei Ma[2], Jiancheng Luo[2*] and Yingying Zhu[3*]

*Correspondence:
luojc@aiyi.link;
julianzyy@hotmail.com

[1] Department of Nephrology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430030, Hubei Province, China
[2] School of Mathematics and Statistics, Wuhan University, Wuhan 430070, Hubei Province, China
[3] Department of Oncology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430030, Hubei Province, China

Abstract

Background: Clustering analysis is widely used to interpret biomedical data and uncover new knowledge and patterns. However, conventional clustering methods are not effective when dealing with sparse biomedical data. To overcome this limitation, we propose a hierarchical clustering method called polynomial weight-adjusted sparse clustering (PWSC).

Results: The PWSC algorithm adjusts feature weights using a polynomial function, redefines the distances between samples, and performs hierarchical clustering analysis based on these adjusted distances. Additionally, we incorporate a consensus clustering approach to determine the optimal number of classifications. This consensus approach utilizes relative change in the cumulative distribution function to identify the best number of clusters, resulting in more stable clustering results. Leveraging the PWSC algorithm, we successfully classified a cohort of gastric cancer patients, enabling categorization of patients carrying different types of altered genes. Further evaluation using Entropy showed a significant improvement ($p = 2.905e{-}05$), while using the Calinski–Harabasz index demonstrates a remarkable 100% improvement in the quality of the best classification compared to conventional algorithms. Similarly, significantly increased entropy ($p = 0.0336$) and comparable CHI, were observed when classifying another colorectal cancer cohort with microbial abundance. The above attempts in cancer subtyping demonstrate that PWSC is highly applicable to different types of biomedical data. To facilitate its application, we have developed a user-friendly tool that implements the PWSC algorithm, which canbe accessed at http://pwsc.aiyimed.com/.

Conclusions: PWSC addresses the limitations of conventional approaches when clustering sparse biomedical data. By adjusting feature weights and employing consensus clustering, we achieve improved clustering results compared to conventional methods. The PWSC algorithm provides a valuable tool for researchers in the field, enabling more

accurate and stable clustering analysis. Its application can enhance our understanding of complex biological systems and contribute to advancements in various biomedical disciplines.

Keywords: Hierarchical clustering, Polynomial weight, Consensus clustering, Sparse biomedical data

Introduction

In biomedical data processing, cluster analysis is an essential tool that can be utilized to classify and predict various types of biological molecule data [1].By grouping similar data points together, cluster analysis forms clusters with distinct features that can differentiate different biological entities, including gene sequences, proteins, and more [2–5]. Through the use of cluster analysis, vast amounts of biological data can be organized and analyzed in a meaningful way, leading to more precise biomedical information [6]. Additionally, cluster analysis can also be employed to analyze the structure and function of biological systems, providing new approaches and perspectives for biomedical research [7].

Commonly used methods in cluster analysis include K-means clustering and hierarchical clustering [8, 9]. The former involves dividing data points into K groups, with the center of each group being the average value of all data points within it [10, 11]. The latter method involves gradually grouping data points based on their similarities, forming a tree-like structure [12, 13]. Additionally, there have been some clustering methods specifically designed for certain problems. For example, Kath Nicholls et al. proposed the Biclustering algorithm to address the issue that genes cluster differently in heterogeneous samples and cannot achieve effective clustering [14]. Juan Wang et al. improved the clustering quality of multi-cancer samples based on gene expression data by applying the graph regularized low-rank representation under symmetric and sparse constraints (sgLRR) method [15].

However, the data in the biomedical domain is often high-dimensional and sparse, primarily due to the complexity of multiple biomolecules, tissues, and organs in living organisms [16]. The high dimensionality and sparsity of biomedical data result in samples being sparsely distributed in a high-dimensional clustering space, making it challenging for conventional clustering methods to effectively capture similarities and affinities between samples [17, 18]. This can lead to issues such as overfitting or underfitting [19].

To address the problem that conventional clustering methods are difficult to handle due to the sparsity of biomedical data, we propose a new clustering algorithm. This algorithm recalculates the distances between samples by adjusting the weights of features, and performs clustering analysis based on this. At the same time, we use a consensus clustering method to select the optimal number of classifications, thus obtaining

the most stable clustering results. With this integrated approach, we have successfully achieved effective clustering of biomedical data with good classification results, avoiding the overfitting or oversimplification problems that can occur in conventional methods.

Methods

Polynomial weight-adjusted sparse clustering

We redefined the distances between samples based on the hierarchical clustering method [12, 13].

By reading the data, we can obtain the following sparse matrix, the rows of which represent the performance values of one of its features in the sample and the columns represent the performance values of different features of one of the samples.

$$D = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & & d_{2n} \\ \vdots & & \ddots & \vdots \\ d_{m1} & d_{m2} & \cdots & d_{mn} \end{bmatrix} \in R^{mn} \tag{1}$$

We process this sparse matrix and count the frequency of its features in each row in the sample to obtain $\{\eta_1, \eta_2 \ldots \eta_m\}$, and next, we build the polynomial:

$$P_n^i(\eta_i) = a_n \eta_i^n + a_{n-1} \eta_i^{n-1} + \cdots + a_1 \eta_i^1 + a_0, \tag{2}$$

With the help of that established polynomial, we re-establish the weights $\{\overline{\eta}_1, \overline{\eta}_2 \ldots \overline{\eta}_m\}$:

$$\overline{\eta}_i = \frac{P_n^i(\eta_i)}{\sum_{k=1}^{m} P_n^k(\eta_k)}, \tag{3}$$

We adjust the weights of the sparse matrix to obtain a correction matrix $\overline{D}$, expressed as follows:

$$\overline{D}(i, \cdot) = D(i, \cdot) \times \overline{\eta}_i, \tag{4}$$

$$\overline{D} = \begin{bmatrix} d_{11}\overline{\eta}_1 & d_{12}\overline{\eta}_1 & \cdots & d_{1n}\overline{\eta}_1 \\ d_{21}\overline{\eta}_2 & d_{22}\overline{\eta}_2 & & d_{2n}\overline{\eta}_2 \\ \vdots & & \ddots & \vdots \\ d_{m1}\overline{\eta}_m & d_{m2}\overline{\eta}_m & \cdots & d_{mn}\overline{\eta}_m \end{bmatrix}, \tag{5}$$

After that, we perform hierarchical clustering on the corrected sparse matrix using the method of sum of squares of differences, with the algorithm shown as follows:

Input: Sparse Matrix D

Output: The Result of The Clustering

1. $\eta = [\ ]$
2. $[m, n] = size(D)$
3. for $i = 1:$ m

$$\eta[i] = sum(D(i,) \neq 0)/n$$

end for.

4. $\eta = P(\eta)/sum(P(\eta))$
5. $D(i,\cdot) = D(i,\cdot) \times \eta[i]$
6. Create groups $G = \{G_1, G_2, \ldots, G_n\}$ by the column of the D, each column is divided into separate categories.
7. Set $c = 1$
8. While $size(G) == 1$

for $k = 1: size(G)$

for $h = k+1: size(G)$

$$\text{compute} \quad W_k = \sum\nolimits_i (G_k(i) - \bar{G}_k)^2$$

$$\text{compute} \quad SW_{k+h} = \sum\nolimits_i (G_k(i) - \bar{G}_{k+h})^2 + \sum\nolimits_j (G_h(j) - \bar{G}_{k+h})^2$$

end for

end for

Get $p, q \in argmin(SW_{k+h} - W_k - W_h)$

Create new group $G_{n+c} = \{G_p, G_q\}$

Refresh $G = G - G_p - G_q + G_{n+c}$, $c = c + 1$

end while

9. return G

Algorithm 1 PWSC (Polynomial Weight-adjusted Sparse Clustering)

Consensus clustering

The Monti consensus clustering algorithm is a well-known method for determining the number of clusters, K, in a dataset of N points [20, 21]. This algorithm involves resampling and clustering the data for each K, resulting in an $N \times N$ consensus matrix that indicates how often pairs of samples were clustered together. A perfectly stable matrix would contain only zeros and ones, indicating that all sample pairs either always clustered together or never did. By comparing the stability of the consensus matrices for different K values, the optimal K can be determined.

To be more precise, let $D = \{e^1, e^2, \ldots, e^N\}$ be the set of points to cluster, and let $D^1, D^2, \ldots, D^H$ be the H perturbed datasets resulting from resampling the original data. Let M^h be the $N \times N$ connectivity matrix obtained by clustering D^h, with entries defined as follows:

$$M^h(i,j) = \begin{cases} 1, & \text{if points i and j belong to the same cluster,} \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Let U^h be the $N \times N$ indicator matrix with (i,j) entry equal to 1 if points i and j are in the same perturbed dataset D^h, and 0 otherwise. This matrix is used to keep track of which samples were selected during each resampling iteration for the normalization step. Thconsensus matrix C for a given K is defined as the normalized sum of all connectivity matrices of all the perturbed datasets:

$$C(i,j) = \frac{\sum_{h=1}^{H} M^h(i,j)}{\sum_{h=1}^{H} U^h(i,j)} \tag{7}$$

In other words, the entry (i,j) in the consensus matrix is the number of times points i and j were clustered together, divided by the total number of times they were selected together. The consensus matrix is symmetric, and each element falls in the range [0,1]. A consensus matrix is calculated for each K value to be tested, and the stability of each matrix is assessed to determine the optimal K. One way to measure the stability of the Kth consensus matrix is by examining its cumulative distribution function (CDF) curve.

Determination of optimal classification

Şenbabaoğlu et al. discovered that the original delta K metric used in the Monti algorithm performed poorly when deciding on the optimal number of clusters, and proposed a superior metric for measuring the stability of consensus matrices using their CDF curves [21]. In the CDF curve of a consensus matrix, the lower left portion represents sample pairs that are rarely clustered together, while the upper right portion represents those that are almost always clustered together. The middle segment represents those with ambiguous assignments in different clustering runs. The proportion of ambiguous clustering (PAC) score measures the fraction of sample pairs with consensus indices falling in the interval $(u_1, u_2) \in [0, 1]$, where u_1 is a value close to 0 and u_2 is a value close to 1. A low value of PAC indicates a flat middle segment and a low rate of discordant assignments across permuted clustering runs. The optimal number of clusters can be inferred by selecting the K value that has the lowest PAC.

Acquisition of two testing datasets

Gene mutation data for gastric cancer

We obtained somatic gene mutation data of 437 gastric cancer patients through the Xena Browser (https://xenabrowser.net/datapages/) for The Cancer Genome Atlas (TCGA), which is a landmark cancer genomics program that molecularly characterized over 11,000 cases of primary cancer samples. Based on the work of Dechao Bu et al. [22], we removed 35 samples missing survival information and 71 samples with high tumor mutation burden (TMB) and screened out 69 genes that were actionable. Finally, we have constructed a matrix with dimensions of 69 rows and 331 columns. Each row represents a gene, and each column represents a patient. The non-zero elements in this matrix account for approximately 1.63%, clearly indicating that it is a sparse matrix.

Gut microbial data for colon cancer

Gut microbiota abundance and composition affect the occurrence and progression of colorectal cancer, which can be used for subtyping of colorectal cancer patients. We obtained the gut microbial data of 195 colon cancer patients through the The Cancer Microbiome Atlas (TCMA, https://tcma.pratt.duke.edu/), which is a collection of curated, decontaminated microbial compositions for multiple types of cancers. Finally, we have constructed a matrix with dimensions of 221 rows and 195 columns, where each row represents a gene, and each column represents a patient. The non-zero elements in this matrix account for approximately 3.96%, fully demonstrating its sparsity.

Assessing coefficients

Entropy

In cluster analysis, the concept of entropy is used to assess the stability and reliability of the clustering results [23–26]. When entropy is small, it means that most of the samples are clustered in one large cluster, while the others are grouped into a few small clusters [23]. In this case, the clustering results are less stable, as any point of perturbation or missing data may cause the samples that have been grouped into small clusters to be reallocated to the large clusters, resulting in large changes in the clustering results [27, 28]. In addition, the smaller number of samples in the small clusters results in the extracted features possibly lacking sufficient representation, reducing the reliability of the classification. Conversely, when the entropy value is large, it indicates that the number of samples included in each classification is relatively large, and therefore the classification results are more stable and more reliable. Therefore, entropy is widely used in cluster analysis to assess the quality and stability of clustering results [29].

We can calculate the entropy of the classification results for dataset D in this way [26]:

$$Ent(D) = \sum_{k} p_k log(p_k) \tag{8}$$

where p_k is the probability of the sample being classified into the Kth cluster.

Calinski–Harabasz index

The Calinski–Harabasz index (CHI) is an internal evaluation metric for cluster analysis, designed to measure the tightness and separation of clustering results [30–33]. It is calculated based on the intra-class and inter-class variance of the clusters, allowing assessment of the quality and effectiveness of the clustering. In calculating the intra-class variance, the metric takes into account the sum of the squares of the distances from each sample point to the center of the class to which it belongs, i.e., the intra-class sum of squares, with smaller values indicating tighter data points within the class. When calculating the between-class variance, the metric takes into account the sum of the squares of the distances from the centroid of each cluster to the center of the entire data set, i.e., the between-class sum of squares, with larger values indicating

greater distances between different clusters, i.e., better separation between clusters [34]. Therefore, the Calinski–Harabasz index is a metric for evaluating the quality of clustering based on the ratio of the intra-class sum of squares to the inter-class sum of squares, where a higher index value indicates better quality of clustering results.

We can calculate the Calinski–Harabasz index in this way [34]:

$$s = \frac{tr(B_k)}{tr(W_k)} \times \frac{n_E - k}{k - 1} \tag{9}$$

$$W_k = \sum_{q=1}^{k} \sum_{x \in G_q} (x - c_q)(x - c_q)^T \tag{10}$$

$$B_k = \sum_{q=1}^{k} n_q (c_q - c_E)(c_q - c_E)^T \tag{11}$$

where n_E is the number of training samples, k is the number of categories, B_k is the between-category covariance matrix, W_k is the within-category data covariance matrix, and $tr(\cdot)$ is the trace of the matrix.

Implementation

PWSC is available as a open source software package for the R programming framework. It relied on the R packages cluster, clusterSim, pheatmap, ConsensusClusterPlus, fpc, clv, clvalid. To facilitate its application, we have developed a user-friendly web tool that implements the PWSC algorithm, which is constructed using the Python Flask framework. It takes Nginx as a reverse proxy server to handle a large number of concurrent requests. AJAX dynamic data, REACT frontend framework, and Ant-design component library are used to create user-friendly layout and visualizations. The PWSC web server is now hosted on an elastic cloud server from the Aliyun Cloud running an Centos Linux system (7.9.2009 with 16 CPU and 32 GB memory. It can be accessed at http://pwsc.aiyimed.com/from any platform by using modern Web browsers (recommended but not limited to the latest version of Safari, Chrome and Firefox).

Result

Framework of PWSC

The Fig. 1 shows the procedure of PWSC. Data pre-processing is first performed to obtain a sparse matrix, which is used as input to the clustering algorithm. Then, a polynomial $P_n^i(\eta_i)$ is defined to calculate a correction matrix D, which is used to more accurately represent the degree of affinity between different samples. The correction matrix is clustered using the hierarchical clustering algorithm and the quality of the clustering results is assessed by means of CDF plots and consensus matrix heatmaps to select the best number of clusters. Finally, the clustering heatmap was redrawn and the occurrence of different genes in each cluster was counted to identify the most valuable genes in each cluster, while the clustering results were assessed using assessing coefficients such as Calinski–Harabasz index, entropy, etc.

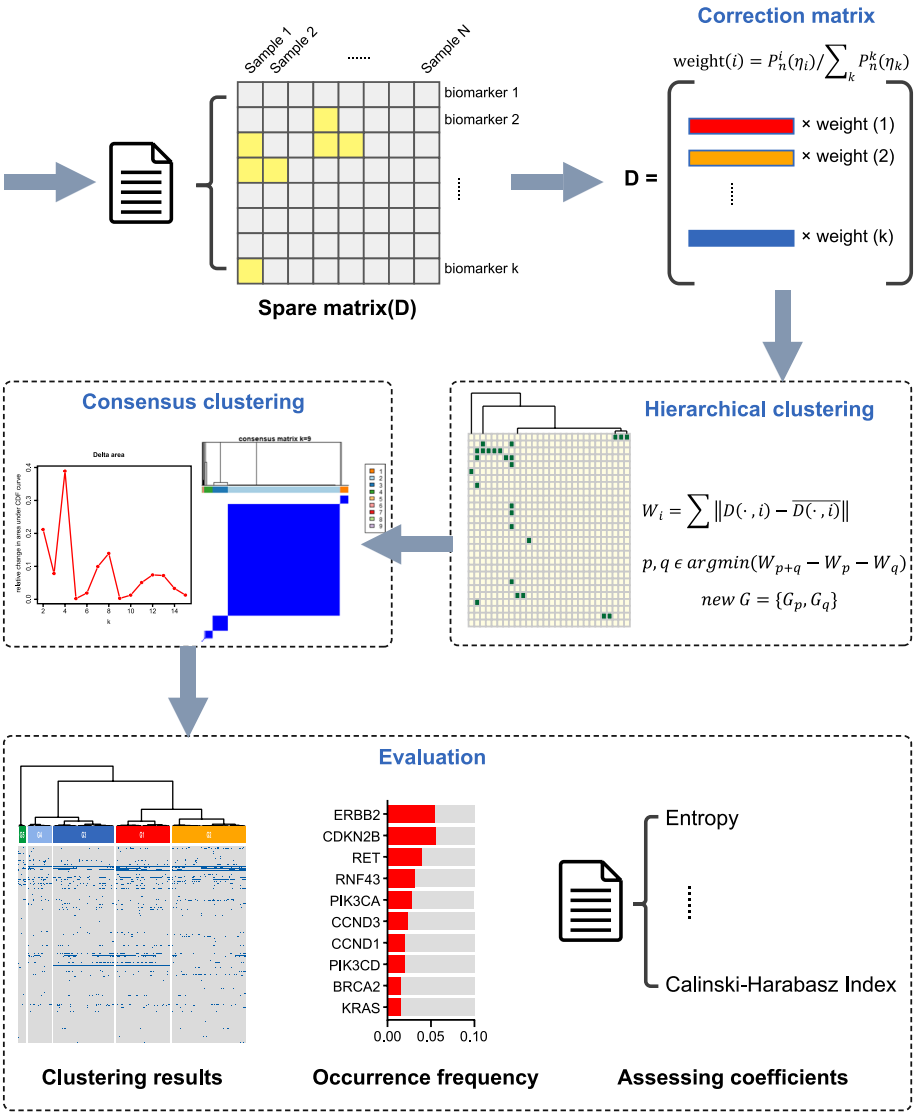

Fig. 1 Frame work of PWSC

Subtyping application on tumor mutational data

Clustering results and assessing coefficients

We applied the PWSC algorithm to perform cancer subtyping on a cohort of 331 gastric cancer patients. The orginal input sparse matrix contains the specific gene mutations carried by each patient. We set the weight function $P_n^i(\eta_i) = \eta_i^4, \overline{\eta}_i = \frac{\eta_i^4}{\sum_{k=1}^m \eta_k^4}$ brought into Algorithm 1 for clustering calculation. The clustering results obtained are shown below. By looking at the clustering heatmap (Fig. 2a), we found that some of the genes would be concentrated in a certain region in the clustering heatmap, which indicates that these genes have an important role in determining the clustering results, and they are likely to be an important basis for dominating our clustering.

In addition, we observed the entropy values and the results showed that the PWSC algorithm also had a more significant increase in entropy values compared to the

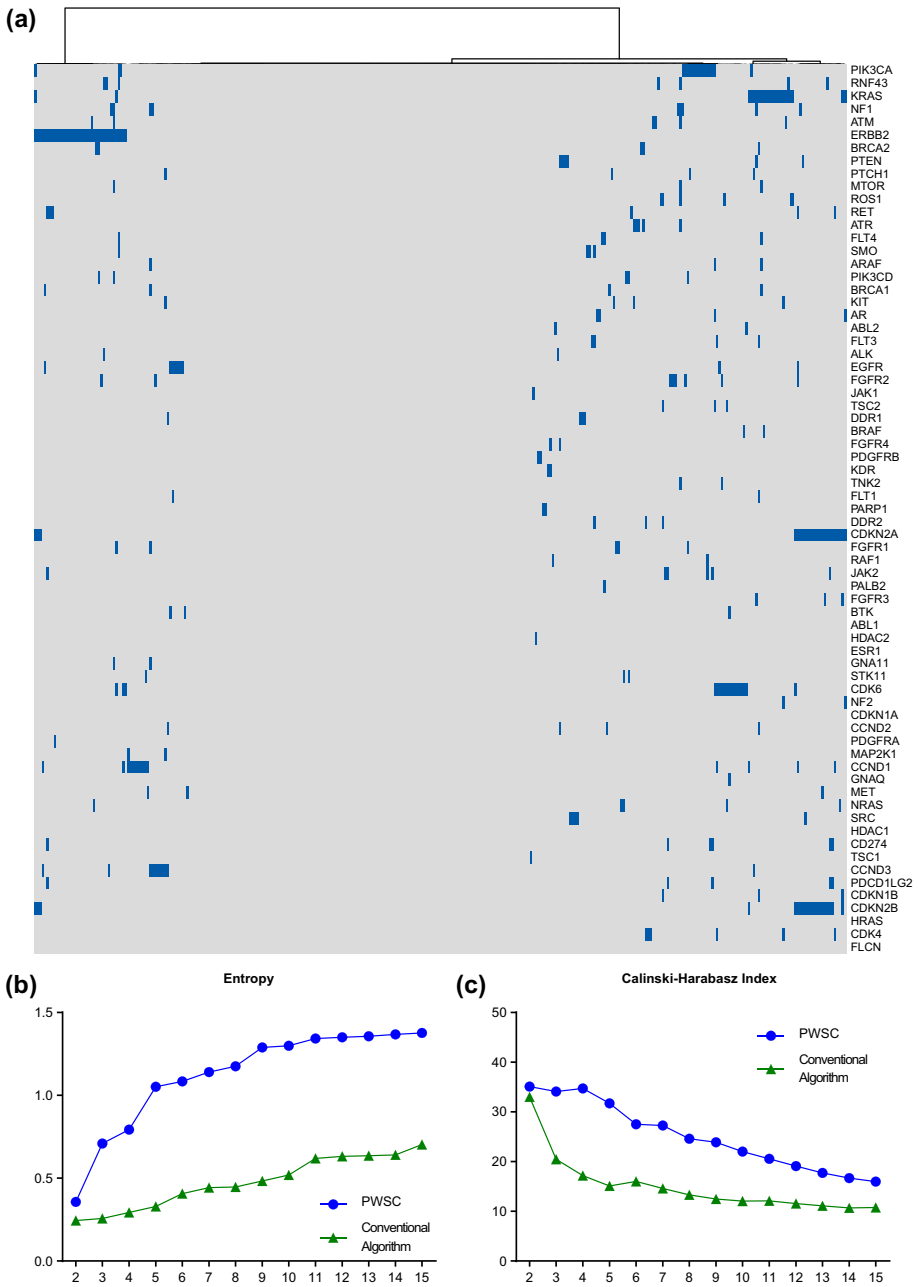

Fig. 2 Clustering results and assessing coefficients. **a** The clustering heatmap of biomedical data. **b** The Entropy of PWSC and conventional algorithm when k is from 2 to 15. **c** The CHI of PWSC and conventional algorithm when k is from 2 to 15

conventional algorithm of utilizing euclidean distance (Mann–Whitney U-test, W = 186, p value = 2.905e−05, Fig. 2b), which indicates that the clustering results of PWSC are more complex, stable and have better clustering results [35, 36].

Next, we calculated the Calinski–Harabasz index when the number of classifications k took on a range of values from 2 to 15. This is shown in Fig. 2c below. The analysis shows that the Calinski–Harabasz index values for the same number of classifications are significantly higher when using the PWSC algorithm for clustering compared to the

conventional algorithm (Mann–Whitney U test, $W = 192.5$, $p = 0.0004871$), which indicates that the PWSC algorithm can perform better clustering analysis and improve the accuracy and reliability of clustering analysis [35, 36].

Best clustering results

We combined the consensus clustering approach with the above methods to calculate the consensus CDF curve (Fig. 3a) and the consensus matrix (Fig. 3b). Upon analyzing the curve, we found that it peaked at $k = 4$, 8, and 13. Considering the cost of clustering, a larger number of classifications can lead to a less stable clustering result [37]. Therefore, we selected the 4-cluster classification as the optimal number of classifications based on the Calinski–Harabasz index, which reached its maximum value at $k = 4$ and was significantly higher than at $k = 8$ and $k = 13$.

Furthermore, the consensus matrix heatmap was very clear at $k = 4$, indicating that the probability of each sample being misclassified during the consensus clustering test was low, demonstrating the high stability of the clustering at this point [21]. Thus, selecting $k = 4$ as the optimal number of clusters is appropriate. A remarkable 100% improvement in the quality of the best classification were observed compared to conventional algorithms (Fig. 2c).

We divided the clustering results based on the distance between the individual clusters, dividing the samples into four groups and presenting the results as a clustering heatmap shown in Fig. 3c. We found that this result further validated our findings in the previous section: genes with a concentrated distribution were present in some of the groups and these genes played an important role in determining the grouping.

Finally, to investigate the practical implications of this classification, we counted the number of occurrences of genes in each group and identified those genes that were significantly different from other genes as most valuable genes. The analysis of Fig. 3d shows that ERBB2, KRAS, CDKN2B and CDKN2A are the most valuable genes in Group1, Group3 and Group4, respectively, which are all the hot oncogenes in gastric cancer. They have a significant difference in occurrence compared to other genes; while Group2 shows a mixed distribution of multiple genes, with multiple genes playing a role in the patient's disease [38–40]. In contrast, Group2 showed a mixed distribution of genes, with multiple genes playing a role in the disease.

Through the analysis of the most valuable genes, we can see that, on the one hand, PWSC can extract the genes that are dominant in the clustering, and through the extraction of these genes, we can have a deeper understanding of the disease, and also achieve more precise treatment according to the different gene expression of the patients [41–45]. The PWSC can help identify the most valuable biomarkers. These biomarkers are not only essential factors for classifying different groups, but also cover a wide scope of instances.

Subtyping application on tumor microbial data

The gut microbiome is a key player in the immunomodulatory and protumorigenic microenvironment during colorectal cancer (CRC), as different gut-derived microbes can induce tumor growth. Thus, it has been used for subtyping of colorectal cancer

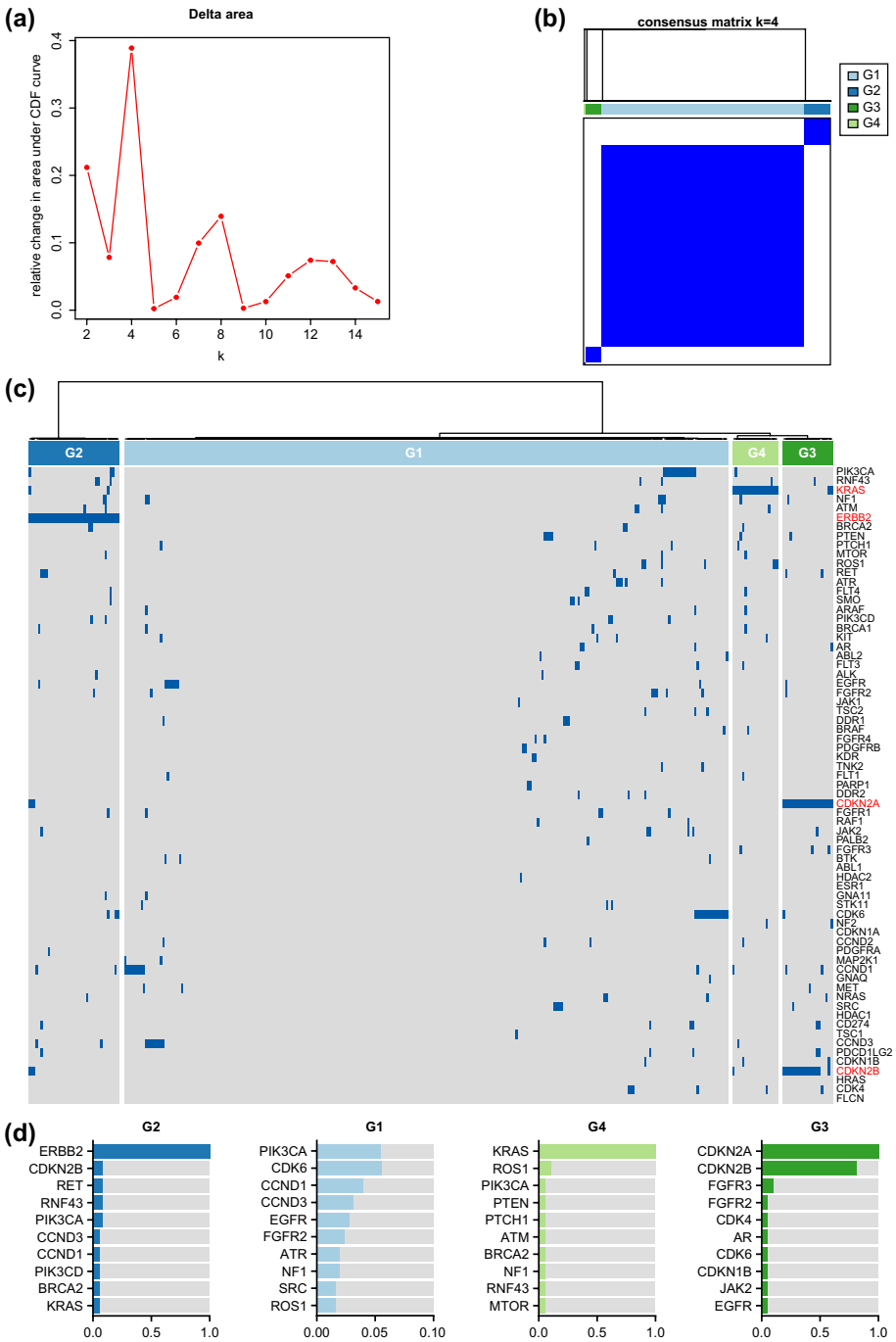

Fig. 3 Optimal clustering and occurrence of genes for the gastric cancer cohort. **a** The consensus clustering CDF curve when k is from 2 to 15. **b** The consensus matrix when k is 4. **c** The clustering heatmap of gene mutations for four groups. **d** The most valuable genes with high occurrence in each group

patients. We applied the PWSC algorithm on the original input sparse matrix, which contain the 221 microbial abundance for 195 patients, with 3.96% of non-zero elements.

Considering the relative change in area under CDF curve (Fig. 4a) and visiualization of consensus matrix (Fig. 4b) for each K, we selected the 5-cluster classification

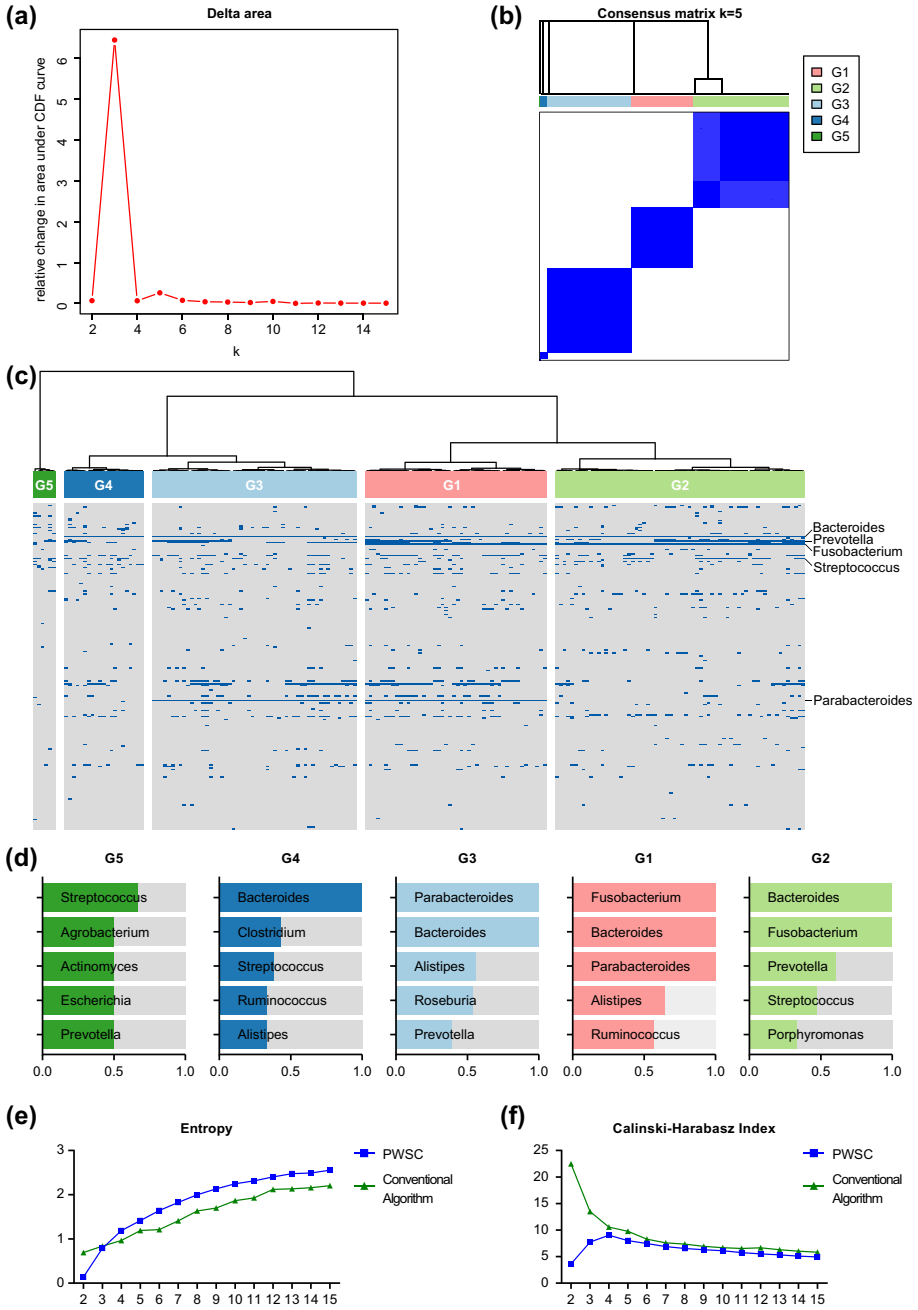

Fig. 4 Subtyping of colorectal cancer cohort using tumor microbial data. **a** The consensus clustering CDF curve when k is from 2 to 15. **b** The consensus matrix when k is 4. **c** The clustering heatmap of microbial data for five groups. **d** The most valuable biomarkers with high occurrence in each group. **e** The Entropy of PWSC and conventional algorithm when k is from 2 to 15. **f** The CHI of PWSC and conventional algorithm when k is from 2 to 15

as the optimal number of classifications. Under this classification, G1 contains 48, G2 contains 66, G3 contains 54, G4 contains 21, and G5 contains 6 samples (Fig. 4c). We have tallied the occurrence of microorganisms in each group and found that different groups are dominated by quite different bacteria. Bacteroides is a genus of bacteria

that naturally exists in the microbiota of the human gut. They are highly present in G1–G4, but not in G5. Fusobacteriums are highly present in G1, G2, but not in G3, G4, G5. Parabacteroides are highly present in G1, G3, while not in G2, G4, G5 (Fig. 4d).

Next, we calculated the entropy values and Calinski–Harabasz index when the number of classifications k took on a range of values from 2 to 15. A significantly increased Entropy ($p = 0.0336$) (Fig. 4e), as well as comparable CHI (Fig. 4f), were observed compared with the conventional algorithm. The above attempts in cancer subtyping demonstrate that PWSC is highly applicable to different types of biomedical data.

Online web service

To make it easier for users to utilize our accomplishments, we have created an online web service (Fig. 5). Users are required to input the data that needs to be clustered and choose the desired assessing coefficients. We will then generate the corresponding consensus clustering results and assessing coefficients. Based on these results, users should enter the number of classifications that fulfill their specific requirements. Lastly, we will produce the clustering result heatmap and the corresponding most valuable biomarkers for each group.

The PWSC website utilizes the NGINX service as its power source. The methods have been refactored into RESTful APIs, and AJAX is used to dynamically refresh data on the web page. To enhance accessibility, we have adopted the REACT frontend framework, along with the Ant-design component library to create user-friendly layouts and display data tables. In addition, Echarts is utilized for interactive chart display.

You can upload the biomedical data that needs to be clustered with the assistance of PWSC. Then, select the desired test coefficients and input the number of categories to

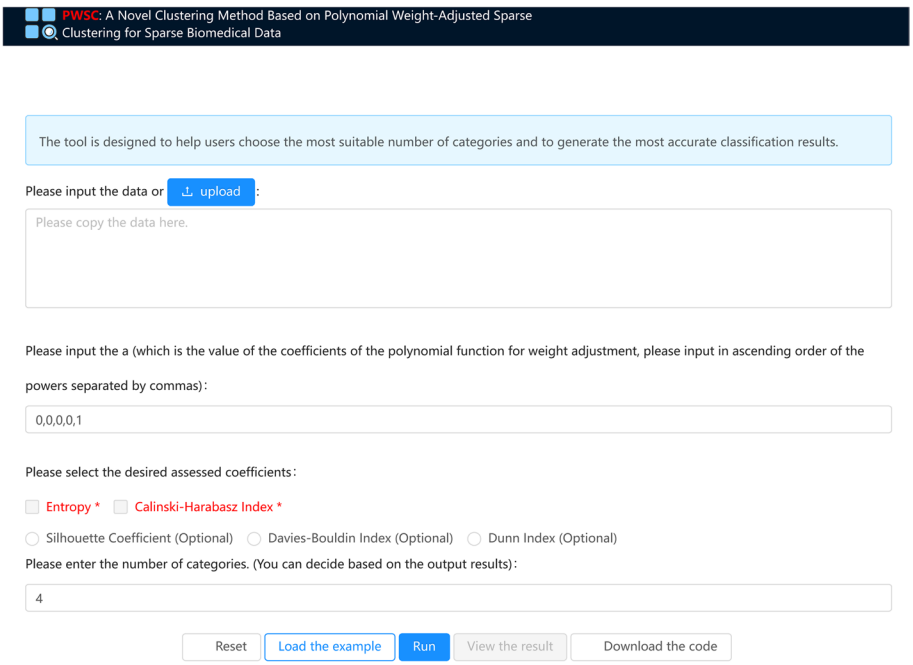

Fig. 5 Online web service interface presentation

be tested. Finally, click on 'Run' to obtain the clustering results. Alternatively, you can click on 'Load the example' to view an example. The data in the example is the experimental data of this article, and all experimental results can be downloaded after running the example. Additionally, for your convenience in further research and study, you can click on 'Load the code' to access the program's code. This website can be accessed through http://pwsc.aiyimed.com/.

Discussion

When clustering sparse matrices, we observed that the sparsity of the data matrix leads to a less compact distribution of sample points in the high-dimensional space characterized by the data. It becomes challenging to find a consensus on an effective partition criterion to assign some samples to a specific category. In hierarchical clustering, the principle of clustering is to assign samples with close "relatedness" to the same category and those with distant "relatedness" to different categories, thus studying the relationship between samples. However, when samples exhibit sparse distribution in space, it means that the "relatedness" between samples is distant, making it difficult to find a closure to partition samples with certain features from the original samples [46]. This results in conventional clustering methods being unable to classify samples in biomedical research when the data matrix is sparse, making it impossible to conduct targeted studies on samples.

The article is based on this point and proposes the use of PWSC to solve this problem. The advantage of PWSC lies in its ability to better separate data by modifying the data matrix through the establishment of polynomial weights. Compared with conventional algorithms and methods that directly use gene occurrence frequency as weights, PWSC can make samples of the same class more compact and those of different classes more dispersed, thus more accurately reflecting the "relatedness" between data. This data modification method can improve the accuracy and stability of clustering results.

On the one hand, PWSC uses polynomial functions for weight modification, avoiding the potential problem of numerical overflow when using the Soft-max function as weights. By manually adjusting the polynomial function, the function value and adjusted weight value can always be kept within a suitable range, avoiding numerical overflow and underflow and ensuring the stability and reliability of the algorithm [47, 48]. On the other hand, compared with the exponential function (Soft-max function), the polynomial function has a slower growth rate. This means that when using the polynomial function to process weights, the differences between data points will not be overly magnified, effectively avoiding the problem of data points being overly stretched and resulting in poor clustering results. This weight modification method can more reasonably and controllably handle differences between data points, thereby improving the effectiveness and stability of clustering.

At the same time, we used the method of consensus clustering to determine the optimal number of clusters. On the one hand, this can help us avoid the influence of subjectivity and subjective bias on the selection of the number of clusters, thus determining the number of clusters more objectively and accurately. On the other hand, it can better reflect the stability and consistency of clustering results under different numbers of clusters. By considering the results of consensus clustering, interference

caused by fluctuations in clustering results due to noise or randomness can be avoided in the selection of the number of clusters.

From the classification results, we can see that the classification results of the PWSC algorithm can help researchers identify prominent genes in different groups, known as "most valuable biomarkers". By statistically analyzing the occurrence of genes, significant differences between genes in different groups can be determined, thus studying the role of different genes in the process of causing disease in humans and helping us to better understand the mechanism of disease occurrence [49]. This can help doctors make more accurate treatment decisions based on the patient's gene expression. Additionally, the PWSC algorithm can assist doctors in distinguishing between disease caused by gene mixing and disease caused by a single gene. By comparing with other groups, PWSC can help doctors better understand the role of different genes in the patient's disease process, thus more accurately classifying and diagnosing patients.

In summary, the PWSC algorithm has the advantages of improving the accuracy, stability, and reliability of clustering analysis through the polynomial weighting correction method. Additionally, the PWSC algorithm can help researchers gain a deeper understanding of the mechanism of disease occurrence and assist doctors in making diagnoses.

Conclusions

PWSC algorithm provides an effective solution for handling sparse biomedical data. By utilizing the PWSC algorithm, researchers can accurately classify genes and identify "star genes" that play a significant role in disease mechanisms. This helps us gain a deeper understanding of the underlying causes of disease and provides valuable insights for medical professionals to make more precise treatment decisions based on a patient's gene expression patterns. With the ability to handle sparse biomedical data and identify important genes, the PWSC algorithm holds great potential in advancing our understanding of disease and improving patient outcomes.

Acknowledgements
This work was supported by the National Natural Science Foundation of China (grant number: 82270748) and Hubei Natural Science Foundation (grant number: 2022CFB219).

Author contributions
MZ and HZ: Methodology, validation, investigation, resources, writing—original draft preparation. ZW: Investigation, resources. XM: Software, visualization. JL and YZ: Conceptualization, supervision, project administration, writing—review and editing. All authors read and approved the final manuscript.

Funding
None.

Availability of data and materials
Genomic data of gastric cancer patients are downloaded from the Xena Browser with identifier of TCGA-STAD (https://xenabrowser.net/datapages/).

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 7 June 2023 Accepted: 4 December 2023
Published online: 21 December 2023

References

1. Xu R, Wunsch DC. Clustering algorithms in biomedical research: a review. IEEE Rev Biomed Eng. 2010;3:120–54.
2. Segal E, Koller D. Probabilistic hierarchical clustering for biological data. In: Proceedings of the sixth annual international conference on Computational biology. Washington: Association for Computing Machinery; 2002, pp. 273–280.
3. Hanage WP, Fraser C, Spratt BG. Sequences, sequence clusters and bacterial species. Philos Trans R Soc Lond B Biol Sci. 2006;361(1475):1917–27.
4. Nascimento MCV, Toledo FMB, de Carvalho ACPLF. Investigation of a new GRASP-based clustering algorithm applied to biological data. Comput Oper Res. 2010;37(8):1381–8.
5. Wei D, et al. A novel hierarchical clustering algorithm for gene sequences. BMC Bioinform. 2012;13(1):174.
6. Huang X, et al. Enhanced clustering of biomedical documents using ensemble non-negative matrix factorization. Inf Sci. 2011;181(11):2293–302.
7. Yin L, et al. Nutritional features-based clustering analysis as a feasible approach for early identification of malnutrition in patients with cancer. Eur J Clin Nutr. 2021;75(8):1291–301.
8. Xu D, Tian Y. A comprehensive survey of clustering algorithms. Ann Data Sci. 2015;2(2):165–93.
9. Xu R, Wunsch D. Survey of clustering algorithms. IEEE Trans Neural Networks. 2005;16(3):645–78.
10. Hartigan JA, Wong MA. Algorithm AS 136: a K-means clustering algorithm. J R Stat Soc Ser C Appl Stat. 1979;28(1):100–8.
11. Likas A, Vlassis N, Verbeek JJ. The global k-means clustering algorithm. Pattern Recognit. 2003;36(2):451–61.
12. Murtagh F, Contreras P. Algorithms for hierarchical clustering: an overview. Wiley Interdiscipl Rev Data Min Knowl Discov. 2012;2(1):86–97.
13. Rani Y, Rohil H. A study of hierarchical clustering algorithm. Int J Inf Comput Technol. 2013;2:113.
14. Nicholls K, Wallace C. Comparison of sparse biclustering algorithms for gene expression datasets. Brief Bioinform. 2021;22(6):bbab140.
15. Wang J, et al. Multi-cancer samples clustering via graph regularized low-rank representation method under sparse and symmetric constraints. BMC Bioinform. 2019;20(Suppl 22):718.
16. Zitnik M, et al. Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. Inf Fusion. 2019;50:71–91.
17. Abdulrauf Sharifai G, Zainol Z. Feature selection for high-dimensional and imbalanced biomedical data based on robust correlation based redundancy and binary grasshopper optimization algorithm. Genes (Basel). 2020;11(7):717.
18. Pes B. Learning from high-dimensional biomedical datasets: the issue of class imbalance. IEEE Access. 2020;8:13527–40.
19. Kuss O. Global goodness-of-fit tests in logistic regression with sparse data. Stat Med. 2002;21(24):3789–801.
20. Monti S, et al. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Mach Learn. 2003;52:91–118.
21. Senbabaoglu Y, Michailidis G, Li JZ. Critical limitations of consensus clustering in class discovery. Sci Rep. 2014;4:6207.
22. Cheng Y, et al. Genomic and transcriptomic profiling indicates the prognosis significance of mutational signature for TMB-high subtype in Chinese patients with gastric cancer. J Adv Res. 2022;51:121–34.
23. Jenssen R, et al. Clustering using Renyi's entropy. In: Proceedings of the international joint conference on neural networks, 2003. 2003.
24. Larson RR. Introduction to information retrieval. J Am Soc Inf Sci Technol. 2010;61(4):852–3.
25. Li T, Ma S, Ogihara M, Entropy-based criterion in categorical clustering. In: Proceedings of the twenty-first international conference on Machine learning. Banff: Association for Computing Machinery; 2004, p. 68.
26. Wehrl A. General properties of entropy. Rev Mod Phys. 1978;50(2):221–60.
27. Nigam K, Lafferty J, McCallum A. Using maximum entropy for text classification. In: IJCAI-99 workshop on machine learning for information filtering. Stockholm, Sweden; 1999.
28. Osborne, M. Using maximum entropy for sentence extraction. In: Proceedings of the ACL-02 workshop on automatic summarization. 2002.
29. Cuzzolin, F. Generalised max entropy classifiers. In: Belief functions: theory and applications: 5th international conference, BELIEF 2018, Compiègne, France, September 17–21, 2018, proceedings 5. Springer; 2018.
30. Ali MFBM, et al. A comprehensive 3-phase framework for determining the customer's product usage in a food supply chain. Mathematics. 2023;11(5):1085.
31. Łukasik S, et al. Clustering using flower pollination algorithm and Calinski–Harabasz index. In: 2016 IEEE congress on evolutionary computation (CEC). 2016.
32. Maulik U, Bandyopadhyay S. Performance evaluation of some clustering algorithms and validity indices. IEEE Trans Pattern Anal Mach Intell. 2002;24(12):1650–4.
33. Wang X, Xu Y. An improved index for clustering validation based on Silhouette index and Calinski–Harabasz index. IOP Conf Ser Mater Sci Eng. 2019;569(5):052024.
34. Caliński T, Harabasz J. A dendrite method for cluster analysis. Commun Stat Theory Methods. 1974;3(1):1–27.
35. McKnight, P.E. and J. Najab, Mann-Whitney U test. In: The Corsini encyclopedia of psychology. 2010, p. 1.
36. Nachar N. The Mann–Whitney U: a test for assessing whether two independent samples come from the same distribution. Tutor Quant Methods Psychol. 2008;4(1):13–20.
37. Ünlü R, Xanthopoulos P. Estimating the number of clusters in a dataset via consensus clustering. Expert Syst Appl. 2019;125:33–9.

38. Yu D, Hung M-C. Overexpression of ErbB2 in cancer and ErbB2-targeting strategies. Oncogene. 2000;19(53):6115–21.
39. Walker GJ, et al. Virtually 100% of melanoma cell lines harbor alterations at the DNA level within CDKN2A, CDKN2B, or one of their downstream targets. Genes Chromosomes Cancer. 1998;22(2):157–63.
40. Liu P, Wang Y, Li X. Targeting the untargetable KRAS in cancer therapy. Acta Pharm Sin B. 2019;9(5):871–9.
41. Tahara E. Genetic pathways of two types of gastric cancer. IARC Sci Publ. 2004;157:327–49.
42. Qu Y, Dang S, Hou P. Gene methylation in gastric cancer. Clin Chim Acta. 2013;424:53–65.
43. Petrovchich I, Ford JM. Genetic predisposition to gastric cancer. Semin Oncol. 2016;43(5):554–9.
44. McLean MH, El-Omar EM. Genetics of gastric cancer. Nat Rev Gastroenterol Hepatol. 2014;11(11):664–74.
45. Lynch HT, et al. Gastric cancer: new genetic developments. J Surg Oncol. 2005;90(3):114–33.
46. Steinbach M, Ertöz L, Kumar V. The challenges of clustering high dimensional data. In: Wille LT, editor. New directions in statistical physics: econophysics, bioinformatics, and pattern recognition. Berlin: Springer; 2004. p. 273–309.
47. Song B, et al. Robustness learning via inference-softmax cross entropy in misaligned distribution of image. Mathematics. 2022;10(19):3716.
48. Duan K, et al. Multi-category classification by soft-max combination of binary classifiers. Multiple Classif Syst. 2003;2709:125–34.
49. Lyman GH, et al. Impact of a 21-gene RT-PCR assay on treatment decisions in early-stage breast cancer. Cancer. 2007;109(6):1011–8.

Publisher's Note