A Novel Biomarker Selection Method Combining Graph Neural Network and Gene Relationships Applied to Microarray Data

Wei Li (liwei@cse.neu.edu.cn)
Northeastern University
Weidong Xie
Northeastern University
Shoujia Zhang
Northeastern University
Linjie Wang
Northeastern University
Jinzhu Yang
Northeastern University
Dazhe Zhao
Northeastern University

Research Article

Keywords: Graph neural networ, Feature selection, Biomarker, Spectral clustering

Posted Date: April 18th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1557151/v1

RESEARCH

A Novel Biomarker Selection Method Combining Graph Neural Network and Gene Relationships Applied to Microarray Data

Weidong Xie[1], Wei Li[2*†], Shoujia Zhang[1], Linjie Wang[1], Jinzhu Yang[2] and Dazhe Zhao[1†]

*Correspondence:
liwei@cse.neu.edu.cn
[2]Key Laboratory of Intelligent Computing in Medical Image (MIIC), Northeastern University, Ministry of Education, China
Full list of author information is available at the end of the article
†Co-corresponding authors

Abstract

Background: The discovery of critical biomarkers is significant for clinical diagnosis, drug research and development. Researchers usually obtain biomarkers from microarray data, which comes from the dimensional curse. Feature selection in machine learning is usually used to solve this problem. However, most methods do not fully consider feature dependence, especially the real pathway relationship of genes.

Results: Experimental results show that the proposed method is superior to classical algorithms and advanced methods in feature number and accuracy, and the selected features have more significance.

Method: This paper proposes a feature selection method based on a graph neural network. A graph structure and real feature dependence combined with Pearson correlation coefficient is used in our method to construct graph structure model, the information propagation and aggregation method of graph neural network is used to characterize node information, and the spectral clustering method is used to filter redundant features. Then, the feature ranking aggregation model using eight feature evaluation methods acts on each clustering sub-cluster for different feature selection.

Conclusion: The proposed method can effectively remove redundant features. The algorithm's output has high stability and classification accuracy, which can potentially select potential biomarkers.

Keywords: Graph neural networ; Feature selection; Biomarker; Spectral clustering

Background

With the development and maturity of microarray technology, researchers can obtain a large number of gene expression values at once by DNA microarray technology, and these data can be used to analyze critical genes for disease diagnosis, drug development, and other tasks [1]. The difficulty of microarray data analysis is the large feature dimensionality and small sample size. Machine learning based feature selection methods can be used to select essential features from high dimensional data to solve this problem.

In the feature selection task, the purpose is to find a set of feature subsets of original features, which are highly redundant with the original features and significantly correlate with the label information. Feature selection is different from feature extraction, which obtains a set of representation information of low-dimensional space

from high-dimensional space. Feature extraction can not explain the meaning of the representation of low-dimensional space and can not be well connected with downstream tasks. Traditional feature selection tasks can be divided into filter, wrapper, and embedded methods.

The filter method does not rely on the machine learning model and solves the best feature ranking through the statistical calculation mode. It has high speed but low accuracy. The common filter methods mainly contain t-test [2], chi-squared test [3], and maximum information efficiency (MIC) [4], fisher score [5]. The wrapper method relies on a specific feature evaluator or machine learning model. It constantly looks for the best feature combination through the heuristic search algorithm. According to the return value of the evaluator as the fitness function, it can find the optimal feature subset under the feature classifier. However, local optimization and high time complexity are the disadvantages of the packaging method. Common wrapper methods incorporates Stability Selection [6], Recursive Feature Elimination (RFE) [7], Genetic Algorithm (GA) [8] ,Artificial Bee Colony (ABC) [9], Ant Colony Optimization (ACO) [10] and Particle Swarm Optimization (PSO) [11]. The embedded method skillfully combines the feature selection process with the machine learning model, and outputs the feature subset through the weight parameters of the model. The effect of this method depends on the machine learning model, and not all models support the output of weight parameters. The common embedded method comprises Decision Tree(DT) [12], Random Forest Algorithm(RF) [13], and Liner Regression(LR) [14].

The hybrid feature selection algorithm combines the advantages of the above three algorithms and is the mainstream algorithm for feature selection tasks. For example, researchers can combine the filtering method and packaging method to realize the rapid filtering of invalid features in the filtering method and reduce the time complexity of the packaging method to design an efficient packaging method for further selection and optimization of features. These methods have been widely used and reported and have achieved excellent results on mainstream microarray data sets. Salem et al. proposed a new method for finding biomarkers from gene microarray data that combines Information Gain (IG) and a Standard Genetic Algorithm (SGA) for feature selection, called IG/SGA [15]. Jain et al. combine Correlation-based Feature Selection (CFS) with an improved Binary Particle Swarm Optimization(iBPOS) to propose a two-stage hybrid feature selection method for biomarker selection of microarray data [16]. Moradi et al. proposed a hybrid feature selection method based on Particle Swarm Optimization (PSO) algorithm [17].

However, most current hybrid feature selection methods assume that the samples are independent and identically distributed or infer the relationship between the samples based on the data model. The most significant feature of DNA microarray data is the close relationship between genes, that is, the characteristics of the data, such as gene pathway, physical interaction, and other information. However, this prior knowledge information is ignored by most algorithms. Previous studies have emphasized and proved the importance of considering the feature interaction relationship in the feature selection task. For example, the method based on the probability graph model uses information entropy and conditional probability to infer the interaction between features. However, the interaction between genes does

not follow the probability distribution, and the natural pathway and coexpression relationship cannot be considered, in which Markov blank (MB) [18]. Although methods based on mutual information, maximum correlation, and minimum redundancy emphasize feature interaction, simple mathematical models cannot infer complex gene interaction relationships.

The graph model adopts the form of nodes and edges, which can well represent the interaction relationship between non-independent and identically distributed data and is well applied to non-Euclidean structure data. Mainstream platforms for analyzing gene or protein interactions, such as GeneMANIA and STRING, are represented by graph structure [19]. The research of [20] and [21] is devoted to finding the characteristic genes of microarray data. Based on the graph structure data, the regularization technology is used to realize the feature selection in the graph structure. However, these methods do not capture the high-order connectivity of graph structure data and do not apply the prior knowledge in the existing database. Graph has been mathematically applied to social science [22, 23], protein interaction network [24], knowledge graph [25], and other research fields [26]. Graph neural network makes each node have global information representation through information dissemination and aggregation between nodes and fully excavates the feature interaction relationship and high-order connectivity information. However, this method has not been applied to the microarray data feature selection task.

This paper proposes an innovative biomar selection method for microarray data. Our previous research has shown that graph neural networks can be a good guide to biomarker selection [27]. In the proposed method, the graph structure is used to establish the interactive information between genes, and each node represents a feature. The numerical correlation of genes and the correlation existing in prior knowledge are considered as the edges between nodes in the graph. The proposed method uses graph neural network technology to spread and aggregate the information of each node and predicts the possible feature interaction through connection prediction technology. Then, in order to delete redundancy features, spectral clustering technology is applied to the graph. Each clustering subgraph is regarded as a feature subset with high self redundancy and low external redundancy. Each feature subset is used as a candidate feature subset to select the final marker gene. In order to ensure the reliability of the results, we use eight different feature evaluators to evaluate the candidate feature subset respectively, input the results into a reliable sorting fusion algorithm, and finally output the feature subset.

The main contributions of this study include the following contents.

(1) A comprehensive feature selection framework is proposed, which makes full use of feature interaction and prior knowledge, and uses graph neural network technology to capture the high-order connectivity between features. The framework integrates eight feature evaluation algorithms and uses a reliable ranking fusion method to obtain the final feature subset.

(2) Graph neural network is innovatively proposed to calculate and analyze the feature correlation. We integrate the prior knowledge and the propagation and aggregation of messages. Each feature node can provide global information representation and makes full use of the existing knowledge and technology to ensure the reliability of the feature correlation

(3) Link prediction has been added to the graph structure to solve the problem of difficulty in applying spectral clustering to sparse plots.

The rest of this paper is organized as follows: the The Results part shows the experimental results, the Method part briefly introduces the overall framework of the proposed algorithm and describes each module in detail. Finally the Discussion and Conclusion parts summarizes the full text.

Results

This section describes the proposed feature selection processing flow, firstly, how to pre-process the data and the initial filtering of features using T-test, followed by our improved binary difference evolution algorithm flow. Finally we present the improvement strategies for the scaling factor and fitness function of the binary difference evolution algorithm.

Cluster quantitative analysis

In the experimental process, firstly, a T-test was performed on all features, 100 groups of features were retained, the gene relationship matrix is obtained from GeneMANIA, and the Pearson correlation coefficient is calculated. The graph structure is established using the 100 groups of features, and the feature selection is carried out according to the proposed method. The number of clusters is set to be 1 to 50, respectively. After feature sorting and fusion, the feature subset is taken as the final feature selection result, SVM is taken as the classifier, and the average *Acc* and *Auc* of 10 fold cross-validation are taken as the final evaluation index.

Figure 1 shows the relationship between the number of clusters and Acc and Auc on four different datasets, respectively. It can be found that as the number of clusters increases, redundant features are continuously introduced into the feature subset, resulting in a decrease in evaluation indicators. A smaller number of features (the number of clustered subclusters) can remove redundant features well, and when the number of features is very small, although a higher Acc index can be obtained, the Auc index may be lower, and the result stability is poor.

Comparison with traditional algorithms

This section compares the proposed method with traditional machine learning feature selection methods, shown as Table 1, including linear regression model (liner), L1 regularization (lasso), random forest (RF), L2 regularization (ridge), feature recursive elimination (RFE) and decision tree (DT), It can be seen that the proposed method is superior to all classical machine learning algorithms when only one feature is adopted, which proves the superiority of the proposed method.

Comparison with advanced methods

This section compares the proposed method with the advanced feature selection method, and the detailed results are shown in Table 2. It can be found from the table that the proposed method is still better than the advanced feature selection algorithm when the number of features is small. Unlike most hybrid methods, which require high time complexity, the proposed method only needs one aggregation calculation of graph neural network and a simple feature evaluation method

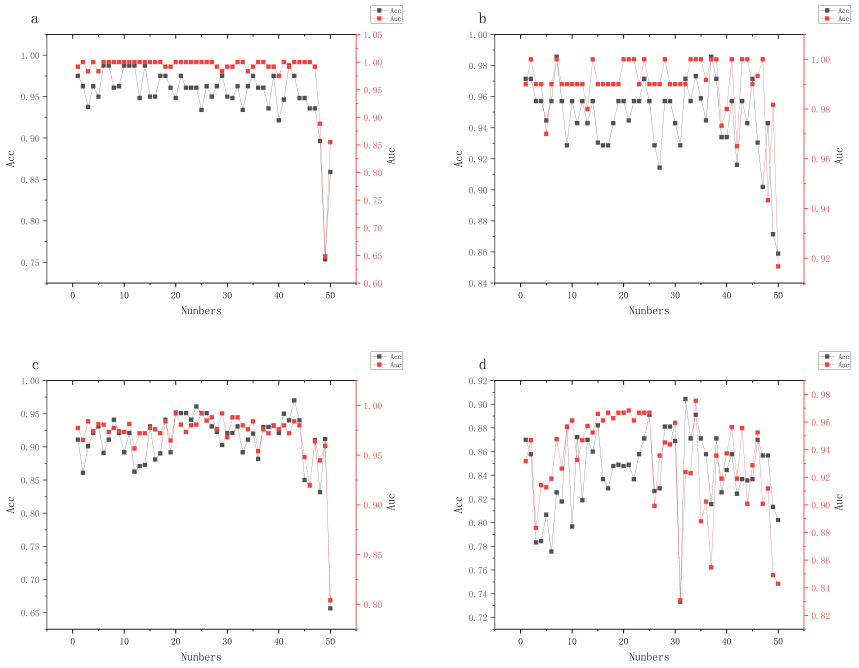

Figure 1 The relationship between the number of features (number of clusters) and classification accuracy and *Auc*. Figure a represents DLBCL data set, figure b represents leukemia data set, figure c represents prostate data set, and figure d represents ALL_4 data set.

Table 1 The proposed method is compared with the classical method. The number of features used is indicated in parentheses.

Method	DLBCL	Leukemia	Prostate	ALL4
Liner (3)	0.820	0.943	0.842	0.849
Lasso (3)	0.909	0.936	0.921	0.830
RF (3)	0.923	0.915	0.932	0.806
Ridge (3)	0.820	0.943	0.842	0.849
RFE (3)	0.753	0.971	0.843	0.839
DT (3)	0.766	0.915	0.756	0.826
Ours (1)	**0.975**	**0.971**	**0.911**	**0.870**

to achieve efficient feature selection. In addition, considering the prior knowledge and feature dependence, the features selected by the proposed method have better interpretability and lower redundancy.

Biomarker analysis

In this section, we further analyze the selected features of the proposed method. Figure 2 (a) (b) (c) (d) shows the results of the four data sets, respectively. The distribution of the four most essential probe ids selected by the proposed method in positive and negative samples is plotted in each data set. It can be seen that the features selected by the proposed method can effectively distinguish positive and negative samples, which have high diagnostic significance.

Discussion

It can be seen from the experimental results that using a priori knowledge as feature dependency can significantly improve the classification accuracy when the number of features is low. However, the feature relationships obtained from GeneMANIA

Table 2 Comparison between the proposed method and the advanced method.

Datasets	Papers	Year	Features	Acc
DLBCL	Agarwalla et al.[28]	2018	15	0.900
DLBCL	Medjahed et al.[29]	2017	15	0.894
DLBCL	Wang et al. [30]	2017	15	0.809
DLBCL	Apolloni et al. [31]	2016	15	0.929
DLBCL	Wang et al.[32]	2015	15	0.936
DLBCL	Maulik et al.[33]	2013	15	0.918
DLBCL	Yu et al.[27]	2021	15	0.946
DLBCL	**Ours**	2022	**1**	**0.975**
Leukemia	Lu et al.[34]	2019	9	0.952
Leukemia	Sun et al.[35]	2018	3	0.927
Leukemia	Wang et al.[30]	2017	8.3	0.961
Leukemia	Tumuluru et al.[36]	2017	/	0.946
Leukemia	**Ours**	2022	**2**	**0.971**
Prostate	Samson et al.[37]	2021	3	0.830
Prostate	Khani et al.[38]	2020	5	0.922
Prostate	Musheer et al.[39]	2019	4	0.763
Prostate	Theera et al.[40]	2018	5	0.874
Prostate	Paredes et al.[41]	2017	24.32	0.928
Prostate	Gunavathi et al.[42]	2014	10	0.927
Prostate	**Ours**	2022	**5**	**0.931**

are rich and include some predicted relationships, which are of great significance to feature selection, especially biomarkers. However, these dependencies are rarely used in current methods. Therefore, we will focus on the actual feature dependency in our future work.

In addition, it can be found from the results in figure 1 that with the increase of the number of features, the evaluation index will not improve significantly but will show a downward trend. This is because more redundant features are added to confuse the classification model, reflecting the importance of feature engineering. Therefore, it is essfinding feature dependencies and effectively removing redundant features is essential.

Existing research works basically adopt simple classification models, such as SVM and decision tree. The reason is that most algorithms need to apply the classifier indicators for feature evaluation repeatedly. Complex models will lead to high time complexity. However, graph structure allows us to directly use graph structure to build a classification model and use feature dependency to make classification decisions. In future research, we will focus on constructing a classifier using a graph structure.

Conclusion

This paper proposes a biomarker selection algorithm based on a graph neural network. This method effectively uses the dependence between features and integrates a priori knowledge to select features together. The algorithm removes redundant features by clusting and uses eight feature evaluators to achieve accurate and efficient feature selection. The results show that the integration and prediction of the natural interaction between genes can effectively improve the accuracy and interpretability of the results. In addition, we also analyze the relationship between the number of features and classification accuracy and prove the effectiveness and reliability of the features selected by the proposed method.

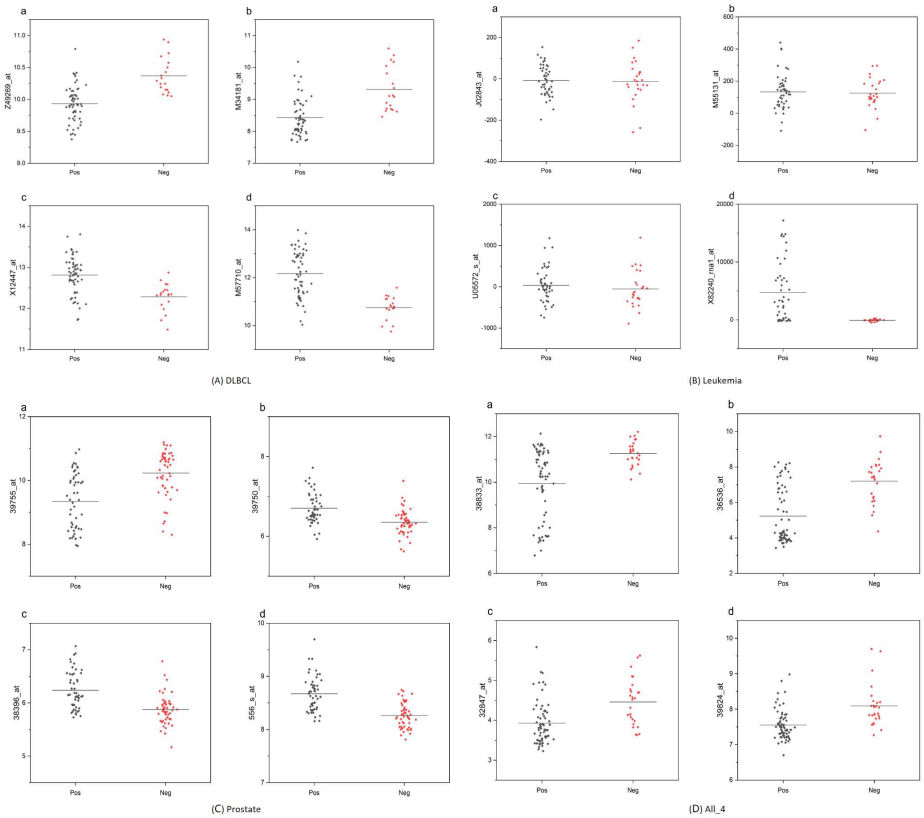

Figure 2 The distribution of biomarkers selected by the proposed method in positive and negative samples. Figure A represents DLBCL data set, Figure B represents leukemia data set, Figure C represents prostate data set, and Figure D represents ALL_4 data set.

Method

Dataset

Four DNA microarray datasets were used in this paper, namely DLBCL, Leukemia, Prostate and ALL_4, the details of these datasets are shown in Table 3. ALL_4 is a leukemia dataset containing 26 samples of Acute Myeloid Leukemia (AML) and 67 samples of Acute Lymphocytic Leukemia (ALL); each sample contains 12625 genes. Prostate is a dataset containing 52 tumor samples and 50 normal samples; each sample contains 12625 genetic information. DLBCL is a lymphoma data set, including 58 samples of diffuse large B-cell lymphoma (DLBCL) and 19 samples of follicular lymphoma (FL). Each sample includes 7129 gene information. Datasets and GPL files can be downloaded from https://github.com/xwdshiwo/BioFSDatasets.

Table 3 The dataset used in this paper, Ur means Unbalance rate.

Dataset	Samples	Pos	Neg	Features	Ur
DLBCL	77	58	19	7129	3.05
Leukemia	72	47	25	7129	1.88
Prostate	102	52	50	12625	1.04
ALL_4	93	26	67	12625	0.38

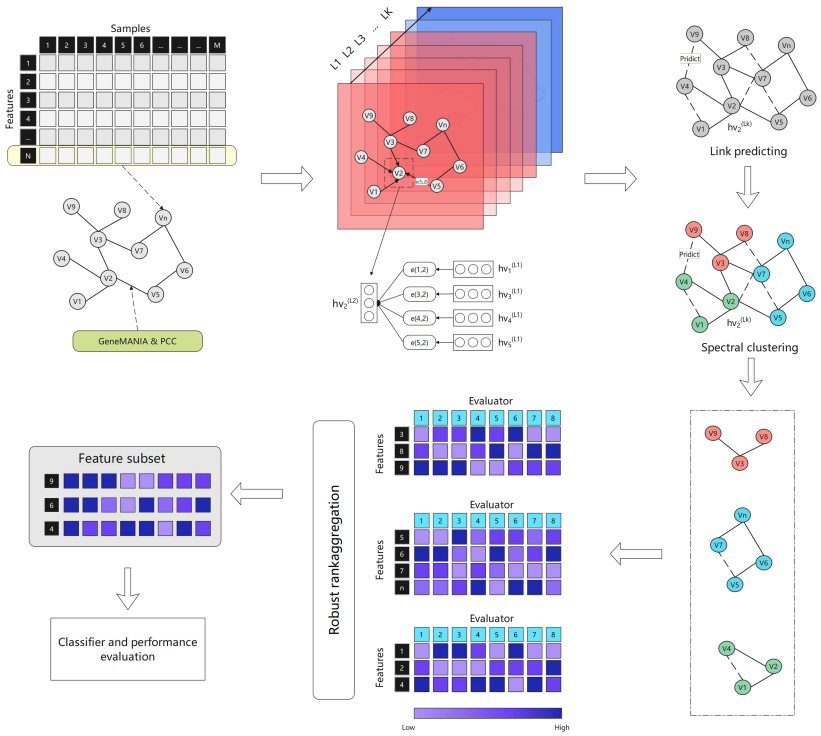

Figure 3 The overall framework of the proposed approach: The gene relationship data is obtained from GeneMANIA, the expression of each gene in positive and negative samples is embedded as node information, and the gene relationship data and Pearson correlation coefficient are embedded as edges after passing through a layer of softmax function. The graph neural networks' information dissemination and aggregation process is carried out. The dependency relationship is predicted by the link prediction method, and spectral clustering is carried out to delete redundant features. The feature of each subgraph is evaluated, eight kinds of evaluators are used, the ranking information is aggregated by the robust ranking method, and the feature subset is finally output.

The proposed framework of our method

The feature selection framework designed in this paper is shown in Figure 3. The first step of the algorithm is to construct the graph structure. The characteristic information from the microarray data is used as the initial embedding representation of the node, and the physical interaction information from GeneMANIA and the Pearson correlation coefficient of the node are used as the edge information of the node after passing through a layer of softmax function. Then, we use the information propagation and aggregation function to embed the nodes to enrich the node information deeply. Then we construct positive samples by randomly deleting the head and tail links of known link nodes and construct negative samples by randomly adding some links. We realize the link prediction of the edge by training a loss function and cluster on the graph after link prediction to delete redundant features. In each clustering subgraph, eight feature evaluators are used to evaluate the feature weight, and the RRA method is used to sort the feature weight comprehensively. Finally, the final feature subset is generated, and the classification model is established to analyze further and evaluate the feature subset.

Graph structure establishment

The sample-set in microarray data is defined as $S = \{S_1, S_2, \ldots, S_M\}$, in which a n-dimensional feature vector $S_i = \left\{F_1^i, F_2^i, \ldots, F_N^i\right\}$, represents each sample. Each feature is taken as node v in the graph. The physical correlation between two features i and j obtained from GeneMANIA is expressed as $e_{(vi,vj)}^{Gm}$. The Pearson correlation coefficient between two nodes with physical correlation is expressed as $e_{(vi,vj)}^{Pe}$. The correlation of the two nodes is calculated by the function shown in Equation 1 and used as edge $e_{(vi,vj)}$.

$$e_{(vi,vj)} = \text{MEAN}\left(e_{(vi,vj)}^{Pe} + e_{(vi,vj)}^{Gm}\right) \tag{1}$$

Then we can get graph $G = \{V, E\}$, where V represents the set of all nodes and E represents the set of all edges. The initial embedding of each node is expressed as the eigenvector $h_{vi}^0 = \left[S_1^{Fi}, S_2^{Fi}, \ldots, S_M^{Fi}\right]$ composed of the eigenvalues on each sample. In this way, we get the graph structure representation composed of the original microarray expression matrix and a priori knowledge information.

Information propagation and aggregation

Information propagation is one of the essential components of a graph neural network. Its purpose is to make each node have the feature vector representing the global information to better carry out the following task. In the process of message propagation, firstly, as described in the structure establishment part of the figure, initialize the eigenvector $h_{vi}^0 = \left[S_1^{Fi}, S_2^{Fi}, \ldots, S_M^{Fi}\right]$ of each node and define $N_{(vi)}$ to represent the first-order neighborhood of node vi. then the aggregation operation is shown in Equation 2 to obtain the state vector of the next layer of the node.

$$h_{vi}^K = \sum_{vj \in N_{(vi)}} \left(h_{vj}^{K-1} * e_{(vi,vj)}\right) / \text{Num}\left(N_{(vi)}\right) \tag{2}$$

where K represents the number of layers of the current graph neural network, and Num$(*)$ represents the number of first-order neighborhood nodes. We believe that when the difference of eigenvectors after two aggregations is less than the given threshold ε, the current graph reaches a stable state, the next layer of propagation will not be carried out.

In the information aggregation stage, each node splices the current layer's state vector with the previous layer's state vector and obtains the final state vector representation of the current layer through the nonlinear activation layer, as shown in Equation 3.

$$h_{vi}^K = \sigma\left(COUNCAT\left(h_{vi}^{K-1}, h_{vi}^K\right)\right) \tag{3}$$

where σ is the nonlinear activation function, representing the vector splicing operation. Then, the normalized representation of the node vector is carried out as shown in Equation 4, and the k-th layer state vector of the node is updated.

$$h_{v_i}^K \leftarrow h_{v_i}^K / \left\| h_{v_i}^K \right\|_2, v_i \in v \tag{4}$$

In the experiment, the above process is repeated until the difference between k layer and $k-1$ layer state vectors of all nodes is less than the given threshold ε, then stop the iteration and record the number of iteration layers L, and finally get the L-layer state vector of all nodes.

Link prediction

The purpose of link prediction is to predict the hidden relationship between two nodes, taking advantage of feature correlation and node high-order connectivity to prepare for further analysis. Feature selection uses the hidden state information of the node for prediction. After the information dissemination and aggregation of the graph neural network, the node has a state vector representing the global information, which can better carry out the prediction task.

In the process of link prediction, we first need to build positive and negative samples. Taking node *vi* as an example, we break any head and tail links connected to node *vi* in graph G, and randomly take *vi* as the central node to sample several new edges e_{new}. if $e_{new} \in E$, it will be marked as positive samples, otherwise it will be marked as negative samples. Then we record the similarity between vi as the central node and all the nodes connected to the new edge (taking node *vr* as an example, it is a node connected to node *vi* through e_{new}). The calculation method is shown in Equation 5.

$$\operatorname{sim}(v_j, v_r) = \frac{\sum_{\varphi=1}^{\pi} z_{v_j}^{\varphi} \times \sum_{\varphi=1}^{\pi} z_{v_r}^{\varphi}}{\sqrt{\sum_{\varphi=1}^{w} \left(z_{v_j}^{\varphi}\right)^2} \times \sqrt{\sum_{\varphi=1}^{w} \left(z_{v_r}^{\varphi}\right)^2}} \tag{5}$$

where $z_{v_j}^{\varphi}$ represents the value of eigenvector v_j, and w represents the eigenvector dimension. We set the positive sample set as Pos and the negative sample set as Neg and establish the loss function as shown in Equation 6.

$$L = MEAN_{(v_j, v_r) \in \mathrm{Pos}} \left[-\log\left(\sigma\left(\operatorname{sim}(v_j, v_r)\right)\right) - \sum_{(\overline{v_j}, \overline{v_r}) \in \mathrm{Neg}} \log\left(\sigma\left(\operatorname{sim}(\overline{v_j}, \overline{v_r})\right)\right) \right] \tag{6}$$

where L represents the loss value of the loss function, $(v_j, v_r) \in$ Pos represents the edge of any group of positive data samples,$(\bar{v}_j, \bar{v}_r) \in$ Neg represents the edge of any group of positive data samples, σ is a nonlinear activation function. The random gradient descent algorithm is used to train the model, and the loss value L in training is retained. When the difference between the loss values of the two training is less than ε, the training is stopped. At the same time, we calculate the mean predictive rank (MRR) of each prediction graph in training. The calculation

method is shown in Equation 7, and select the optimal graph as the final result according to the MRR. In this way, we get graph G^*, which has more prosperous relational attributes than graph G.

$$MRR = \frac{1}{\varepsilon}\sum_{\tau=1}^{\varepsilon}\frac{1}{\text{rank}_\tau} \quad \tau = 1, 2, \cdots, \varepsilon \tag{7}$$

MRR represents the average reciprocal rank, and rank represents the rank number of the scores from highest to lowest when the τ-th edge in the positive sample set scores the corresponding ε-th edge in the negative sample set.

Spectral clustering and feature ranking

After obtaining the new gene relationship graph G^*, we can use its prosperous gene relationship to cluster redundant features and find the feature gene with the most abundant information in each sub-cluster. We use spectral clustering to cluster features. The process is as follows:

Define all nodes in the new gene graph G^* as E, that is, $E = (e_1, e_2, \cdots, e_\zeta)$, ζ represents the total number of nodes in the gene graph G^*. Use Equation 8 to calculate the similarity between any two nodes (v_i, v_j), and $W_(v_i, v_j)$ will form an ζ dimensional similarity matrix W,

$$w_{vi,vj} = \sum_{vi=1,vj=1}^{\zeta}\exp\frac{-\left\|e_{vi} - e_{vj}\right\|^2}{2\Omega^2}, \quad e_{vi}, e_{vj} \in E \tag{8}$$

Ω uses to control the neighborhood width of nodes. Calculate the sum of all elements in each row of the similarity matrix w to obtain $\{d_1, d_2, \cdots, d_\eta, \cdots d_\zeta\}$, where d_ζ represents the sum of all elements in the row, and use $\{d_1, d_2, \cdots, d_\eta, \cdots d_\zeta\}$ to construct the diagonal matrix with D dimension, then use Equation 9 to calculate laplacian matrix L_{reym} .

$$L_{\text{reym}} = D^{\frac{-1}{2}}(D - W)D^{\frac{-1}{2}} \tag{9}$$

Calculate the eigenvalues of the Laplace matrix L_{reym} , and sort the eigenvalues in the order from small to large. According to the number μ of clustering clusters, take the first μ eigenvalues and calculate the corresponding eigenvector $\{\chi_1, \chi_2, \cdots, \chi_\mu\}$. use the μ eigenvectors $\{\chi_1, \chi_2, \cdots, \chi_\mu\}$ to form the matrix U of rows and columns, that is, $U = \{\chi_1, \chi_2, \cdots, \chi_\mu\}$.

The K-means clustering algorithm is used to cluster the eigenvectors in each row of matrix U to obtain $\{C_1, C_2, \cdots, C_v, \cdots, C_\mu\}$, where C_v represents the cluster clustered by the eigenvectors in row V. According to the obtained cluster $\{C_1, C_2, \cdots, C_v, \cdots, C_\mu\}$, all nodes in the new gene relationship graph $G*$ are divided into μ groups to obtain μ subgraph, which is recorded as Equation 10.

$$G^* = [G_1, G_2, \cdots, G_v, \cdots, G_\mu] = \left[\left(v_1', \varepsilon_1'\right), \left(v_2', \varepsilon_2'\right), \cdots, \left(v_v', \varepsilon_v'\right), \cdots, \left(v_\mu', \varepsilon_\mu'\right)\right] \tag{10}$$

where G_v represents the v subgraph, the v subgraph represents (v_v', ε_v'), v_v' represents all node sets in the subgraph G_v, and EB represents all edges in the subgraph G_v.

There are eight kinds of evaluator RRA ranking fusion in each subfigure. RRA is a widely used feature ranking fusion method, which can synthesize the results of multiple evaluators and output the best feature subset. Usually, the clusters can be selected according to downstream tasks or determined according to elbow rules. In the Results part, we analyzed the possible number of K.

Abbreviations
GNN: Graph Neural Network
RFE: Recursive Feature Elimination
GA: Genetic Algorithm
KNN: K-Nearest Neighbor
PSO: Particle Swarm Optimization
DT: Decision Tree
RF: Random Forest
LR: Lasso Regression
MIN: Mutual Information Maximization
SGA: Standard Genetic Algorithm
CFS: Correlation-based Feature Selection
MB: Markov Blank

Acknowledgements
Not applicable.

Funding
This work is the results of the research project funded by National key research and development program, China (2021YFC2701003), the Fundamental Research Funds for the Central Universities (N2016006).

Authors' contributions
WDX proposed experimental ideas, evaluated experimental data, and drafted manuscripts. WL and SJZ designs experimental procedures collects data, and assists in manuscript writing. LJW proposed the overall structure of the article and supplemented the experimental chart. JZY and DZZ revises the manuscript and evaluates the data. All authors read and approved the final manuscript.

Corresponding author
Correspondence to Wei Li and Dazhe Zhao.

Consent for publication
Not applicable.

Data Availability Statement
The public data set used in our experiment is from the GEO (Gene Expression Omnibus) database, which can be obtained through the following website:
https://www.ncbi.nlm.nih.gov/geo.
https://github.com/xwdshiwo/BioFSDatasets

Declarations
Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Author details
[1]School of Computer Science and Engineering, Northeastern University, China. [2]Key Laboratory of Intelligent Computing in Medical Image (MIIC), Northeastern University, Ministry of Education, China.

References

1. Kavitha, K., Prakasan, A., Dhrishya, P.: Score-based feature selection of gene expression data for cancer classification. In: 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), pp. 261–266 (2020). IEEE
2. Zhou, N., Wang, L.: A modified t-test feature selection method and its application on the hapmap genotype data. Genomics, proteomics & bioinformatics **5**(3-4), 242–249 (2007)
3. Liu, H., Setiono, R.: Chi2: Feature selection and discretization of numeric attributes, 388–391 (1995). IEEE
4. Lin, C., Miller, T., Dligach, D., Plenge, R., Karlson, E., Savova, G.: Maximal information coefficient for feature selection for clinical document classification. In: ICML Workshop on Machine Learning for Clinical Data. Edingburgh, UK (2012)
5. He, X., Cai, D., Niyogi, P.: Laplacian score for feature selection., vol. 18 (2005)
6. Haury, A.-C., Mordelet, F., Vera-Licona, P., Vert, J.-P.: Tigress: trustful inference of gene regulation using stability selection. BMC systems biology **6**(1), 145 (2012)
7. Yan, K., Zhang, D.: Feature selection and analysis on correlated gas sensor data with recursive feature elimination. Sensors and Actuators B: Chemical **212**, 353–363 (2015)
8. Li, X., Xiao, N., Claramunt, C., Lin, H.: Initialization strategies to enhancing the performance of genetic algorithms for the p-median problem. Computers & Industrial Engineering **61**(4), 1024–1034 (2011)
9. Karaboga, D.: An idea based on honey bee swarm for numerical optimization, technical report - tr06. Technical Report, Erciyes University (2005)
10. Dorigo, M., Maniezzo, V.: Ant system: optimization by a colony of cooperating agents. IEEE Trans. on SMC-Part B **26**(1), 29 (1996)
11. A, M.F.T., B, Y.C.L., C, M.S., D, G.G.: A particle swarm optimization algorithm for makespan and total flowtime minimization in the permutation flowshop sequencing problem - sciencedirect. European Journal of Operational Research **177**(3), 1930–1947 (2007)
12. Stein, G., Chen, B., Wu, A.S., Hua, K.A.: Decision tree classifier for network intrusion detection with ga-based feature selection. In: Proceedings of the 43rd Annual Southeast Regional conference-Volume 2, pp. 136–141 (2005)
13. Chen, K.-H., Wang, K.-J., Tsai, M.-L., Wang, K.-M., Adrian, A.M., Cheng, W.-C., Yang, T.-S., Teng, N.-C., Tan, K.-P., Chang, K.-S.: Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm. BMC bioinformatics **15**(1), 49 (2014)
14. Fonti, V., Belitser, E.: Feature selection using lasso. VU Amsterdam Research Paper in Business Analytics **30**, 1–25 (2017)
15. Salem, H., Attiya, G., El-Fishawy, N.: Classification of human cancer diseases by gene expression profiles. Applied Soft Computing **50**, 124–134 (2017)
16. Jain, I., Jain, V.K., Jain, R.: Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification. Applied Soft Computing **62**, 203–215 (2018)
17. Moradi, P., Gholampour, M.: A hybrid particle swarm optimization for feature subset selection by integrating a novel local search strategy. Applied Soft Computing **43**, 117–130 (2016)
18. Jl, A., Iyc, B., Chj, C.: An efficient multivariate feature ranking method for gene selection in high-dimensional microarray data. Expert Systems with Applications **166** (2020)
19. Warde-Farley, D., Donaldson, S.L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., Franz, M., Grouios, C., Kazi, F., Lopes, C.T., *et al.*: The genemania prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic acids research **38**(suppl_2), 214–220 (2010)
20. Gu, Q., Han, J.: Towards feature selection in network. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management. CIKM '11, pp. 1175–1184. Association for Computing Machinery, New York, NY, USA (2011). doi:10.1145/2063576.2063746. https://doi.org/10.1145/2063576.2063746
21. Tang, J., Liu, H.: Feature Selection with Linked Data in Social Media, pp. 118–128. doi:10.1137/1.9781611972825.11. https://epubs.siam.org/doi/pdf/10.1137/1.9781611972825.11. https://epubs.siam.org/doi/abs/10.1137/1.9781611972825.11
22. Monti, F., Bronstein, M., Bresson, X.: Geometric matrix completion with recurrent multi-graph neural networks. In: Advances in Neural Information Processing Systems, pp. 3697–3707 (2017)
23. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
24. Fout, A., Byrd, J., Shariat, B., Ben-Hur, A.: Protein interface prediction using graph convolutional networks. In: Advances in Neural Information Processing Systems, pp. 6530–6539 (2017)
25. Hamaguchi, T., Oiwa, H., Shimbo, M., Matsumoto, Y.: Knowledge transfer for out-of-knowledge-base entities: A graph neural network approach. arXiv preprint arXiv:1706.05674 (2017)
26. Khalil, E., Dai, H., Zhang, Y., Dilkina, B., Song, L.: Learning combinatorial optimization algorithms over graphs. In: Advances in Neural Information Processing Systems, pp. 6348–6358 (2017)
27. Yu, K., Xie, W., Wang, L., Zhang, S., Li, W.: Determination of biomarkers from microarray data using graph neural network and spectral clustering. Scientific reports **11**(1), 1–11 (2021)
28. Agarwalla, P., Mukhopadhyay, S.: Bi-stage hierarchical selection of pathway genes for cancer progression using a swarm based computational approach. Applied Soft Computing **62**, 230–250 (2017)
29. Medjahed, S.A., Saadi, T.A., Benyettou, A., Ouali, M.: Kernel-based learning and feature selection analysis for cancer diagnosis. Applied Soft Computing **51**, 39–48 (2016)
30. Wang, A., An, N., Yang, J., Chen, G., Li, L., Alterovitz, G.: Wrapper-based gene selection with markov blanket. Computers in Biology & Medicine **81**(Complete), 11–23 (2017)

31. Apolloni, J., Leguizamón, G., Alba, E.: Two hybrid wrapper-filter feature selection algorithms applied to high-dimensional microarray experiments. Applied Soft Computing **38**, 922–932 (2016)
32. Wang, A., An, N., Chen, G., Li, L., Alterovitz, G.: Accelerating wrapper-based feature selection with k-nearest-neighbor. Knowledge-Based Systems **83**(jul.), 81–91 (2015)
33. Lu, H., Chen, J., Yan, K., Jin, Q., Xue, Y., Gao, Z.: A hybrid feature selection algorithm for gene expression data classification. Neurocomputing **256**(sep.20), 56–62 (2016)
34. Lin, S., Xz, A., Yq, C., Jx, A., Sz, A.: Feature selection using neighborhood entropy-based uncertainty measures for gene expression data classification. Information Sciences **502**, 18–41 (2019)
35. Sun, L., Zhang, X.Y., Qian, Y.H., Xu, J.C., Zhang, S.G., Tian, Y.: Joint neighborhood entropy-based gene selection method with fisher score for tumor classification. Applied Intelligence **49** (2018)
36. Tumuluru, P., Ravi, B.: Goa-based dbn: Grasshopper optimization algorithm-based deep belief neural networks for cancer classification. International Journal of Applied Engineering Research **12**, 14218–14231 (2017)
37. P, S.A.B., Annavarapu, C.S.R., Dara, S.: Clustering-based hybrid feature selection approach for high dimensional microarray data. Chemometrics and Intelligent Laboratory Systems **213**, 104305 (2021). doi:10.1016/j.chemolab.2021.104305
38. Khani, E., Mahmoodian, H.: Phase diagram and ridge logistic regression in stable gene selection. Biocybernetics and Biomedical Engineering **40**(3) (2020)
39. Musheer, R.A., Verma, C.K., Srivastava, N.: Novel machine learning approach for classification of high-dimensional microarray data. Soft Computing A Fusion of Foundations Methodologies & Applications (2019)
40. Jinthanasatian, P., Auephanwiriyakul, S., Theera-Umpon, N.: Microarray data classification using neuro-fuzzy classifier with firefly algorithm. In: 2017 IEEE Symposium Series on Computational Intelligence (SSCI) (2018)
41. Alarcón-Paredes, A., Alonso, G.A., Cabrera, E., Cuevas-Valencia, R.: Simultaneous gene selection and weighting in nearest neighbor classifier for gene expression data. In: International Conference on Bioinformatics and Biomedical Engineering (2017)
42. Gunavathi, C., Premalatha, K.: Performance analysis of genetic algorithm with knn and svm for feature selection in tumor classification (2014)