2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 978-1-6654-6819-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/BIBM55620.2022.9994959

Feature Selection for Microarray Data via Community Detection Fusing Multiple Gene Relation Networks Information

1st Shoujia Zhang
College of Computer Science and Engineering
Northeastern University
Shenyang, Chain
2001868@stu.neu.edu.cn

2th Wei Li*
Key Laboratory of Intelligent Computing in Medical Image
Northeastern University
Shenyang, Chain
liwei@cse.neu.edu.cn

3nd Weidong Xie
College of Computer Science and Engineering
Northeastern University
Shenyang, Chain
1910638@stu.neu.edu.cn

4rd Linjie Wang
College of Computer Science and Engineering
Northeastern University
Shenyang, Chain
1971660@stu.neu.edu.cn

***Abstract*—In recent decades, the rapid development of gene sequencing and computer technology has increased the growth of high-dimensional microarray data. Some machine learning methods have been successfully applied to it to help classify cancer. In most cases, high dimensionality and the small sample size of microarray data restricted the performance of cancer classification. T his p roblem u sually i s s olved b y s ome feature selection methods. However, most of them neglect the exploitation of relations among genes. This paper proposes a novel feature selection method by fusing multiple gene relation network information based on community detection (MGRCD). The proposed method divides all genes into different communities. Then, the genes most associated with cancer classification a re s elected from each community. The proposed method satisfies b oth maximum relevances gene with cancer and minimum redundancy among genes for the selected optimal feature subset. The experiment results show that the proposed gene selection method can effectively improve classification performance.**

***Index Terms*—Feature Selection, Community Detection, Gene Relation Networks, Graph Neural Networks, Microarray Data**

I. Introduction

Gene microarray data is widely used to analyze cancer or other genetic diseases, typically including thousands of genes and a small sample size. Not all genes are associated with diseases; instead, only a small percentage of genes are helpful for cancer classification [1]. M eanwhile, medical researchers are eager to use machine learning algorithms to help them quickly identify disease-related genes. Therefore, efficient g ene s election f rom m icroarray i s a n e ssential and urgent issue [2].

Feature selection methods for microarray data select an optimal genes subset from abundant genes. The basic idea of most of these methods is to find the genes most relevant to the cancer classification and have a small redundancy among these genes. The expression level of the selected genes constitutes the characteristics of a sample. Then some classification methods such as Logistic Regression (LR), k-Nearest Neighbours (KNN), Naive Bayes (NB), and Support Vector Machine (SVM) are commonly used in microarray data classification [3].

Many feature selection methods were introduced in previous work about cancer classification in microarray data, which can be classified into three categories depending on the combination with the prediction model: filter [4], wrapper [5] and embedded [6]. However, most of these methods use only the gene expression level to select genes. The structural relations (e.g., physical, biological, functional, etc.) among genes are not considered. The prior knowledge reflects multiple relations information among genes and can better help us classify cancers [7]. Hence, developing an effective method to learn and exploit multiple relations among genes for feature selection in microarray data is crucial.

The limitations of existing research inspired us to look for other methods that allow efficient analysis and deal with problems of structural relations among genes. The properties of community structures in networks are connected in tight groups, where there are only looser connections between different communities [8]. Fig.1 shows an example of community division in a network. The similarity between nodes from the same community is high. In contrast, the similarity between nodes from different communities is low. Community detection is a fundamental issue in network science that aims to discover the community structure. Therefore, the joint community detection and network structure among genes provide span-new thinking for feature selection. For example,

*Corresponding author: liwei@cse.neu.edu.cn (Wei Li). This work supported by 2021YFC2701003, N2016006, 2021-MS-085.

Rostami et al. utilized the community detection and centrality of nodes to find the optimal gene subset, which ensures both maximum relevance with cancer and minimum redundancy among genes [9].

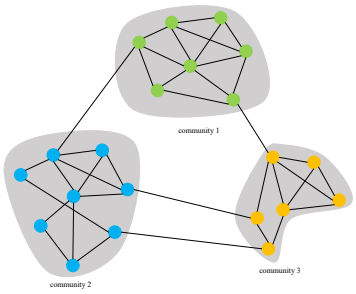

Fig. 1. An example of community division in a network. In this network, the nodes of each community (shaded part of the figure) are tightly connected, while the connections between different communities are sparse.

Recently, Graph Neural Networks (GNN) have gained remarkable success in leveraging the node feature and graph structure feature to improve the representation learning of the target graph [10]. Some researchers also proposed a few community detection algorithms by graph neural networks [11]. However, feature selection algorithms that combine community detection and graph neural networks are almost blank.

On the other hand, determining how to extract better and fuse the information of multiple gene relations is also a key consideration. Therefore, we propose a feature selection method MERCD based on community detection and graph neural networks to learn the multiple gene relation networks information. In summary, the main contribution of this paper is side-stepping the high dimensionality problem by using representative genes by clustering genes into disjoint exhaustive subsets ("communities") based on structure relations among genes.

II. Methodology

A. Notation And Background

Define the multiple gene relation networks data as $G = \{\mathcal{V}, \mathbf{E}^1, \ldots, \mathbf{E}^V, \mathbf{A}^1, \ldots, \mathbf{A}^V\}$. The normalized adjacency matrix of v-th gene relation $\tilde{\mathbf{A}}^v = (\mathbf{D}^v)^{-\frac{1}{2}}(\mathbf{A}^v + \mathbf{I})(\mathbf{D}^v)^{-\frac{1}{2}}$. The maximize the modularity index Q of the network was first introduced by Newman [12], which is defined by the following:

$$Q = \frac{1}{2M}\sum_{i,j}\left[\left(a_{ij} - \frac{k_i k_j}{2M}\right)\mathcal{Z}(i,j)\right] \tag{1}$$

where M is the total number of edges of the network.

We also define the modularity matrix $\mathbf{B} = [b_{ij}^v]$ of v-th gene relation network according to the work of Qiu et al. [11]:

$$b_{ij}^v = a_{ij}^v - \frac{k_i^v k_j^v}{2m^v} \tag{2}$$

where a_{ij}^v is the element of the adjacency matrix of v-th gene relation, k_i^v is the degree of node i of v-th gene relation and the m is the number of edge of the v-th gene network.

TABLE I
Summary of Notation.

Notation	Description
G	Multiple gene relation networks data
$\mathcal{V}$	Sets of gene
V	Number of relations
N	Number of genes
Q	Maximize modularity index
$\mathbf{B}^v$	Simplified maximize modularity of v-th gene relation
$\mathbf{E}^v$	Weight matrix of v-th gene relation
e_{ij}^v	Weight between gene i and gene j of v-th relation
$\mathbf{X}$	Gene expression value matrix of microarray data
$\mathbf{A}^v$	Adjacency matrix of v-th gene relation
$\mathbf{D}^v$	Degree matrix of v-th gene relation
$\mathbf{I}$	Identity matrix
$\tilde{\mathbf{A}}^v$	Normalized adjacency matrix of v-th gene relation
$\mathbf{Z}$	Community membership matrix
λ^v	Weight of v-th gene relation

In this way, each node has a modularity relationship with all the other nodes, whether they have connected edges.

All notations are summarized in Table I.

B. Variational Fusion-based Autoencoders Reconstruction

We design the inference model:

$$q(\mathbf{Z}^v \mid \mathbf{B}^v, \mathbf{A}^v, \mathbf{E}^v, \mathbf{X}) = \prod_{i=1}^{N} q\left(\mathbf{z}_i^v \mid \mathbf{B}^v, \mathbf{A}^v, \mathbf{E}^v, \mathbf{X}\right) \tag{3}$$

where $\mathbf{Z}^v = \left[\mathbf{z}_{ij}^v\right] \in \mathbb{R}^{N\times K}$ which each row $\mathbf{z}_i^v$ is the community membership vector of v-th gene ralation, and K is the dimension of the node community membership vector. $q(\mathbf{z}_i^v|\mathbf{B}^v, \mathbf{A}^v, \mathbf{E}^v, \mathbf{X})$ is a variational approximation of node i's true posterior distribution based on Gaussian family as:

$$q\left(\mathbf{z}_i^v \mid \mathbf{B}^v, \mathbf{A}^v, \mathbf{E}^v, \mathbf{X}\right) = \mathcal{N}\left(\mathbf{z}_i^v \mid \boldsymbol{\mu}_i, \operatorname{diag}\left(\boldsymbol{\sigma}_i^2\right)\right) \tag{4}$$

Then, we use two graph neural networks and autoencoders $\mu^{gcn} = GCN_\mu(\mathbf{B}^v, \mathbf{A}^v)$, $log\sigma^{gcn} = GCN_\sigma(\mathbf{B}^v, \mathbf{A}^v)$, $\mu^{ae} = AE_\mu(\mathbf{E}^v, \mathbf{X})$, $log\sigma^{ae} = AE_\sigma(\mathbf{E}^v, \mathbf{X})$ as encoders to fit the mean vectors μ and the standard deviation vectors σ for gene i in gene relation graph v. And the encoders have the following unified form:

$$\mathrm{GCN}(\mathbf{B}^v, \mathbf{A}^v) = \widetilde{\mathbf{A}}^v \tanh\left(\widetilde{\mathbf{A}}^v \mathbf{B}^v \mathbf{W}_0^v\right)\mathbf{W}_1^v \tag{5}$$

$$\mathrm{AE}(\mathbf{E}^\mathbf{v}, \mathbf{X}) = \mathcal{E}^v LeakyReLU\left(\mathcal{E}^v \mathcal{W}_0^v\right)\mathcal{W}_1^v \tag{6}$$

Finally, we combine the mean vectors μ and the standard deviation vectors σ from AE and GCN with a linear combination operation:

$$\mu = \alpha\mu^{gcn} + (1-\alpha)\mu^{ae} \tag{7}$$

$$\log\sigma = \alpha[\log\sigma]^{gcn} + (1-\alpha)[\log\sigma]^{ae} \tag{8}$$

where α is a learnable coefficient that selectively determines the importance of GCN and AE according to the property of different datasets, α is initialized as 0.5 and then tuned automatically with a gradient descent method. W_0 and W_1 represent the weight matrix for the first layer and second layer,

respectively, and W_0 is shared between GCN_σ and GCN_μ. $\mathcal{W}_0$ and $\mathcal{W}_1$ represent the weight matrix for the first layer and second layer, respectively, $\mathcal{W}_0$ also is shared between AE_σ and AE_μ. $\tilde{\mathbf{A}}^v$ is the normalized adjacency matrix and $\mathcal{E}^v = CONCAT[\mathbf{E}^v, \mathbf{X}]$.

Then, we use the graph convolution-like operation (i.e., message passing operation) [13] to enhance the initial embedding $\mathbf{Z}^v \in \mathbb{R}^{N\times K}$ by considering the local structure within the data:

$$\mathbf{Z}_L^v = \tilde{\mathbf{A}}^v \mathbf{Z}^v \tag{9}$$

$\mathbf{Z}_L^v \in \mathbb{R}^{N\times K}$ denotes the local structure enhanced $\mathbf{Z}^v$. After that, a self-correlated learning mechanism exploits the non-local relationship among genes' preliminary information space. Specifically, the normalized self-correlation matrix $\mathbf{S} \in \mathbb{R}^{N\times N}$ is calculated through (10):

$$\mathbf{S}_{ij} = \frac{e^{\left(\mathbf{Z}_L^v \mathbf{Z}_L^{v\,\mathbf{T}}\right)_{ij}}}{\sum_{k=1}^{N} e^{\left(\mathbf{Z}_L^v \mathbf{Z}_L^{v\,\mathbf{T}}\right)_{ik}}} \tag{10}$$

with $\mathbf{S}$ as coefficients, the $\mathbf{Z}_L$ recombined by considering the global correlation among genes: $\mathbf{Z}_G^v = \mathbf{S}\mathbf{Z}_L^v$.

Finally, the skip connection encourages information to pass smoothly:

$$\widetilde{\mathbf{Z}}^v = \beta \mathbf{Z}_G^v + \mathbf{Z}_L^v \tag{11}$$

where β is a scale parameter, we initialize it as 0 and learn its weight while training the network. The graph convolution-like operation considers the gene correlation from both the perspective of the local and global levels. Thus, it has the potential benefit of finely refining the information for learning latent consensus representations.

The conditional distribution of $p(\widetilde{\mathbf{B}}_{ij}^v | \mathbf{z}_i^v, \mathbf{z}_j^v)$ according to VGAE [14] is defined as the (12), where $\widetilde{\mathbf{B}}_{ij}$ is a reconstructed entry.

$$p\left(\widetilde{\mathbf{B}}_{ij}^v \mid \widetilde{\mathbf{z}}_i^v, \widetilde{\mathbf{z}}_j^v\right) = \left(\sigma\left(\widetilde{\mathbf{z}}_i^v \widetilde{\mathbf{z}}_j^{v\top}\right)\right)^a \left(\left(1-\sigma\left(\widetilde{\mathbf{z}}_i^v \widetilde{\mathbf{z}}_j^{v\top}\right)\right)\right)^{(1-a)} \tag{12}$$

where $\sigma(*) = \frac{1}{1+e^*}$ is a sigmoid function and the $a = \sigma\left(b_{ij}\right)$. The meaning of the above equation is clearer by understanding the following equation.

$$\begin{aligned}\log p\left(b_{ij} \mid \mathbf{z}_i^v, \mathbf{z}_j^v\right) = \sigma\left(b_{ij}\right) \log\left(\sigma\left(\mathbf{z}_i^v \mathbf{z}_j^{v\top}\right)\right) \\ + \left(1-\sigma\left(b_{ij}\right)\right) \log\left(1-\sigma\left(\mathbf{z}_i^v \mathbf{z}_j^{v\top}\right)\right)\end{aligned} \tag{13}$$

And the $p(\mathbf{B}^v \mid \widetilde{\mathbf{Z}}^v)$ is as follows:

$$p(\mathbf{B}^v \mid \widetilde{\mathbf{Z}}^v) = \prod_{i=1}^{N}\prod_{j=1}^{N} p\left(\widetilde{\mathbf{B}}_{ij}^v \mid \widetilde{\mathbf{z}}_i^v \widetilde{\mathbf{z}}_j^v\right) \tag{14}$$

We first give the variational lower bound $\mathcal{L}(\varphi^v, \theta^v)$ derived from the maximization objective function as follow:

$$\begin{aligned}&\mathbb{E}_{\mathcal{B}^v}\left[\log p_{\boldsymbol{\theta}}(\mathbf{B}^v)\right] = \\ &\mathbb{E}_{\mathcal{B}^v}\left[\mathbb{E}_{q_{\boldsymbol{\phi}}(\widetilde{\mathbf{Z}}^v|\mathbf{B}^v,\mathbf{A}^v,\mathbf{E}^v,\mathbf{X})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{B}^v, \widetilde{\mathbf{Z}}^v) - \log q_{\boldsymbol{\phi}}(\widetilde{\mathbf{Z}}^v \mid \mathbf{B}^v)\right]\right] \\ &\geq \mathcal{L}(\phi^v, \boldsymbol{\theta}^v) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\widetilde{\mathbf{Z}}^v|\mathbf{B}^v,\mathbf{A}^v,\mathbf{E}^v,\mathbf{X})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{B}^v \mid \widetilde{\mathbf{Z}}^v)\right] \\ &- \mathrm{KL}\left[q_{\boldsymbol{\phi}}(\widetilde{\mathbf{Z}}^v \mid \mathbf{B}^v, \mathbf{A}^v, \mathbf{E}^v, \mathbf{X}) \| p(\widetilde{\mathbf{Z}}^v)\right]\end{aligned} \tag{15}$$

where $\mathcal{B}^v$ is a modularity set of $G(\varphi^v, \theta^v) \in \{\mathbf{W}_\mathbf{0}^v, \mathbf{W}_\mathbf{1}^v, \mathbf{W}_\mathbf{2}^v, \mathcal{W}_0^v, \mathcal{W}_1^v, \mathcal{W}_2^v\}$ is the parameter space, and take a Gaussian prior

$$\prod_i p\left(\widetilde{\mathbf{z}}_i^v\right) = \prod_i \mathcal{N}\left(\widetilde{\mathbf{z}}_i^v \mid 0, \mathbf{I}\right) \tag{16}$$

Then optimization task is:

$$\begin{aligned}&\arg\max_{\boldsymbol{\phi}^v, \boldsymbol{\theta}^v} \sum_{v=1}^{V} \mathcal{L}(\boldsymbol{\phi}^v, \boldsymbol{\theta}^v) = \\ &\arg\max_{\boldsymbol{\phi}^v, \boldsymbol{\theta}^v} \sum_{v=1}^{V} \mathbb{E}_{q_{\boldsymbol{\phi}}(\widetilde{\mathbf{Z}}^v|\mathbf{B}^v,\mathbf{A}^v,\mathbf{E}^v,\mathbf{X}) \| p(\widetilde{\mathbf{Z}}^v)}\left[\log p_{\boldsymbol{\theta}}(\mathbf{B}^v \mid \widetilde{\mathbf{Z}}^v)\right] \\ &- \mathrm{KL}\left[q_{\boldsymbol{\phi}}(\widetilde{\mathbf{Z}}^v \mid \mathbf{B}^v, \mathbf{A}^v, \mathbf{E}^v, \mathbf{X}) \| p(\widetilde{\mathbf{Z}}^v)\right]\end{aligned} \tag{17}$$

The first term of this lower bound is a reconstruction loss, and we now consider the previous (12) form as follows:

$$\begin{aligned}&\mathbb{E}_{q_{\boldsymbol{\phi}}(\widetilde{\mathbf{Z}}^v|\mathbf{B}^v,\mathbf{A}^v,\mathbf{E}^v,\mathbf{X})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{B}^v \mid \widetilde{\mathbf{Z}}^v)\right] = \\ &\mathbb{E}_{q_{\boldsymbol{\phi}}(\widetilde{\mathbf{Z}}^v|\mathbf{B}^v,\mathbf{A}^v,\mathbf{E}^v,\mathbf{X})}\left[\sum_{i=1}^{N}\sum_{j=1}^{N} \log p\left(b_{ij} \mid \widetilde{\mathbf{z}}_i^v \widetilde{\mathbf{z}}_j^v\right)\right]\end{aligned} \tag{18}$$

C. Relation Weight Learning

We introduce a weighting factor to distinguish the contributions of different gene relation networks. Initially, all gene relation networks have equal weights $\lambda^v = \frac{1}{V}$. The optimization of the objective function is modified to be defined as follows:

$$\begin{aligned}&\arg\max_{\boldsymbol{\phi}^v, \boldsymbol{\theta}^v} \sum_{v=1}^{V} (\lambda^v)^\gamma + \lambda^v \mathbb{E}_{q_{\boldsymbol{\phi}}(\widetilde{\mathbf{Z}}^v|\mathbf{B}^v,\mathbf{A}^v)}\left[\log p_{\boldsymbol{\theta}}(\mathbf{B}^v \mid \widetilde{\mathbf{Z}}^v)\right] \\ &- \lambda^v \mathrm{KL}\left[q_{\boldsymbol{\phi}}(\widetilde{\mathbf{Z}}^v \mid \mathbf{B}^v, \mathbf{A}^v, \mathbf{E}^v, \mathbf{X}) \| p(\widetilde{\mathbf{Z}}^v)\right]\end{aligned} \tag{19}$$

For each gene relation v, we define

$$\begin{aligned}M^v = &\mathbb{E}_{q_{\boldsymbol{\phi}}(\widetilde{\mathbf{Z}}^v|\mathbf{B}^v,\mathbf{A}^v)}\left[\log p_{\boldsymbol{\theta}}(\mathbf{B}^v \mid \widetilde{\mathbf{Z}}^v)\right] \\ &- \mathrm{KL}\left[q_{\boldsymbol{\phi}}(\widetilde{\mathbf{Z}}^v \mid \mathbf{B}^v, \mathbf{A}^v, \mathbf{E}^v, \mathbf{X}) \| p(\widetilde{\mathbf{Z}}^v)\right]\end{aligned} \tag{20}$$

The optimization of the objective function is simplified as

$$\arg\max_{\boldsymbol{\phi}^v, \boldsymbol{\theta}^v} \sum_{v=1}^{V} \lambda^v M^v + (\lambda^v)^\gamma \tag{21}$$

Algorithm 1 Overall process of MGRCD algorithm

Input: multiple gene relation networks data $G = \{\mathcal{V}, \mathbf{E}^1, \ldots, \mathbf{E}^V, \mathbf{A}^1, \ldots, \mathbf{A}^V\}$, the gene expression value matrix $\mathbf{X}$, smooth parameter γ, the number of selected features k.

Output: optimal feature subset F.

$\lambda^v = \frac{1}{V}$;
$\tilde{\mathbf{A}}^v = (\mathbf{D}^v)^{-\frac{1}{2}}(\mathbf{A}^v + \mathbf{I})(\mathbf{D}^v)^{-\frac{1}{2}}$;
while convergence condition does not meet **do**
  calculate the variational lower bound $\mathcal{L}(\phi^v, \boldsymbol{\theta}^v)$ for each gene relation network v;
  Update $\mathbf{W}_{\mathbf{0}}^v, \mathbf{W}_{\mathbf{1}}^v, \mathbf{W}_{\mathbf{2}}^v, \mathcal{W}_0^v, \mathcal{W}_1^v, \mathcal{W}_2^v$ in (5) and (6) via Adam according to (21);
  for each gene relation network v **do**
    Update λ^v in (22);
  end for
end while
Calculate the $\mathbf{Z}$ matrix according to (23);
Get the community division results according to the $\mathbf{Z}$ matrix with k-means;
Calculate p-value as a correlation score of genes with cancer in each community;
Select the genes with the smallest p-value in each community;
Obtain the optimal feature subset F.

By setting its derivation to zero and normalization, we get

$$\lambda^v = \left(\frac{M^v}{\gamma}\right)^{\frac{1}{\gamma-1}} = \frac{e^{\lambda^v}}{\sum_{v=1}^{V} e^{\lambda^v}} \tag{22}$$

where γ is smooth parameter.

We alternatively update $\{\mathbf{W}_{\mathbf{0}}^v, \mathbf{W}_{\mathbf{1}}^v, \mathbf{W}_{\mathbf{2}}^v, \mathcal{W}_0^v, \mathcal{W}_1^v, \mathcal{W}_2^v\} \in G(\phi^v, \theta^v)$ and λ^v until convergence.

D. Community Division and Feature Selection

Finally, we can obtain the community membership matrix $\mathbf{Z}$ to be defined as follows:

$$\mathbf{Z} = \sum_{v=1}^{V} \lambda^v \widetilde{\mathbf{Z}}^v \tag{23}$$

Then, we get the community division results according to the $\mathbf{Z}$ matrix with cluster algorithms k-means. After that, we used the T-test to obtain the p-value for gene and cancer correlations in each community. Then, we select the genes with the smallest p-value of the T-test in each community as representative genes, forming the optimal feature subset of the microarray data. The complete procedures of MGRCD are outlined in Algorithm 1.

III. Experiment

A. Datasets And Metrics

The five publicly available datasets from NCBI[1] are downloaded for our experiments. The details are described in Table II, including the number of genes, samples, and different classes. Since microarray data are available in many forms depending on the laboratory or platform, we preprocess the data used in the experiments. First, we removed some of the missing data and then logarithmized the data after standard normalization to ensure data consistency and continuity. The different types of gene networks of each microarray data are obtained by GeneMANIA[2]. We use four common types of edge (Co-expression, Co-localization, Genetic Interaction, and Physical Interactions) as the input of our model. Fig.2 shows an example of gene relation networks from GeneMANIA.

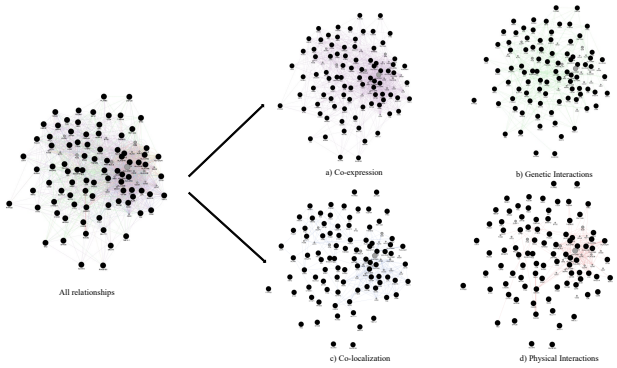

Fig. 2. An example of gene relation networks from GeneMANIA. On the left is the gene relation network containing all types, and on the right is the gene relationship network for each type, including Co-expression, Genetic Interactions, Co-localization, and Physical Interactions, respectively.

TABLE II
DESCRIPTION OF MICROARRAY DATASETS.

Dataset	Genes	Samples	Pos	Neg
ALL4	12625	93	67	26
DLBCL	7129	77	19	58
Leukaemia	7129	72	47	25
Myeloma	12625	173	36	137
Prostate	12600	102	50	52

We use the SVM classifier to evaluate the performance of our model by the average classification accuracy, precision, recall, and F1 score of 10-fold cross-validation. These metrics are defined as follows:

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{24}$$

$$Precision = \frac{TP}{TP + FP} \tag{25}$$

$$Recall = \frac{TP}{TP + FN} \tag{26}$$

[1]https://www.ncbi.nlm.nih.gov/geo/
[2]http://genemania.org/

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{27}$$

where TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative, respectively.

B. Experimental Setting

Since many genes are not related to cancer in the gene microarray data. Therefore, we calculate p-values for gene and cancer correlations using the T-test to eliminate most cancer-related genes. Furthermore, we only reserve the top 1000 genes for each dataset as input to the model according to the positive order of p-values. The matrices $\mathbf{B}^1, \ldots, \mathbf{B}^V$ according to (2) and the gene relation networks are obtained from GeneMANIA. We initialize the weight $\lambda^v = \frac{1}{V}$ of each relation in the microarray dataset and the parameter $\gamma = -1$ in (22).

After obtaining the $\mathbf{Z}$ matrix, we used the k-means clustering algorithm to obtain the community divisions for all genes in microarray data. However, although there are some criteria to evaluate the performance of community divisions, it is still essentially unsupervised learning. We do not know how many communities to partition into to get the best cancer classification performance. Therefore, we traverse k from 1 to 20 to get the best cancer classification results, providing more flexibility for choosing the number of genes and classification performance.

C. Experimental Results

Fig.3 shows the cancer classification performance of MGRCD with different numbers of selected genes in different microarray datasets. We can observe that the cancer classification performance in the microarray datasets is more sensitive to a smaller number of selected genes. In contrast, the sensitivity of different datasets to changes in the number of selected genes varies. As the number of selected genes increases, the overall evaluation metric gradually stabilizes, with some datasets (e.g., ALL4 and Leukaemia) showing a decreasing trend, which indicates that it is necessary to initialize an appropriate number of communities in our model for different datasets. This demonstrates that too small a choice of the number of communities is not too wise for gene selection in microarray data.

To demonstrate the significance of multiple gene relation networks in community detection, we used the k-means algorithm to divide communities directly in microarray data by gene clustering. The results are shown in Table III. We set the same number of communities $k = 5$ (number of selected genes) for both methods. As seen from this Table, MGRCD achieves better cancer classification performance in allover evaluated metrics in the tested dataset, which proves the effectiveness of multiple gene relation networks in gene selection in microarray data by community detection. This result arises because our community detection algorithm fully learns multiple gene network structures and gene expression attribute values through graph neural networks and autoencoders.

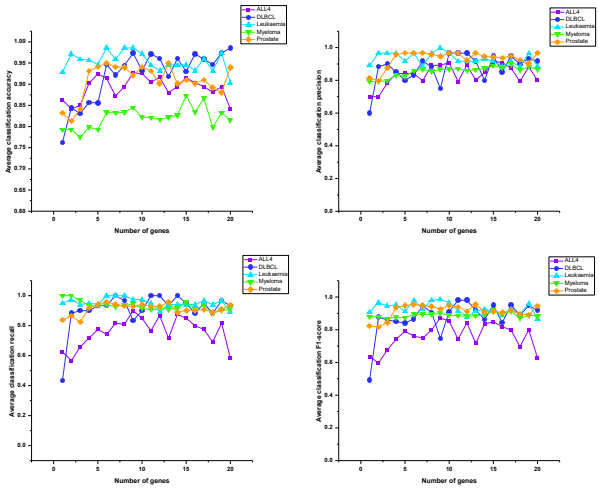

Fig. 3. Average accuracy, recall, precision, and F1 score of MGRCD on all microarray datasets as a function of the number of features selected.

TABLE III
WHEN THE NUMBER OF COMMUNITIES IS SET TO 5, THE PERFORMANCE (MEAN ± STANDARD DEVIATION) OF FEATURE SELECTION BY DETECTING COMMUNITIES IS COMPARED USING ONLY K-MEANS AND MGRCD.

Dataset	Metric	k-means	MGRCD
ALL4	Accuracy	0.8300±0.1005	**0.9244±0.0523**
	Precision	0.6100±0.4006	**0.8417±0.3202**
	Recall	0.5667±0.3784	**0.7750±0.3144**
	F1 score	0.5617±0.3578	**0.7914±0.2917**
DLBCL	Accuracy	0.8911±0.1115	**0.9054±0.0941**
	Precision	0.7833±0.2491	**0.8000±0.2194**
	Recall	0.9333±0.1405	**0.9333±0.1405**
	F1 score	0.8233±0.1764	**0.8400±0.1481**
Leukaemia	Accuracy	0.9304±0.0736	**0.9464±0.0694**
	Precision	0.8833±0.1933	**0.9167±0.1800**
	Recall	0.9300±0.1054	**0.9417±0.1245**
	F1 score	0.8981±0.1193	**0.9124±0.1224**
Myeloma	Accuracy	0.8150±0.0855	**0.8333±0.0862**
	Precision	0.8420±0.1050	**0.8582±0.0971**
	Recall	0.8954±0.0902	**0.9404±0.0460**
	F1 score	0.8781±0.0681	**0.8949±0.0612**
Prostate	Accuracy	0.8909±0.0996	**0.9409±0.0839**
	Precision	0.8955±0.1508	**0.9667±0.1054**
	Recall	0.9114±0.1343	**0.9429±0.0999**
	F1 score	0.8890±0.1023	**0.9479±0.0764**

D. Comparison Experiments

This section compares the average cancer classification accuracy of MGRCD with other feature selection published state-of-the-art methods in microarray data. Table IV shows the comparison experiment results. For different gene selection methods on the same dataset, using fewer genes to accomplish higher classification performance is often considered more efficient. Most of these traditional methods use gene relations such as GNNSC [15], and some use community detection methods such as CDNC [9]. We can observe that MGRCD achieved better classification accuracy with fewer genes, indicating the advantage and effectiveness of our model in using multiple gene relation networks for the gene feature selection

task of microarray data.

TABLE IV
COMPARISON OF THE CANCER CLASSIFICATION PERFORMANCE OF MGRCD AND OTHER STATE-OF-THE-ART FEATURE SELECTION METHODS IN MICROARRAY DATA.

Dataset	Method	Accuracy	Genes
DLBCL	GNNSC [15]	0.9464	15
	FCBF [16]	0.9610	44.3
	IG-GA [17]	0.9487	110
	EnCFSH [18]	0.9481	10.4
	MGRCD	**0.9732**	**7**
Leukaemia	FJMI [19]	0.8813	3.19
	FSDNE [20]	0.9520	9
	GSFSJNE [21]	0.9273	3
	CDNC [9]	0.9116	3.27
	EnCFSH [18]	0.9583	45.3
	MGRCD	**0.9714**	**2**
Prostate	FJMI [19]	0.7913	2.63
	BHAPSO [22]	0.8281	2.83
	AHEDL [23]	0.7972	1.64
	ABCD [24]	0.8267	2.71
	CDNC [9]	0.8391	2.81
	GNNSC [15]	0.8255	2
	MGRCD	**0.8518**	**2**

IV. CONCLUSION

This paper provides an unsupervised feature selection method by community detection fusing multiple gene relation network information. The method uses graph neural networks and autoencoders to learn genes' attributes and multiple structural information in microarray data and reconstruct the modularity matrix using the inferential model. And then, based on the results of the community division, a gene most relevant to cancer classification is selected from each community to constitute the optimal feature subset. We conducted experiments on several public microarray datasets to validate the performance of the selected genes on the cancer classification task. The experimental results show that our method is effective. Meanwhile, this study also validates the positive effect of gene relations in microarray data for cancer classification.

ACKNOWLEDGMENT

This document is the results of the research project funded by the National key research and development program, China (2021YFC2701003), the Fundamental Research Funds for the Central Universities (N2016006), Natural Science Foundation of Liaoning Province under grant 2021-MS-085.

REFERENCES

[1] O. A. Alomari, S. N. Makhadmeh, M. A. Al-Betar, Z. A. A. Alyasseri, I. A. Doush, A. K. Abasi, M. A. Awadallah, and R. A. Zitar, "Gene selection for microarray data classification based on gray wolf optimizer enhanced with triz-inspired operators," *Knowl. Based Syst.*, vol. 223, p. 107034, 2021.

[2] M. Abd-Elnaby, M. Alfonse, and M. Roushdy, "Classification of breast cancer using microarray gene expression data: A survey," *J. Biomed. Informatics*, vol. 117, p. 103764, 2021.

[3] E. Alhenawi, R. Al-Sayyed, A. Hudaib, and S. Mirjalili, "Feature selection methods on gene expression microarray data for cancer classification: A systematic review," *Computers in Biology and Medicine*, vol. 140, p. 105051, 2022.

[4] W. Duch, T. Wieczorek, J. Biesiada, and M. Blachnik, "Comparison of feature ranking methods based on information entropy," in *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No. 04CH37541)*, vol. 2. IEEE, 2004, pp. 1415–1419.

[5] R. Ruiz, J. C. Riquelme, and J. S. Aguilar-Ruiz, "Incremental wrapper-based gene selection from microarray data for cancer classification," *Pattern Recognition*, vol. 39, no. 12, pp. 2383–2392, 2006.

[6] N. Hoque, H. Ahmed, D. Bhattacharyya, and J. Kalita, "A fuzzy mutual information-based feature selection method for classification," *Fuzzy Information and Engineering*, vol. 8, no. 3, pp. 355–384, 2016.

[7] H. Azzawi, J. Hou, R. Alanni, and Y. Xiang, "A hybrid neural network approach for lung cancer classification with gene expression dataset and prior biological knowledge," in *Machine Learning for Networking - First International Conference, MLN 2018, Paris, France, November 27-29, 2018*, vol. 11407. Springer, 2018, pp. 279–293.

[8] S. Fortunato and D. Hric, "Community detection in networks: A user guide," *CoRR*, vol. abs/1608.00163, 2016.

[9] M. Rostami, S. Forouzandeh, K. Berahmand, M. Soltani, M. Shahsavari, and M. Oussalah, "Gene selection for microarray data classification via multi-objective graph theoretic-based method," *Artif. Intell. Medicine*, vol. 123, p. 102228, 2022.

[10] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, vol. 29, 2016, pp. 3837–3845.

[11] C. Qiu, Z. Huang, W. Xu, and H. Li, "VGAER: graph neural network reconstruction based community detection," *CoRR*, vol. abs/2201.04066, 2022.

[12] M. E. Newman, "Modularity and community structure in networks," *Proceedings of the national academy of sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

[13] W. Tu, S. Zhou, X. Liu, X. Guo, Z. Cai, E. Zhu, and J. Cheng, "Deep fusion clustering network," in *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI*. AAAI Press, 2021, pp. 9978–9987.

[14] T. N. Kipf and M. Welling, "Variational graph auto-encoders," *CoRR*, vol. abs/1611.07308, 2016.

[15] K. Yu, W. Xie, L. Wang, S. Zhang, and W. Li, "Determination of biomarkers from microarray data using graph neural network and spectral clustering," *Scientific Reports*, vol. 11, no. 1, pp. 1–11, 2021.

[16] A. Wang, H. Liu, J. Liu, H. Ding, J. Yang, and G. Chen, "Stable and accurate feature selection from microarray data with ensembled fast correlation based filter," in *IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2020, Virtual Event, South Korea, December 16-19, 2020*. IEEE, 2020, pp. 2996–2998.

[17] H. Salem, G. Attiya, and N. A. El-Fishawy, "Classification of human cancer diseases by gene expression profiles," *Appl. Soft Comput.*, vol. 50, pp. 124–134, 2017.

[18] A. Wang, H. Liu, J. Yang, and G. Chen, "Ensemble feature selection for stable biomarker identification and cancer classification from microarray expression data," *Comput. Biol. Medicine*, vol. 142, p. 105208, 2022.

[19] X. Tang, Y. Dai, and Y. Xiang, "Feature selection based on feature interactions with application to text categorization," *Expert Syst. Appl.*, vol. 120, pp. 207–216, 2019.

[20] L. Sun, X. Zhang, Y. Qian, J. Xu, and S. Zhang, "Feature selection using neighborhood entropy-based uncertainty measures for gene expression data classification," *Inf. Sci.*, vol. 502, pp. 18–41, 2019.

[21] L. Sun, X. Zhang, Y. Qian, J. Xu, S. Zhang, and Y. Tian, "Joint neighborhood entropy-based gene selection method with fisher score for tumor classification," *Appl. Intell.*, vol. 49, no. 4, pp. 1245–1259, 2019.

[22] E. Pashaei, E. Pashaei, and N. Aydin, "Gene selection using hybrid binary black hole algorithm and modified binary particle swarm optimization," *Genomics*, vol. 111, no. 4, pp. 669–686, 2019.

[23] X. Zheng, W. Zhu, C. Tang, and M. Wang, "Gene selection for microarray data classification via adaptive hypergraph embedded dictionary learning," *Gene*, vol. 706, pp. 188–200, 2019.

[24] V. Coleto-Alcudia and M. A. Vega-Rodríguez, "Artificial bee colony algorithm based on dominance (ABCD) for a hybrid gene selection method," *Knowl. Based Syst.*, vol. 205, p. 106323, 2020.