Supplementary manuscript of

Label reusing based graph neural network for unbalanced classification of personalized driver genes in cancer

Han-Wen Wan1,2,#, Meng-Han Wu3,#, Wen-Shan Zhao3,Han Cheng3,Ying Bi1, Xian-Fang Wang 4,Xiang-Rui Zhang3,Yan Li5, Wei-Feng Guo1,6,7\*

1School of Electrical and Information Engineering, Zhengzhou University, Zhengzhou 450001, China

2 School of International College, Zhengzhou University, Zhengzhou 450001, China

3 School of Life Sciences, Zhengzhou University, Zhengzhou 450001, China

4The school of computer science and engineering, Henan Institute of Technology, Xinxiang 453000, China

5 Key Laboratory of Information Fusion Technology of Ministry of Education, School of Automation, Northwestern Polytechnical University, Xian, 710072, China

6 State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Sun Yat-sen University Cancer Center, Guangzhou 510060, China

7 State Key Laboratory of Intelligent Agricultural Power Equipment, Zhengzhou University, Luoyang 471000, China

\* To whom correspondence should be addressed.

[Tel:86-0371-677810](Tel:86-21-54920100)18; Fax: [86-0371-677810](Tel:86-21-54920100)18; Email: [guowf@zzu.edu.cn](mailto:liangjing@zzu.edu.cn,)

#The authors should be regarded as Joint First Authors.

**Supplementary note 1:The statistic information of positive and negative genes in PGIN**

Table S1 The statistic information of positive and negative genes in PGIN

|  |  |  |  |
| --- | --- | --- | --- |
|  | BRCA | LUSC | LUAD |
| #posi | 18.4602+-1.5749 | 27.7292+-3.0479 | 29.4286+-3.7005 |
| #neg | 191.9115+-15.8389 | 419.7292+-23.0389 | 412.5000+-24.4861 |

**Supplementary note 2: Class weighted loss function**

To balance the loss of imbalanced classes or labels in PGIN, we adopted a class-balanced softmax cross-entropy loss function 34,35 by introducing a normalized weighting factor, 1/where is the effective number of genes for class *y* in PGIN. The effective number of genes for class *y* is defined as ),=*N*/(*N*-1) where *ny* is the number of genes in the class or label *y* and *N* is the number of genes in PGIN. The class weighted loss function can be written as



wheredenotes the model's estimated class probabilities. The weighting factor is a normalized vector=*C; C* is the number of classes (here *C* = 2); is softmax cross-entropy loss function. If the gene *x* is within a positive label set, its label , otherwise, .

**Supplementary note 3:The parameters in PersonalizedGNN**

The parameters are set in Table S2: The number of layers was 3, and the number of nodes in one layer was 750. The learning rate was set as and weight decay was set as 1e-7. The number of attention heads for the final classification layers and other layers was set as . Furthermore, during training of the GAT with label reuse strategy in our PersonalizedGNN model, the coefficient of Dropout was , while the coefficients of DropEdge and DropAttention were for coping with the training set in PersonalizedGNN.

Table S2 The parameters of PersonalizedGNN on cancer driver gene prediction

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| The number of layers | The number of nodes in each layer | weight decay | learn rate | coefficient of Dropout | coefficient of DropEdge | coefficient of DropAttention | The number of attention heads | epoch |
| 3 | 750 | 1E-7 | 0.02 | 0.25 | 0.75 | 0.75 | 5 | 2000 |

To further demonstrate the effectiveness of the setting parameters in our PersonalizedGNN, we tuned one of the parameters over suitable ranges of parameter values while we fixed other parameters in BRCA, LUSC, LUAD data sets. For each cancer data set, we used the same parameters among all individual patients and calculated the average gene ranking score among all individual patients for each cancer data. Based on the average gene ranking score of the cancer data, we obtained the gene ranking result and calculated the *Average* *Precision* of different parameter choices on each cancer data. The parameter setting results on BRCA, LUAD, and LUSC were shown in **Fig. S1**. As shown in Fig. S1, PersonalizedGNN has the best performance for our setting parameters in PersonalizedGNN.

**para check**

**Fig S1** The error bar in terms of *Average* *Precision* for different parameter choices among top *k* (*k*=1, 2, ..., 30) ranking genes on BRCA, LUSC, LUAD data sets.

**Supplementary note 4: Parameters in other methods**

* ChebNet: the graph convolution kernel size is 2; the number of layers is 3; the size of each hidden layer is 750; the number of epochs is 2000; bias in graph convolution is enabled, the activation function is ReLU; the dropout rate is 0.25; the learning rate is 0.02, and the optimizer is Adam.
* GraphSAGE: the number of layers is 3; the size of each hidden layer is 750; the number of epochs is 2000; the number of sampled neighbors is 10; the batch size is 50; and the learning rate is 0.02.
* GCN: the number of layers is 3; the size of hidden layers is 750; the num of epochs is 2000; the the dropout rate is 0.25; learning rate is 0.02; and the weight decay is 1e-7.
* GAT without label reuse: the heads of attention mechanism is 5; the number of layers is 3; the size of each hidden layer is 500; the number of epochs is 2000; the coefficient of Dropout, DropEdge and DropAttention is 0.25, 0.75 and 0.75 respectively; the learning rate is 0.02; the weight decay is 1e-7; and the activation function is LeakyReLU with the slope alpha equals 0.2.