USER MANUAL FOR IOAT

Chapter 1. Software Overview

This system is a data analysis tool based on machine learning. In multi-omics data using bioinformatics, the clinical data survival time and survival status of multi-omics data can be combined with multiple omics data such as gene expression data , Methylated data, copy number combination, pre-processing data through relevant machine learning methods, feature screening, clustering the filtered features to use the clustered results as the true labels of data for survival analysis, its main process as shown in Figure 1. The function of each module of the tool of the system is now fully explained:

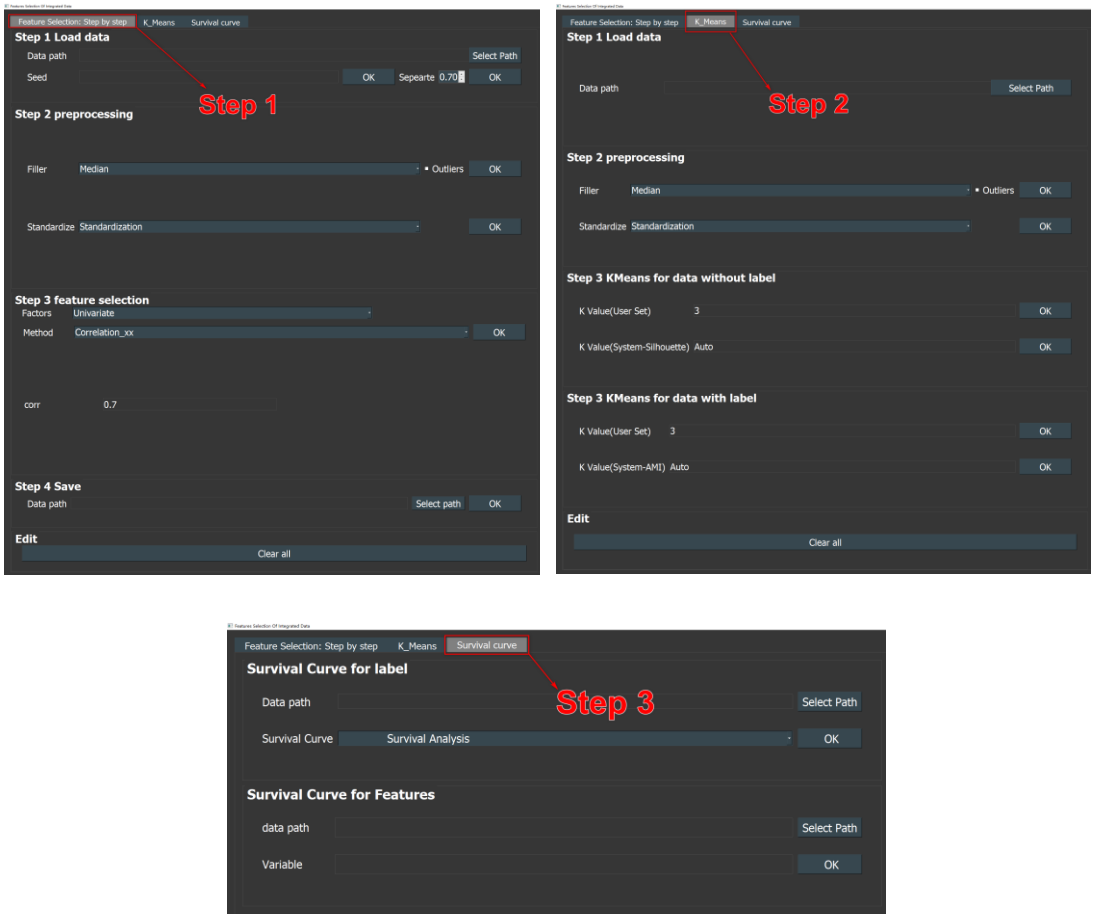

Figure1 Tool overall process

Chapter 2. Installation

The chapter explains how to download and install IOAT on the user's computer.

2.1 Requirement

1) Hardware requirements

   a) Intel Pentium III/800 MHz or higher (or compatible) although one should probably not go below a dual core processor.

b) 2 GB RAM minimum.

2) Software requirements

a) Supported operating system (OS) versions (32-bit or 64-bit)

Windows 7 SP1

Windows Server 2008 R2 SP1

Windows Server 2008 SP2

Windows Server 2012 R2

Windows 8

Windows 10

b) Python v3.5.6 or higher (for Windows) .

c) R v3.5.1 or higher (for Windows) .

2.2 Configuration of IOAT Environment

2.2.1 Installing R-3.5.1-win.exe

Users should install R-3.5.1-win.exe of these IOAT-software before starting IOAT. After installing R-3.5.1-win.exe, the user needs to configure the environment R variables. R-3.5.1-win.exe download link is https://pan.baidu.com/s/1PnZlfxYI11lQJ6yecABGtg, password:xkd0.

复制这段内容后打开百度网盘手机 App，操作更方便哦

2.2.2 Installing IOAT.exe of python

After installing R-3.5.1-win.exe, the user needs to execute IOAT.exe in the IOAT-software file to run the IOAT tool. IOAT-software file download link is https://pan.baidu.com/s/1PnZlfxYI11lQJ6yecABGtg, password:xkd0.

2.3 Download

IOAT can be freely downloaded from https://pan.baidu.com/s/1PnZlfxYI11lQJ6yecABGtg, password:xkd0. Compress the zip package (or 7z) into a specified file folder.

IOAT and the graphical user interface (GUI) of IOAT will be shown in Fig. 2

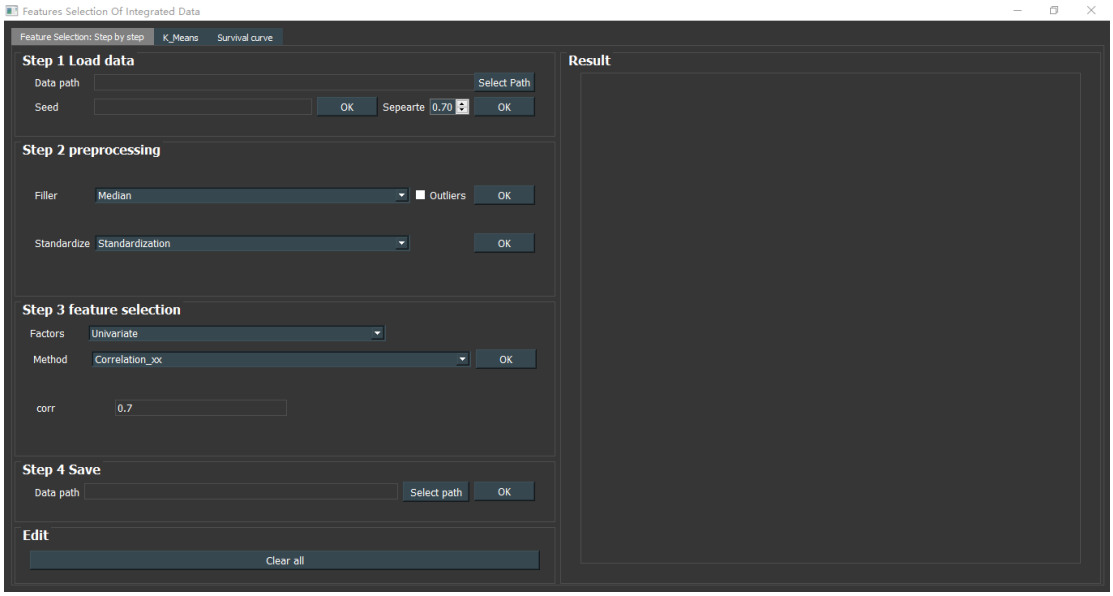

Figure 2. The GUI of IOAT

2.4 Data format requirements

This tool is mainly for data in CSV format, and requires the input data format to be in the first column of data with a label of 'time', which is the patient's survival time; the second column of data with a label of 'status', which is the patient's survival status; the next is multi-omics features of data fusion.

Chapter 3. Main Functions

3.1 Feature selection

3.1.1 Read data

The user reads the data to be analyzed through the Select Path button (the data format is: the label in the first column is the time name, the second column is the status label, and the other columns are features); and a random seed point is set to enable split testing and training The data of the set, including the results of each run in the case of setting the same seed point in the subsequent model training process. In the future, users can also choose to set the size of the split training test data set. The default value is 0.70, that is, the data set is divided into 70% training set and 30% test set, as shown in Figure 3.

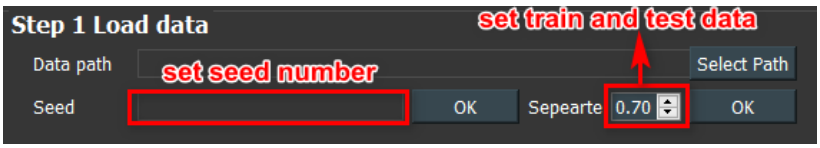

Figure 3 Download and segmentation of the data set

3.1.2 Preprocess data

The user can select the Outliers button to remove outliers and fill in the missing values in two ways: use the average of the eigenvalues of each column to fill the NaN data; use the median of the eigenvalues of each column to perform the NaN data Fill. Then the user can perform feature scaling on the data filled with missing values in two ways, so that the features of different dimensions are at the same numerical level, reducing the influence of features with large variance, and making the subsequent model more accurate. The method includes Standardization and MinMaxScaler, StandardScaler method is to scale the feature to the range of mean 0 and variance 1. MinMaxScaler is to scale the feature to the range between 0 and 1, as shown in Figure 4.

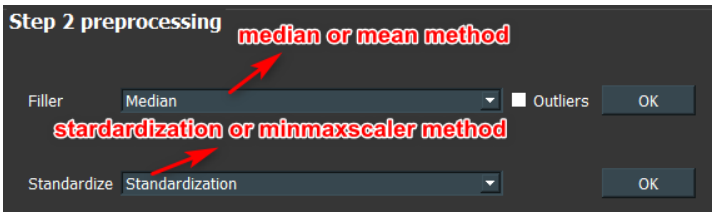

Figure 4 Data preprocessing

3.1.3 Select features

After data preprocessing, users can select the Correlation_xx method in single factor analysis and set the correlation threshold to find the relationship between features and features to filter out the features whose correlation coefficient is less than the threshold, as shown in Figure 5;

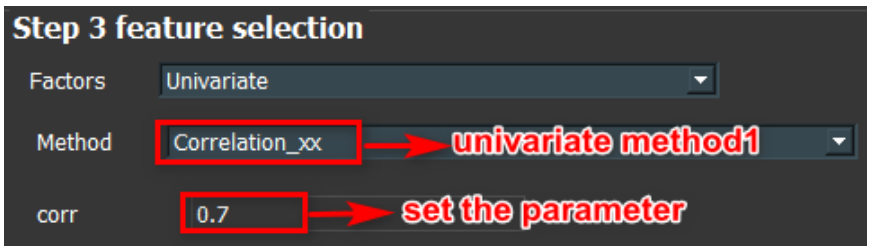

Figure 5 Single factor feature screening method 1

You can also choose the single factor Cox regression method in single factor analysis to find the relationship between features and survival time and survival status to filter out the statistically significant features with p-value less than or equal to the threshold, as shown in Figure 6;

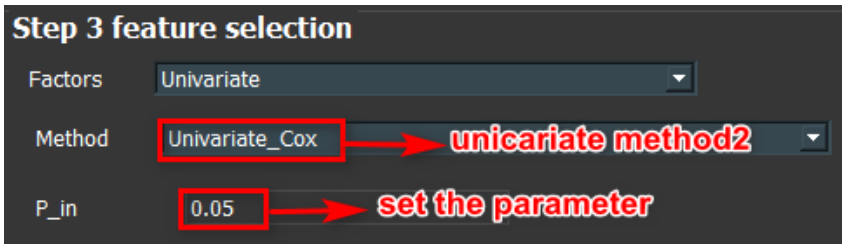

Figure 6 Single factor feature screening method 2

You can also choose the logrank test method in the single factor analysis method to find the relationship between features and survival time and survival status to filter out the features with a p value of less than or equal to the threshold value, as shown in Figure 1-6.

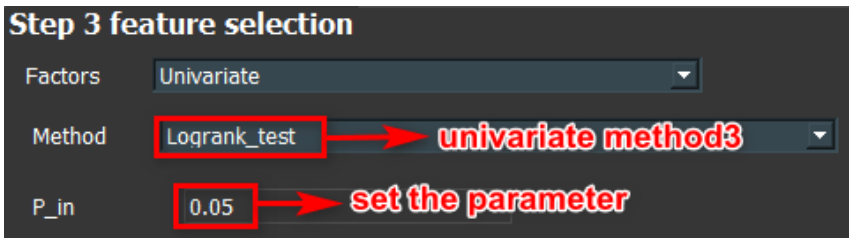

Figure 7 Single factor feature screening method 3

After the preliminary single-factor feature screening, further screening features can be carried out (of course, users can also choose to stop here). Multi-factor feature screening, users can choose multi-factor Cox regression to perform feature screening, and save the filtered features or calculation Cox risk value radscore. When the user

selects the Select button, the radscore value is calculated for the selected features, and if not selected, the radscore value is calculated for all features, as shown in Figure 8;

Figure 8 Multi-factor feature screening method 1

You can also choose Lasso's feature selection method for feature screening, which is a Lasso adaptive method based on the Cox model. As shown in Figure 9.

Figure 9 Multi-factor feature screening method 2

Each feature selection will be based on the last method to perform a feature selection on the remaining features, after each single-factor or multi-factor feature selection will draw a coefficient map of the relationship between the features The heat map of the sample features, and the Lasso path map is given when the Lasso feature is filtered, so that the user can see the change of each regression coefficient with the penalty coefficient, and the order of the independent variables exiting the model, and provide the user with the visualization results of the method.

After getting the filtered feature, the user can select the location where the feature is saved by selecting the path button, and click OK to save it to the specified location for the next operation.

3.1.4 Save result

This module can save the result of feature selection to the location specified by the user, click the Select path button to select the save location, and click OK to save the features retained by the feature filter to the specified location,as shown in Figure10.

Figure 10 Save filter results

3.1.5 Display steps

The tool displays all user operations in the result column and returns the results of each operation, as shown in Figure 11.

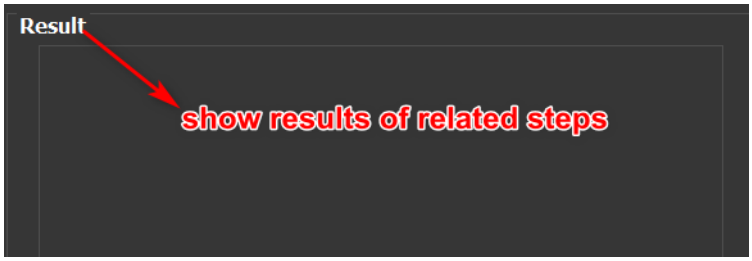

Figure 11 User operation and result display box

3.1.6 Clear all

When the user wants to analyze the new data, the user can click the clear all button to clear all the display results in the result box, so that the user can analyze the new data, as shown in Figure 12.

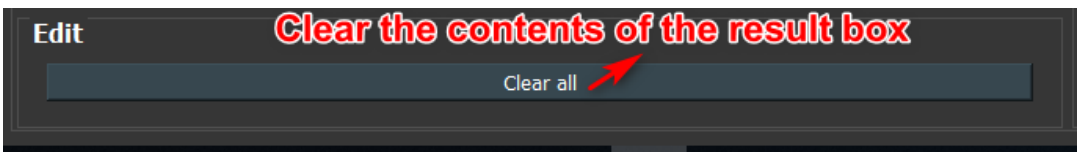

Figure 12 Clear all operations

3.2 K-Means clustering

3.2.1Read data

Perform k-Means clustering on the filtered features or other feature data, and the user reads in the data, as shown in Figure 13.

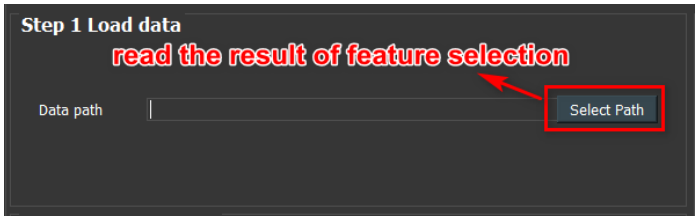

Figure 13 Reading and preprocessing of feature data

3.2.2 Preprocess data

Then perform data preprocessing on the data, as shown in Figure14.

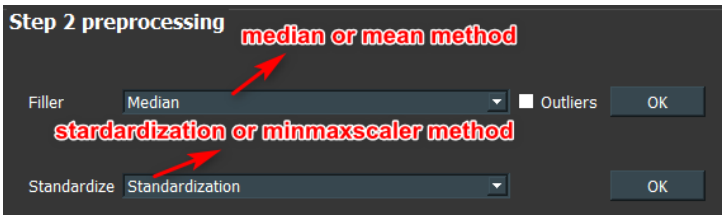

Figure 14 Feature data preprocessing

3.2.2 Cluster

After data and processing, the user can choose the number of clusters. The default value is 3, as shown in Figure 15.

Figure 15 User set the number of clusters

For unlabeled data (not classified), the system can use the silhouette method to find the best number of clusters, whose range is within 3-9, as shown in Figure 16, and draw the correlation coefficient and k value obtained by silhouette In the figure, the higher the correlation coefficient value obtained from the silhouette, the better the

clustering effect.

Figure 16 The system automatically finds the best number of clusters

For labeled data (divided into layers), the system can find the best number of clusters through the AMI mutual information method, and the range is within 3-9.

The results of the clustered classes are automatically retained on the user's desktop. The user can keep the results of the number of categories selected by the system and the number of categories selected by the system at the same time, and compare it through subsequent survival analysis for subsequent research.

3.2.3 Display steps

The tool displays all user operations in the result column and returns the results of each operation, as shown in Figure 17.

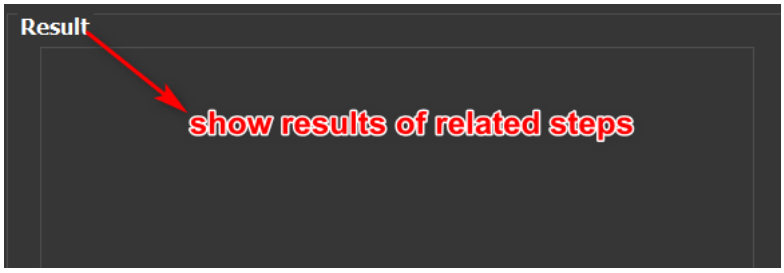

Figure 17 User operation and result display box

3.2.3 Clear all

When the user wants to analyze the new data, the user can click the clear all button to clear all the display results in the result box, so that the user can analyze the new data, as shown in Figure 18.

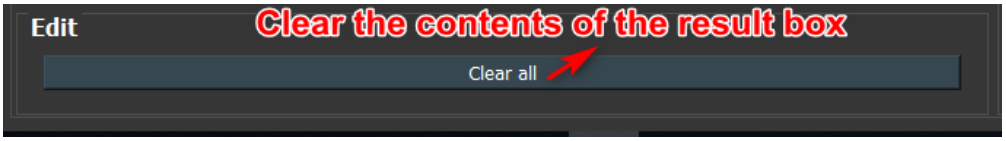

Figure 18 Clear all operations

3.3 Survival analysis

3.3.1 Read data

Perform a survival analysis on the clustering results, and the user can read in the obtained clustering result data, as shown in Figure 19.

Figure 19 Reading in clustering result data

3.3.2 Survival analysis

After reading in the clustering results, click the analysis button to perform survival analysis and analysis, and get a survival curve graph, so that users can view the clustering results, as shown in Figure 20. Calculate the p value between each category by the logrank test test method. When p is less than or equal to 0.05, there is a significant difference between the clustering results.

Figure 20 Survival analysis of clustering results

Users can also perform survival analysis on a certain feature (discrete) to see whether there is a significant difference between each group, as shown in Figure 21.

Figure 21 Survival analysis of a discrete feature

3.4 Data Visualization

3.4.1 Select feature

This tool uses multi-omics data of gliomas as an example to show the visualization results of related tool processes. The first feature selection uses single factor feature selection using the Cox model to obtain 2147 selected features, and obtain a training

and test set of hierarchical clustering heat map, as shown in Figure 22.

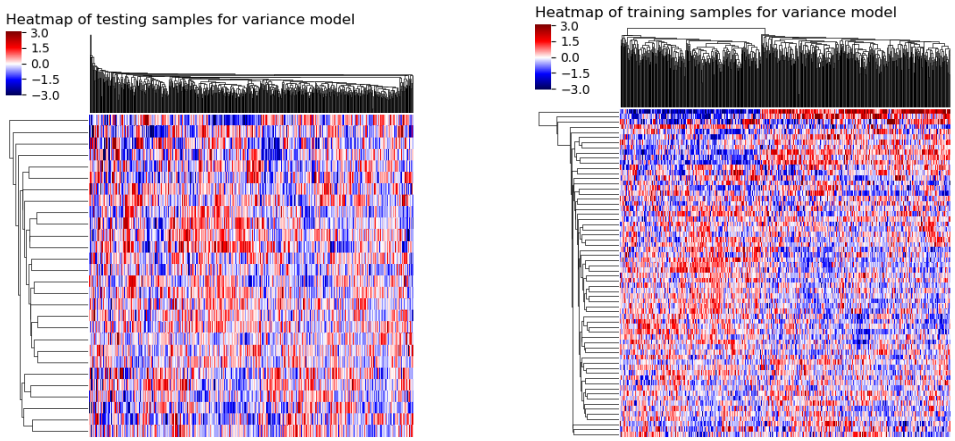

Figure 22 First feature screening training set and test set hierarchical clustering heat map

On this basis, the multi-factor feature screening Lasso method was used for the second feature screening, and the hierarchical clustering heat map of the training and test set was obtained, as shown in Figure 23; the relationship between the penalty coefficient and the features, as shown in the figure 24; The relationship between features and features, as shown in figure 25.

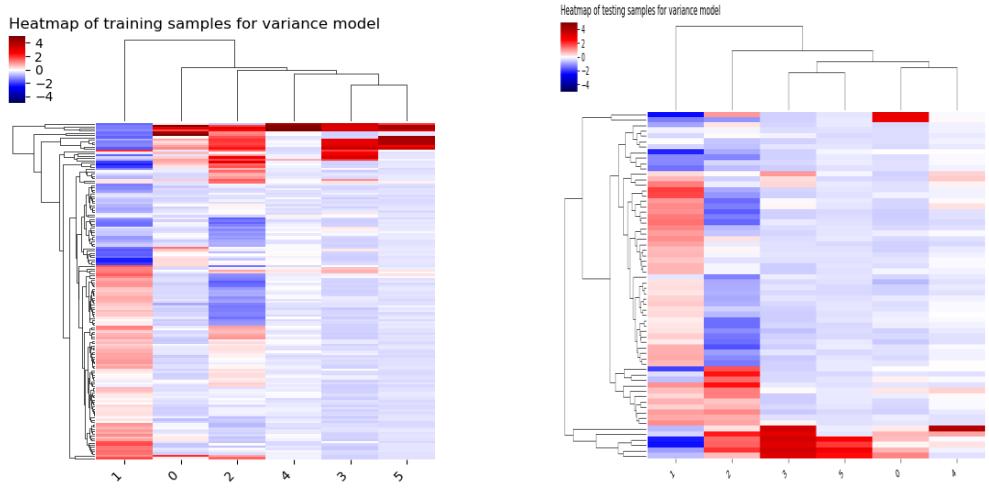

Figure 23 Second feature screening training set and test set hierarchical clustering heat map

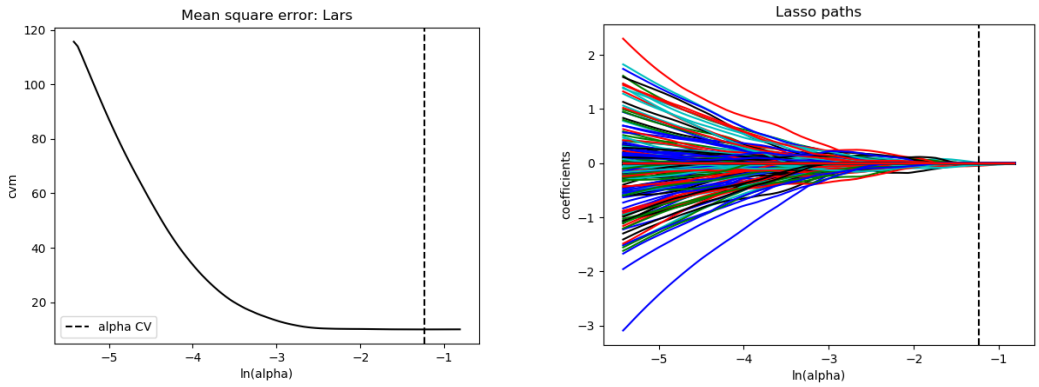

Figure 24 Relationship between penalty coefficients and features

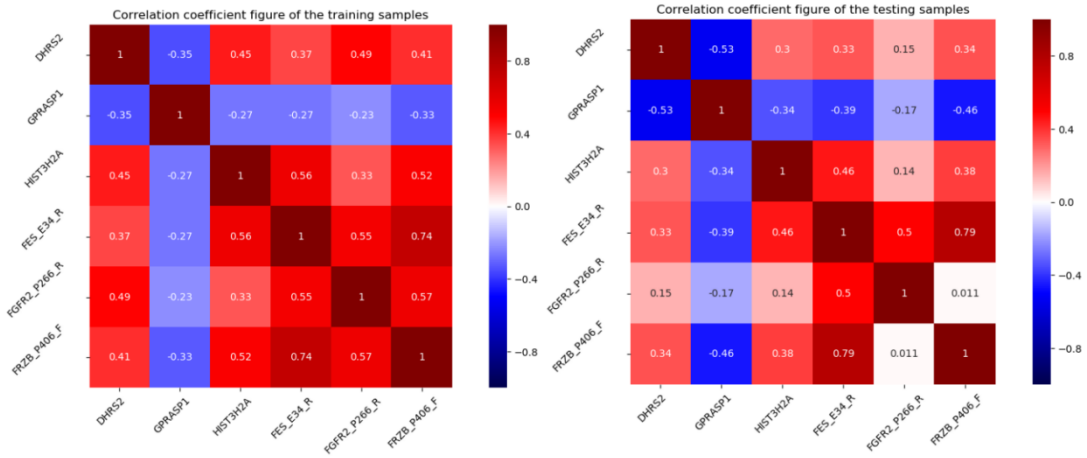

Figure 25 Relationship between features and features

Finally, a total of 6 features ['DHRS2', 'GPRASP1', 'HIST3H2A', 'FES_E34_R', 'FGFR2_P266_R', 'FRZB_P406_F'] (not the best feature selection method) were selected and saved in the user-specified location, such as Figure 26.

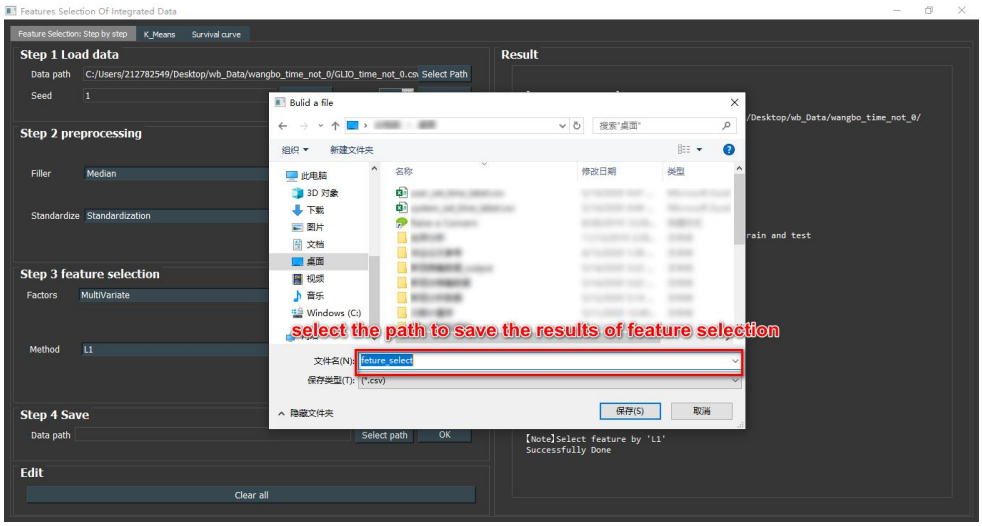

Figure 26 Save the result of feature selection

3.4.2 K-Means cluster

The user can select a number between 3-9 as the k value of the cluster. The tool takes 4 as an example and saves the result to the user's desktop location. At the same time, the user can also let the system help select the K value and save the result to the user's desktop location. When the system helps to select the k value, the optimal k value will be determined through the correlation coefficient, as shown in Figure 27.

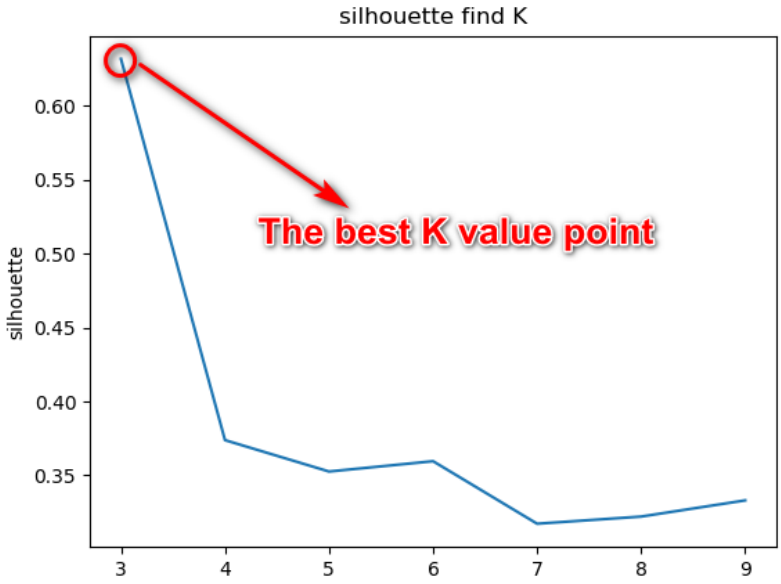

Figure 27 The system selects the k value through the sihouette method

3.4.3 Survival analysis

The user enters the relevant clustering result data and clicks the analysis button to obtain a survival analysis chart. The overall test result is shown in Figure 28. Figure 29 shows the survival analysis chart of the system selecting the optimal K value, and Figure 30 shows the survival analysis chart of the user selecting the K value.

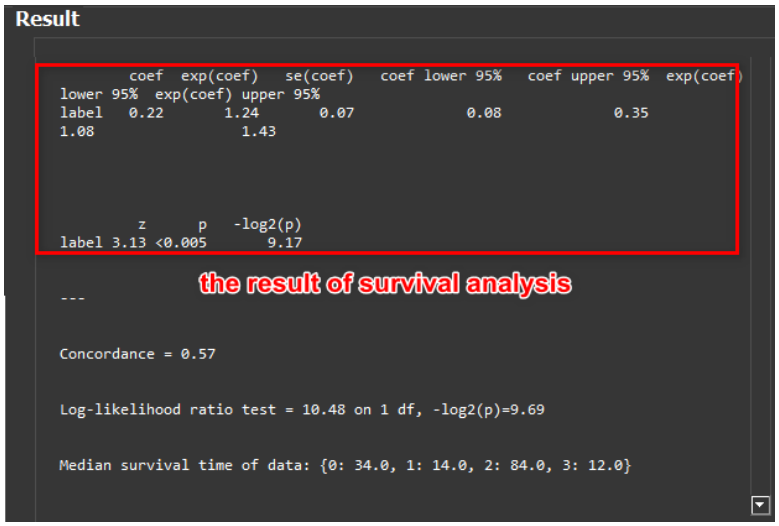

Figure 28 Survival analysis test results

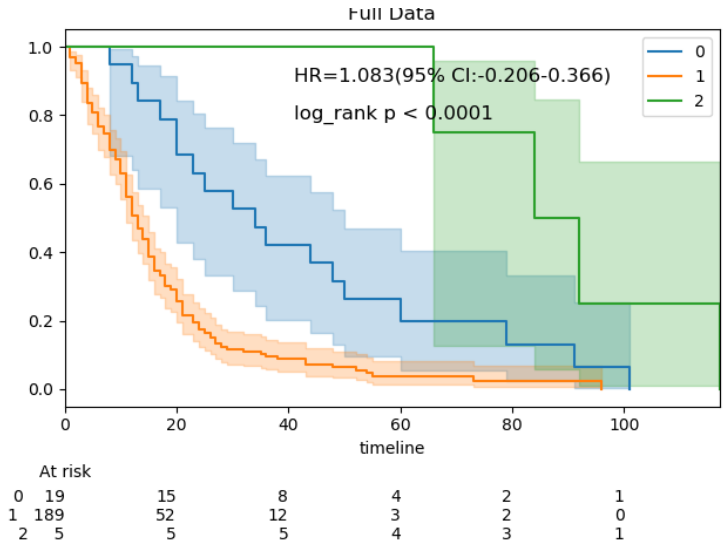

Figure 29 The survival analysis diagram of the system selecting the optimal K value

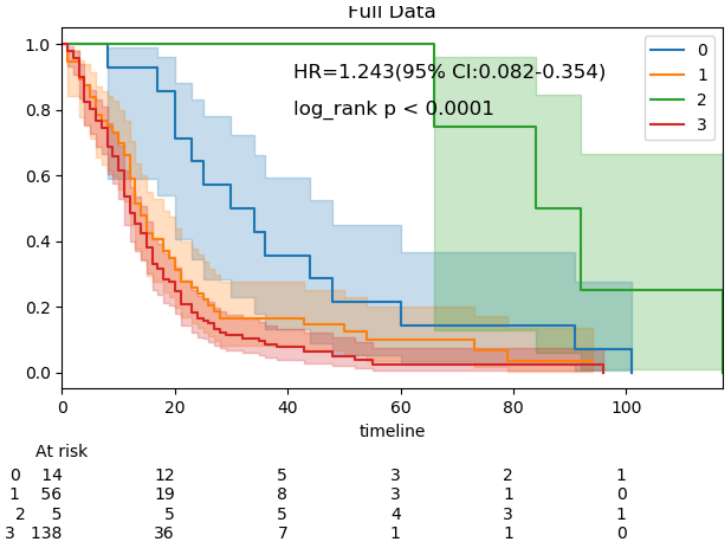

Figure 30 Survival analysis chart of user selecting K value

Chapter 5. References

Taosheng Xu, Thuc Duy Le, Lin Liu2, Ning Su. , et al. (2017) CancerSubtypes: an R/Bioconductor package for molecular cancer subtype identification, validation and visualization. Bioinformatics, 33(19), 2017, 3131–3133.

Yuan, Y., Savage, R. S., and Markowetz, F. (2011). Patient-specific data fusion defines prognostic cancer subtypes. PLoS Comput. Biol. 7:e1002227.

Aure, M. R., Steinfeld, I., Baumbusch, L. O., Liestøl, K., Lipson, D., Nyberg, S., et al. (2013). Identifying in-trans process associated genes in breast cancer by integrated analysis of copy number and expression data. PLoS ONE 8:e53014.
Ashley,E.A. (2015) The precision medicine initiative: a new national effort. JAMA, 313, 2119–2120.

Collins,F.S. and Varmus,H. (2015) A new initiative on precision medicine. N. Engl. J. Med., 372, 793–795.

Xu Aodan. (2019)Fusion model based on high-order path similarity network and multi-omics data integration analysis method [D] .South China University of Technology.

David,C.R. (1972) Regression models and life tables (with discussion). J. R Stat. Soc., 34, 187–220.

Rousseeuw,P.J. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math., 20, 53–65.

Statist. Med. (1998) The Lasso method for variable selection in the cox model.J.Roberttibshirani.
Yang K.(2019) A multidimensional nomogram combining overall stage, dose volume histogram parameters and radiomics to predict progression-free survival in patients with locoregionally advanced nasopharyngeal carcinoma.Oral Oncology ,98,85-91.