University of Macau

Faculty of Science and Technology



Visualizing Decision Rules and Relations for Medical Data Analysis: VisualizationVisualizing Decision Rules and Relations for Medical Data Analysis: Visualization

by

Song HanbaiSong Hanbai, Student No: DB627233DB627233

Final Project Report submitted in partial fulfillment  
of the requirements of the Degree of   
Bachelor of Science in Computer Science

Project Supervisor

Prof. Simon FONG & Prof. Shirley SIU (University of Macau, Macau SAR)

04 June 2020

# 

DECLARATION

I sincerely declare that:

1. I and my teammates are the sole authors of this report,
2. All the information contained in this report is certain and correct to the best of my knowledge,
3. I declare that the thesis here submitted is original except for the source materials explicitly acknowledged and that this thesis or parts of this thesis have not been previously submitted for the same degree or for a different degree, and
4. I also acknowledge that I am aware of the Rules on Handling Student Academic Dishonesty and the Regulations of the Student Discipline of the University of Macau.

Signature : \_\_\_\_Song Hanbai\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Name : Song Hanbai

Student No. : DB627233

Date : 04 June 2020

ACKNOWLEDGEMENTS

I express my heartfelt thanks to UM, because UM provides an opportunity to implement a project to partially meet the requirements of a Bachelor of Science degree.

Throughout the project, I was fortunate to have the guidance and encouragement from my supervisors, Prof. Simon FONG & Prof. Shirley SIU. They gave me a lot of very useful suggestions and used their professional perspective to help me. More importantly, they gave me a lot of encouragement, giving me the confidence to complete the goal.

**ABSTRACT**

The ultimate goal of the project is to make it easy to use. This project will help doctors or other users to conduct research more conveniently and use the interactive interface to get the information they want.

The core technology in the project is Echarts, an open source JavaScript library. The main body of the project is two Java programs, read the data in the csv file, and process the data into a format that meets the requirements of Echarts, and finally generate an html file containing interactive charts.

The project made two interactive charts, input chart and output chart. The input chart visualizes the data before the machine learning modelling and makes a line graph, so that the user can understand the characteristics of the data in advance. The output chart processes the modelled data and is a relationship diagram. It is convenient for users to understand the connection and influence of different factors.

TABLE OF CONTENTS

[University of Macau 1](#_Toc42198898)

[Faculty of Science and Technology 1](#_Toc42198899)

[Visualizing Decision Rules and Relations for Medical Data Analysis: Visualization 1](#_Toc42198900)

[by 1](#_Toc42198901)

[Song Hanbai, Student No: DB627233 1](#_Toc42198902)

[DECLARATION 2](#_Toc42198903)

[ACKNOWLEDGEMENTS 3](#_Toc42198904)

[ABSTRACT 4](#_Toc42198905)

[TABLE OF CONTENTS 5](#_Toc42198906)

[LIST OF TABLES 7](#_Toc42198907)

[LIST OF FIGURES 8](#_Toc42198908)

[CHAPTER 1. INTRODUCTION 9](#_Toc42198909)

[1.1 Project Introduction 9](#_Toc42198910)

[1.2 Group Introduction 10](#_Toc42198911)

[1.3 Project Schedule 10](#_Toc42198912)

[CHAPTER 2. RELATED WORK 12](#_Toc42198913)

[2.1 Advantages of data visualization 12](#_Toc42198914)

[2.1.1 Move Faster 12](#_Toc42198915)

[2.1.2 Discuss the results in a constructive way 12](#_Toc42198916)

[2.1.3 Understand the connection between operations and results 12](#_Toc42198917)

[2.1.4 Accept emerging trends 13](#_Toc42198918)

[2.1.5 Interact with data 13](#_Toc42198919)

[2.1.6 Create new discussion 13](#_Toc42198920)

[2.1.7 Machine learning 13](#_Toc42198921)

[2.2 Data visualization process 14](#_Toc42198922)

[2.2.1 Data collection 14](#_Toc42198923)

[2.2.2 Data processing and transformation 15](#_Toc42198924)

[2.2.3 Visual mapping 15](#_Toc42198925)

[2.2.4 Human-computer interaction 16](#_Toc42198926)

[2.2.5 User perception 17](#_Toc42198927)

[2.3 Data visualization methods 17](#_Toc42198928)

[2.3.1 FineReport 18](#_Toc42198929)

[2.3.2 Power Map 18](#_Toc42198930)

[2.3.3 Dy data 18](#_Toc42198931)

[2.4 Echarts 18](#_Toc42198932)

[2.4.1 The advantages of echarts 19](#_Toc42198933)

[2.4.2 Possible types of images 19](#_Toc42198934)

[2.5 Bayesian Network 22](#_Toc42198935)

[2.5.1 Bayesian network structure 22](#_Toc42198936)

[2.5.2 Conditional probability table 23](#_Toc42198937)

[2.5.3 Benefits of Bayesian networks 24](#_Toc42198938)

[CHAPTER 3. Functional Specification 25](#_Toc42198939)

[3.1 Inputdata.java 25](#_Toc42198940)

[3.2 Htmlpage.java 25](#_Toc42198941)

[CHAPTER 4. Software Design Specification 26](#_Toc42198942)

[4.1 relationships among modules 26](#_Toc42198943)

[4.2 Interactive interface description 26](#_Toc42198944)

[4.2.1 Input data chart 26](#_Toc42198945)

[4.2.2 Output data chart 28](#_Toc42198946)

[CHAPTER 5. Implementation Narrative and Description 31](#_Toc42198947)

[CHAPTER 6. System Quality 32](#_Toc42198948)

[6.1 Input data chart 32](#_Toc42198949)

[6.1.1 Test data 1: tox21-features.csv 33](#_Toc42198950)

[6.1.2 Test data 2: LungCancer\_cancerYesNo.csv 39](#_Toc42198951)

[6.2 Output data chart 41](#_Toc42198952)

[6.2.1 Test data 1: The upper half of the data 42](#_Toc42198953)

[6.2.2 Test data 2: The whole data 43](#_Toc42198954)

[CHAPTER 7. Ethics and Professional Conduct 44](#_Toc42198955)

[7.1 Data 44](#_Toc42198956)

[7.2 Code 44](#_Toc42198957)

[7.3 Contribute 44](#_Toc42198958)

[7.4 Precautions 44](#_Toc42198959)

[CHAPTER 8. SUMMARY & CONCLUSION 45](#_Toc42198960)

[REFERENCES 46](#_Toc42198961)

LIST OF TABLES

[Table 1: Project Schedule 11](#_Toc42160082)

LIST OF FIGURES

[Figure 1: Examples when the project is completed 10](#_Toc42199041)

[Figure 2: Echarts line chart sample 20](#_Toc42199042)

[Figure 3: Echarts bar chart sample 20](#_Toc42199043)

[Figure 4: Echarts Radar chart sample 21](#_Toc42199044)

[Figure 5: Echarts Heatmap sample 21](#_Toc42199045)

[Figure 6: 3D charts sample 22](#_Toc42199046)

[Figure 7: A simple Bayesian network 23](#_Toc42199047)

[Figure 8: Simple conditional probability table 23](#_Toc42199048)

[Figure 9: address read 25](#_Toc42199049)

[Figure 10: flow charts 26](#_Toc42199050)

[Figure 11: Initial state of input chart 26](#_Toc42199051)

[Figure 12: Move the mouse to display detailed information 27](#_Toc42199052)

[Figure 13: Click the button to show or hide the curve 27](#_Toc42199053)

[Figure 14: Initial state of output chart 28](#_Toc42199054)

[Figure 15: Drag node to move 29](#_Toc42199055)

[Figure 16: Click the button to show or hide nodes of the same type 29](#_Toc42199056)

[Figure 17: Use mouse scroll wheel to zoom chart 30](#_Toc42199057)

[Figure 18: Click the reset button to reset 30](#_Toc42199058)

[Figure 19:input chart test data 1: tox21-features.csv 38](#_Toc42199059)

[Figure 20: input chart test data 2: LungCancer\_cancerYesNo.csv 41](#_Toc42199060)

[Figure 21:output chart test data 1: The upper half of the data 42](#_Toc42199061)

[Figure 22: details of output chart test 1 42](#_Toc42199062)

[Figure 23: output chart test data 2: The whole data 43](#_Toc42199063)

[Figure 24:details of output chart test 2 43](#_Toc42199064)

# INTRODUCTION

## Project Introduction

The supervisions of this project is Prof. Simon FONG & Prof. Shirley SIU (University of Macau, Macau SAR) and the project type is Application + Research.

This project focuses on data analysis and data visualization. Visual technology is very useful in data exploration, because human visual system can detect the structure and relationship in the image. This is sometimes referred to as visual data mining or visual mining, which makes learning from the information presented by visualization faster, and is very useful in the exploration stage, because the accurate prediction target may not be very clear, or it needs to be preliminarily determined by studying abstract data graphically. This approach is the opposite of the formal approach to model building and testing, but is ideal for searching data for unexpected or unusual relationships.

Therefore, the goal of this final year project is to analyze the vertical solution of medical data. This solution combines the advantages of various data mining technologies into a system. The system is composed of the following four parts Data recognition, feature subset selection, rule induction subsystem and visualize.

-Data identification-This subsystem converts raw data into a form suitable for further data processing. In addition, based on statistical analysis, noise and redundant data are also removed.

– Feature subset selection, responsible for selecting the best attribute set to clearly generate decision rules.[1]

-Rule induction subsystem, which uses both classic machine learning algorithms and new algorithms to be proposed. Some previous works are in [2].

-Visualize the collected knowledge in a form that humans can easily understand. Some previous work done based on decision trees is published in [3]



Figure 1: Examples when the project is completed

## Group Introduction

There are three members in this project, divided into two groups. One group is a machine learning group with two people. The other group is the data visualization group, with one person. I am responsible for data visualization, visualizing the data analyzed by the machine learning group, and forming a simple interactive diagram. I mainly use JavaScript to achieve this function. In addition, I also realized the analysis of machine learning input data to generate a graph, so that users can have a simple concept before modeling and analysis.

## Project Schedule

Table 1: Project Schedule

|  |  |
| --- | --- |
| 25-09-2019 | Introduce the goals of the project. Find and test different visualization methods. |
| 23-10-2019 | Show a visual result. Discuss users should use the charts simply and conveniently. |
| 13-11-2019 | Installed a Linux virtual machine environment and tested the GraknAI tool. Determined that the software should be as simple as possible. It should decide what kind of data the project needs. |
| 27-11-2019 | Tested Rawgraph. It was discussed that the chart should conform to the Bayesian network. Looking for JavaScript support tools. |
| Before 20-3-2020 | Continue to study Rawgraph, try to build JavaScript model, but because the mainland can not access some websites, the progress is slow. |
| 1-4-2020 | The interaction required for the project is to drag and click directly within the chart, not to operate outside the chart. Find new tools to achieve interactivity. |
| 11-4-2020 | Tried a JavaScript library called Echarts to achieve interactivity. This tool implements the required interactive function very well. It can drag and drop nodes to show or hide nodes according to the label. This beta version is currently only a static web page and cannot read external files as data. Need to implement the file reading function. |
| 20-4-2020 | The data file reading function is realized through the java program. Now I use java to generate html code and write the read csv file data to html. An html web page contains all the functions of interactive charts. Should write a user interface to let users enter the address of the data file to make the project look more professional. |
| 4-5-2020 | The user input interface is implemented. Analyze the data before machine learning modeling and analysis, and display it as a chart in order to prevent unknown errors. |
| 8-5-2020 | The analysis of the input data cannot be adapted to all data sets, so we can only analyze what we are using. |
| 22-5-2020 | Discuss the content of the report. |
| 4-6-2020 | Modify the report. |

# RELATED WORK

## Advantages of data visualization

A large number of research results show that humans can obtain information through graphics much faster than reading information through reading text, and there are many other benefits to displaying numbers in a visual form.

Data visualization refers to displaying data in the form of pie charts and other graphics. This helps users to recognize patterns faster. Interactive visualization enables decision makers to gain insight into the level of detail. This change in presentation allows users to see the facts behind the analysis.

### Move Faster

The human brain processes visual information much more easily than written information. Using diagrams to summarize complex data ensures that understanding of relationships is faster than those of confusing reports or spreadsheets.

This provides a very clear way of communicating, enabling business leaders to understand and process their information faster. Big data visualization tools can provide real-time information, making it easier for stakeholders to evaluate the entire enterprise. Faster adjustments to market changes and rapid identification of new opportunities are the competitive advantages of each industry.

### Discuss the results in a constructive way

Many business reports submitted to senior management are standardized documents that are often exaggerated by static forms and various chart types. It is precisely because it is too detailed to make those executives unable to remember these contents, so it is not necessary for them to see too detailed information.

However, reports from big data visualization tools allow us to represent complex information with some short graphics, even a single graphic. Decision makers can easily interpret a variety of different data sources through interactive elements and new visualization tools like heatmaps, fever charts, and more. Rich but meaningful graphics help busy executives and business partners understand issues and pending plans.

### Understand the connection between operations and results

One benefit of data visualization is that it allows users to track the connection between operations and overall business performance. In a competitive environment, finding correlations between business functions and market performance is critical.

For example, the executive sales director of a software company may immediately see in the bar chart that sales of their flagship product in the Southwest region have fallen by 8%. Supervisors can then gain insight into where these differences occur and start planning. In this way, data visualization allows managers to immediately spot problems and take action.

### Accept emerging trends

The amount of data on consumer behavior that has been collected can open up many new opportunities for adaptable companies. However, this requires them to continuously collect and analyze this information. By using big data visualizations to monitor key indicators, leaders can more easily spot market changes and trends across a variety of big data sets.

### Interact with data

The main benefit of data visualization is that it brings risk changes in a timely manner. But unlike static charts, interactive data visualization encourages users to explore or even manipulate data to discover other factors. This provides better insights into usage analysis.

For example, large data visualization tools can show boat manufacturers the decline in sales of their large processes. This may be due to a number of reasons. But team members are actively exploring related issues and linking them to actual boat sales to identify root causes and find ways to reduce their impact to drive more sales.

### Create new discussion

One advantage of data visualization is that it provides a ready-made way to tell stories from data. Heat maps can show product performance developments in multiple geographic areas, making it easier for users to see products that perform well or perform poorly. This allows executives to drill down to specific locations and see what works well and what doesn't.

They may realize that higher-income market segments do not sell higher-priced products or that traditional cleaning products are less popular than environmentally friendly green products. These insights can be used to brainstorm and brainstorm to support higher sales.

Data visualization tools provide a more efficient way to use operational data. For most business leaders, changes in real-time performance and market metrics are easier to identify and respond to.

### Machine learning

All the hype surrounding machine learning is coming to fruition. In addition to some large companies, such as Amazon and Google, using machine learning to eliminate spam, Pinterest uses machine learning to show users relevant content, and Yelp uses machine learning to organize user-uploaded photos. Even companies like Disqus are using machine learning to remove spam. Get ready to start applying machine learning to your own or your customers' business areas.

Home Depot uses machine learning to help users find products faster. Even small companies like Lyst use machine learning to help customers find relevant information for any kind of query.

Customer service is being transformed by the ability of machine learning to interpret customer emails and categorize them to correct departments or areas within a company. This also means that the way of telephone communication may not be used in the future.

The future of machine learning is infinite.[4]

## Data visualization process

Most people's first impressions of data visualization may be various graphics, such as the bar chart, bar chart, line chart, pie chart, scatter chart, etc. in the Excel chart module, which are not listed one by one. The above is just a concrete manifestation of data visualization, but data visualization does not stop there.

Data visualization is not a simple visual mapping, but a complete process that takes data flow as the main line. It mainly includes data collection, data processing and transformation, visual mapping, user interaction, and user perception. A complete visualization process can be seen as a process in which the data stream passes through a series of processing modules and is transformed. The user obtains knowledge and inspiration from the visualized mapping results through visual interaction.

The modules of the visualization main process are not just purely linear connections, but there is a connection between any two modules. For example, different data collection, data processing and transformation, visualization coding and human-computer interaction methods will generate new visualization results. Users will have new knowledge and inspiration through their perception of new visualization results.

In the following, several key steps in the main process of data visualization are described.

### Data collection

Data collection is the first step in data analysis and visualization. As the saying goes, "the clever woman is hard to cook without rice", the method and quality of data collection largely determine the final effect of data visualization.

There are many classification methods for data collection. From the perspective of the data source, it can be divided into internal data collection and external data collection.

1. Internal data collection:

Refers to the collection of data on the company's internal operating activities. Usually the data comes from business databases, such as the transaction status of orders. If you want to analyze user behavior data and APP usage, you also need some behavior log data. At this time, you need to use the "buried point" method to collect APP or Web data.

2. External data collection:

The index refers to some external data of the enterprise through some methods, and the specific purposes include obtaining data of competing products and some industry data published on the official website of the official institution. To obtain external data, the commonly used data collection method is "web crawler".

The data obtained by the above two types of data collection methods are second-hand data. Collecting data through surveys and experiments belongs to first-hand data, which is more commonly used in market research and scientific research experiments, and is not included in this discussion.

### Data processing and transformation

Data processing and data transformation are the prerequisites for data visualization, including data preprocessing and data mining.

On the one hand, the data obtained through the previous data collection inevitably contains noise and errors, and the quality of the data is low; on the other hand, the characteristics and patterns of the data are often hidden in massive data, which requires further data mining to extract .

Because of the above problems, the collected data is directly analyzed or visualized, and the conclusions drawn often mislead users to make wrong decisions. Therefore, data cleaning and standardization of the collected raw data is an indispensable part of the data visualization process.

The display space for data visualization is usually two-dimensional, such as computer screens and large-screen displays. 3D graphics rendering technology solves the problem of displaying three-dimensional objects on a two-dimensional plane.

But in the era of big data, the data we collect usually has 4V characteristics: Volume, Variety, Velocity, Value. How to mine valuable information from high-dimensional, massive, and diverse data to support decision-making. In addition to cleaning and removing noise from the data, you also need to perform secondary processing on the data according to business purposes.

Common data processing methods include: dimensionality reduction, data clustering and segmentation, sampling and other statistical and machine learning methods.

### Visual mapping

After the data is cleaned, denoised, and processed according to business purposes, the next step is the visual mapping link. Visual mapping is the core of the entire data visualization process, and refers to the process of mapping processed data information into visualization elements.

The visualization element consists of 3 parts: visualization space + marker + visual channel.

1.Visualization space

The display space for data visualization is usually two-dimensional. The visualization of three-dimensional objects, through graphics rendering technology, solves the problems of displaying in two-dimensional planes, such as 3D ring diagrams, 3D maps, and so on.

2. Mark

Markers are the mapping of data attributes to visual geometric elements and are used to represent the classification of data attributes.

According to the difference in spatial degrees of freedom, markers can be divided into points, lines, faces, and volumes, which have zero degrees of freedom, one-dimensional, two-dimensional, and three-dimensional degrees of freedom. For our common scatter plots, line graphs, rectangular tree plots, and three-dimensional histograms, four different types of markers are used: point, line, area, and volume.

3. Visual channel

The mapping of data attribute values to labeled visual presentation parameters is called a visual channel and is usually used to display quantitative information about data attributes.

Commonly used visual channels include: marker position, size (length, area, volume ...), shape (triangle, circle, cube ...), direction, color (hue, saturation, brightness, transparency ...) Wait.

The four graphic examples in Figure 3 make good use of visual channels such as position, size, and color to visualize data information.

"Marking" and "visual channel" are two aspects of visual coding elements. The combination of the two can complete the visual expression of data information, thereby completing the process of visual mapping.

### Human-computer interaction

The purpose of visualization is to reflect the values, characteristics and patterns of the data, present the information behind the data to the target users in a more intuitive and easy-to-understand manner, and assist them to make correct decisions.

But usually, the data we face is complex, and the information contained in the data is rich.

If in the visual graphics, all the information is not organized and filtered, and all the machinery is placed, it will not only make the entire page look particularly bloated and confusing and lack aesthetics; but it will blur the focus, distract users, and reduce users. The ability to obtain information per unit of time.

Common interaction methods include:

1. Scrolling and zooming: When the data cannot be fully displayed on the device with the current resolution, scrolling and zooming is a very effective way to interact, such as information details of maps, line charts, etc. However, the specific effects of scrolling and zooming are not only related to the page layout, but also to the specific display device.

2. Control of color mapping: Some visualization open source tools will provide color palettes, such as D3. Users can configure the color of the visual graphics according to their preferences. This is relatively more in platform tools such as self-service analysis, but for some self-developed visualization products, professional designers are generally responsible for this work, so that the visual communication of the visualization has an aesthetic sense.

3. Control of data mapping method: This refers to the user's choice of data visual mapping elements. Generally, a data set has multiple sets of features. It provides flexible data mapping methods to users, which can facilitate users to follow the dimensions they are interested in Go explore the information behind the data. This is provided in common visual analysis tools, such as tableau, PowerBI and so on.

4. Data detail level control: such as hiding data details, hover or click only appear.

### User perception

The results of visualization can only be transformed into knowledge and inspiration after being perceived by the user.

In the process of perception, in addition to passively accepting visualized graphics, users also actively obtain information through interaction with various modules of visualization.

How to make users better perceive the results of visualization, and transform the results into valuable information to guide decision-making. There are too many influencing factors involved in this. Knowledge in multiple disciplines such as psychology, statistics, human-computer interaction, etc. .

## Data visualization methods

The methods of data visualization can be classified according to the degree of visualization and interaction:

Pure visualization chart classes: Echart, AntV, HIghcharts

Visual report class: FineReport,

Business Intelligence Analysis: Tableau, FineBI, PowerBI

Data map category: Power Map2016, Map Hui

Visualization of large screens: Ali DataV, FineReport, Digital Hail

Data mining class: R-ggplot2, Python

Other: dydata.

Some methods are to directly call the existing modules to analyze the data to form a chart, and some are to provide method and code package support. Users can write programs to achieve their own needs. Next, several methods will be briefly introduced.

### FineReport

FineReport reporting software is an enterprise-level Web reporting tool written in pure Java, which integrates data display (report) and data input (form) functions. Drag and drop operations can design complex Chinese reports and establish a data decision analysis system.

Fully functional and professional. It supports connection access to relational databases and BI multidimensional databases, supports the processing of Chinese-style complex reports, supports offline reports, multi-layer reports and data reports, and has a safe and perfect permission control scheme.

Designing reports is simple and efficient, with low learning costs. An Excel-like interface allows users to develop reports using zero coding without any additional learning costs, and can easily drag and drop data to complete the report in a minute or two.

The industry is rich in accumulation. It has its own unique insights into various industries, and can provide solutions and implement solutions for various topics such as strategy, operations, organization, finance and marketing (including from strategy to internal, top to bottom, and internal to external) Program.

It is a highly complete commercial software, as long as you enter data on the web excel, you can generate charts with one click. However, the disadvantage is that it runs slower and requires a longer time for resource loading.

### Power Map

PowerMap is a data visualization tool developed by Microsoft based on Bing maps. It can draw geographic and time data and vividly display the dynamic development trend. It is specifically for generating data maps, but since Office365 has integrated this plug-in, it is very convenient to use.

There are three main steps required to create a visual map: setting the geographic location variable, selecting the visualization chart type, and setting the display variable

There are three main steps required to create a visual map: setting the geographic location variable, selecting the visualization chart type, and setting the display variable. It can be easily realized by using the insert → three-dimensional map function in excel.

### Dy data

Dydata is actually not a data visualization tool, it is a data supply and demand platform, users provide data or release needs on this platform, some are free, some need to pay. However, it also provides visualization services, not only generating data charts, but also providing resources such as poster templates and resume templates.

## Echarts

ECharts, an open source visualization library implemented with JavaScript, can run smoothly on PC and mobile devices, compatible with most current browsers (IE8 / 9/10/11, Chrome, Firefox, Safari, etc.), the bottom layer relies on vector graphics The library ZRender provides intuitive, rich interactive and highly personalized data visualization charts.

After many failed attempts, I tried the suggestions of a team member, studied the echarts package, and found that it satisfies the needs of the project very well.

### The advantages of echarts

ECharts provides conventional line charts, bar charts, scatter charts, pie charts, K-line charts, box charts for statistics, maps, heat maps, and line graphs for geographic data visualization, for relational data visualization Relational graphs, treemaps, sunburst graphs, parallel coordinates for multidimensional data visualization, as well as funnel graphs for BI, dashboards, and support for mashup between graphs and graphs. In this project, I used the relationship diagram and line chart.

Echarts provides interactive components such as legend, visual mapping, data area zoom, tooltip, and data brushing out of the box. It can perform interactive operations such as multi-dimensional data filtering, view zooming, and displaying details. These interactive components are exactly what I need most.

In addition to the correct interactive functions, ECharts has also optimized mobile interaction in detail. For example, the small screen on the mobile terminal is suitable for zooming and panning in the coordinate system with fingers. On the PC side, you can also use the mouse to zoom (using the mouse wheel), pan, and so on.The fine-grained modularization and packaging mechanism makes ECharts smaller on mobile terminals, while the optional SVG rendering module makes mobile memory no longer occupy space.

In addition to PC and mobile browsers, ECharts can also cooperate with node-canvas on node to perform efficient server-side rendering (SSR). Since 4.0, we have also cooperated with the WeChat Mini Program team to provide ECharts adaptation to Mini Programs!

### Possible types of images

Echarts supports many different types of image formats.

#### Line Chart

The line chart can not only indicate the quantity, but also reflect the development and changes of the same thing at different times. Line charts are very commonly used in life. Although it does not directly give accurate data, as long as you have a certain skill and skilled use of the "coordinate method", you can quickly determine a specific data.

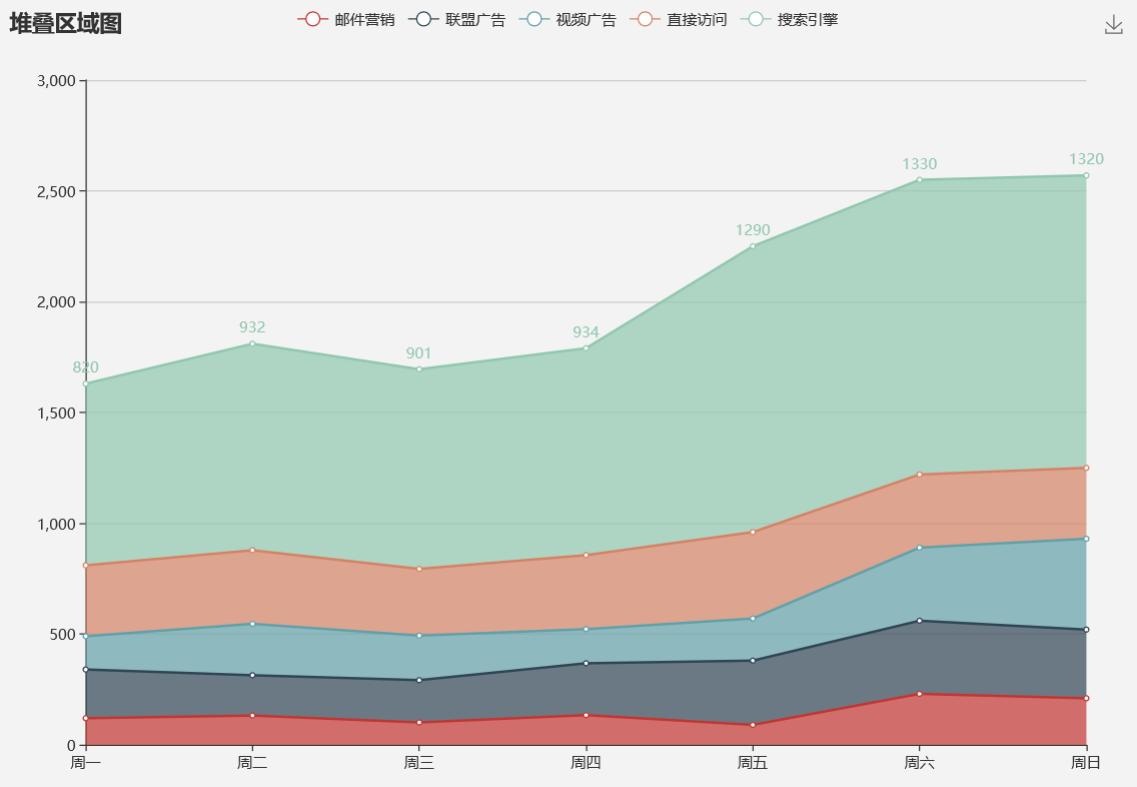


Figure 2: Echarts line chart sample

#### Bar Chart

Bar charts can clearly indicate how many different quantities. It is easy to compare the differences between the data.

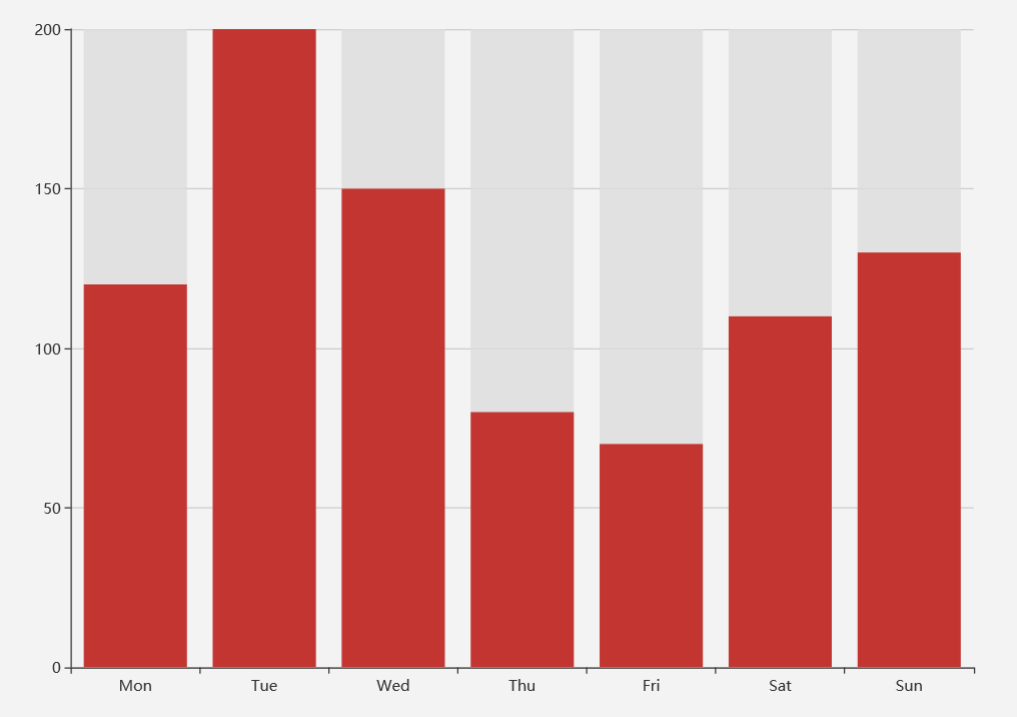


Figure 3: Echarts bar chart sample

#### Radar chart

The radar chart is mainly used to evaluate the profitability, productivity, liquidity, safety and business growth of the business. The combination of the above indicators is named because it is very similar to the shape of the radar.



Figure 4: Echarts Radar chart sample

#### Heatmap

The heat map can display the visitor’s passion page area and the visitor’s geographic area in a particularly prominent format to show what is happening in the non-clickable area. With the help of heat maps, you can intuitively and clearly see the visitors’ interest in each area of the page without having to report on the experience of data analysis, graphical display or page analysis.

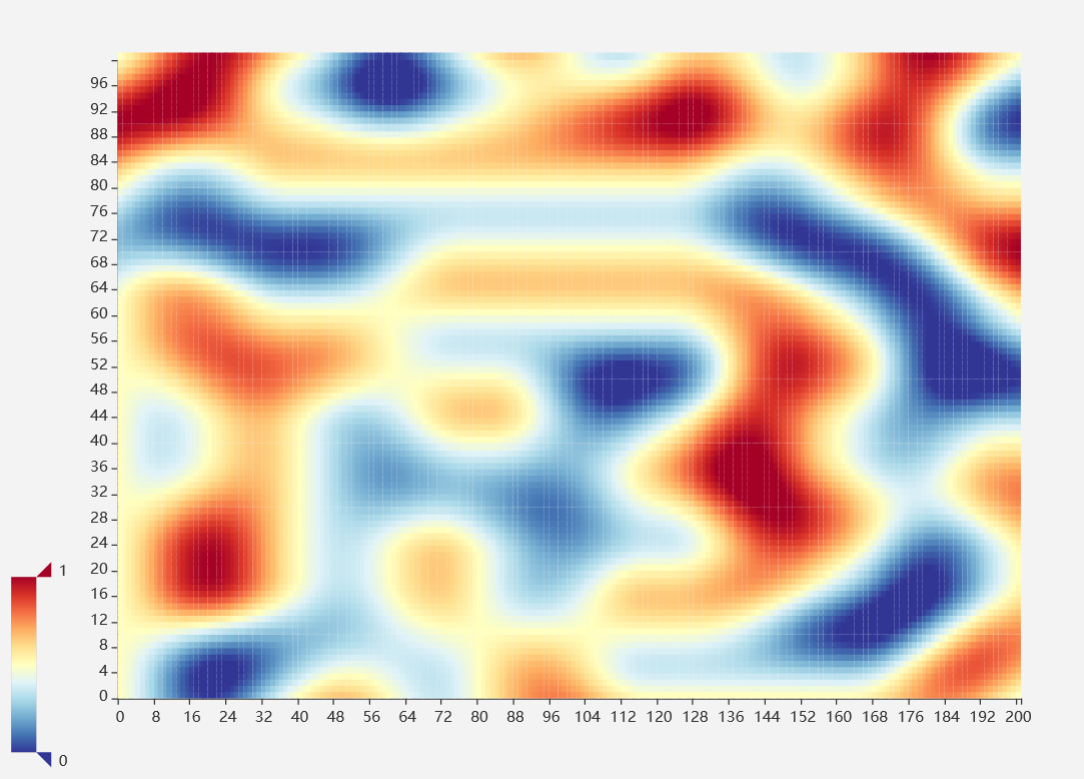


Figure 5: Echarts Heatmap sample

#### 3D Charts

Echarts does not only have 2d charts, but also corresponding 3d versions. Here use a 3d bar chart as an example.

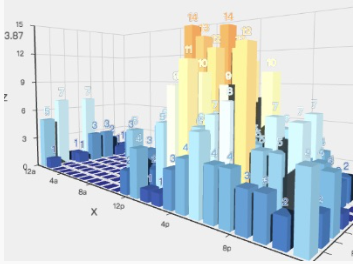


Figure 6: 3D charts sample

## Bayesian Network

Bayesian Network is the core theory of the project. Bayesian Networks, also known as Belif Networks or Causal Networks, is a graphical model that describes the dependencies between data variables and is a model for inference. Bayesian networks provide a convenient framework structure for people to express causality, which makes uncertainty reasoning mo For the Bayesian network, we can look at it two ways: First, Bayesian networks expressed conditional independence relationships between individual nodes, we can intuitively draw independent conditions between the property from them and rely on Bayesian Network relations; in addition Bayesian network can be considered another form expressed by the joint probability distribution of events, we can quickly get each basic event (all property values according to the Bayesian network structure and network conditional probability table (CPT) the probability of a combination). Bayesian learning theory prior knowledge and sample data to obtain an estimate of the unknown sample, the probability (including joint and conditional probabilities) is the a priori information and sample forms of learning information data among the Bayesian theory.re logically clear and understandable.

### Bayesian network structure

The network structure of the Bayesian network is a Directed Acyclic Graph, where each node represents an attribute or data variable, and the arc between the nodes represents the probability dependency relationship between the attributes (data variable). An arc points from one attribute (data variable) A to another attribute (data variable) B. The value of attribute A can affect the value of attribute B. Since it is a directed acyclic graph, it will not appear between A and B. Directed loop. In the Bayesian network, the direct cause node (arc end) A is called its result node (arc head) B's parents (parents), and B is called A's child node. If there is a directional path from a node X to Y, then node X is called the ancestor of node Y, and node Y is also called the descendant of node X.

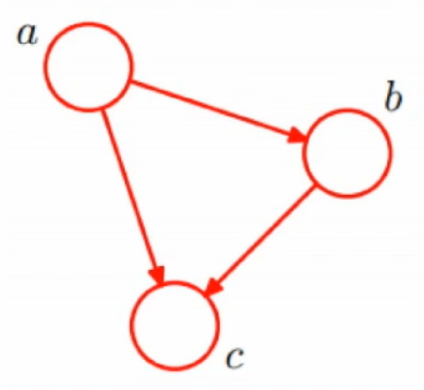


Figure 7: A simple Bayesian network

### Conditional probability table

A conditional probability table in a Bayesian network is a collection of conditional probabilities of nodes. When using Bayesian network for inference, it is actually a process of using the prior probability in the conditional probability table and the known evidence nodes to calculate the posterior probability of the queried target node.

Conditional probabilities can be given by experts in some fields by summing up their past experience. Another method is to use a conditional probability formula to statistically obtain large data.

If the snow at node A1 is used as the evidence node, what is the probability of traffic jam at A2? The following table gives the corresponding conditional probabilities:

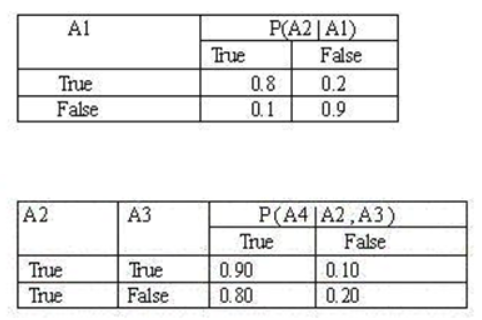


Figure 8: Simple conditional probability table

### Benefits of Bayesian networks

Bayes network makes it easy to describe the relationship between data by means of graphics, with clear semantics and easy to understand. The graphical knowledge representation method makes it easy to keep the consistency and integrity of the probability knowledge base, and can easily reconfigure the network module for the change of conditions.

Bayesian network is easy to deal with incomplete data sets. For the traditional standard supervised learning algorithm, it is necessary to know that all possible data inputs, if one of them is missing, will produce biased difference to the established model. The Bayes network method reflects the probability relationship model between the data in the whole data base, and the lack of a certain data variable can still be built Establish a precise model.

Bayesian networks allow learning causal relationships among variables. In the past data analysis, when the causal relationship of a problem has more interference, the system can not make accurate prediction. And this causality has been included in Bayesian network model. Bayes method has the semantics of causality and probability. It can be used to learn causality in data and learn according to causality.

The combination of Bayesian network and Bayesian statistics can make full use of the information of domain knowledge and sample data. Bayes network uses arcs to represent the dependence among variables, and uses probability distribution table to represent the strength or weakness of dependence relation system. It combines prior information with sample knowledge organically and promotes the integration of prior knowledge and data, which is especially effective when sample data is sparse or hard to obtain .

# Functional Specification

The program of the project is mainly two Java files, and they complete the analysis and visualization of input data and output data through similar cores.

## Inputdata.java

inputdata.java reads the feature values of input data and generates a webpage graph to let users intuitively recognize the features of input data.

Because each data set has different specifications and different requirements, this program cannot adapt to all data sets. Here I only configured the parameters for the test data set. The program reads a column in the data set, counts the number of positive and negative numbers and represents it with two curves, so that the user can understand the data characteristics in advance.

## Htmlpage.java

htmlpage. Java reads the data files node.csv and link.csv that have been modeled and analyzed by the machine learning group. The program can be adapted to any csv file with correct format.

After running the program, a window will pop up for entering the addresses of two csv files. After reading the file content correctly, the program will generate an html webpage based on the content, where the content is a relationship diagram. Users can intuitively see the relationship between different factors, and move nodes or save pictures as needed.

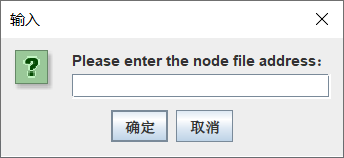


Figure 9: address read

# Software Design Specification

## relationships among modules

The core of the entire project mainly uses 4 technologies: echarts, java generates files, reads csv data and user windows.

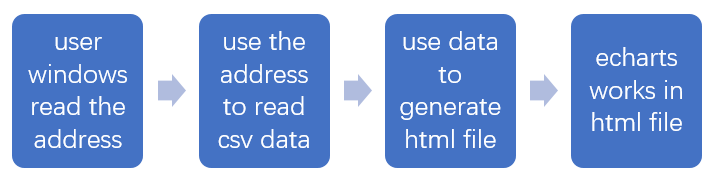


Figure 10: flow charts

## Interactive interface description

The echarts chart provides interactive functions. The two charts have some common contents and differences.

### Input data chart

The original state of the input graph shows two curves of positive and negative numbers. The user can put the mouse on the circular icon of the curve to get detailed information, and also click on the ‘positive’ and ‘negative’ at the top to select the curve to be displayed and hidden.

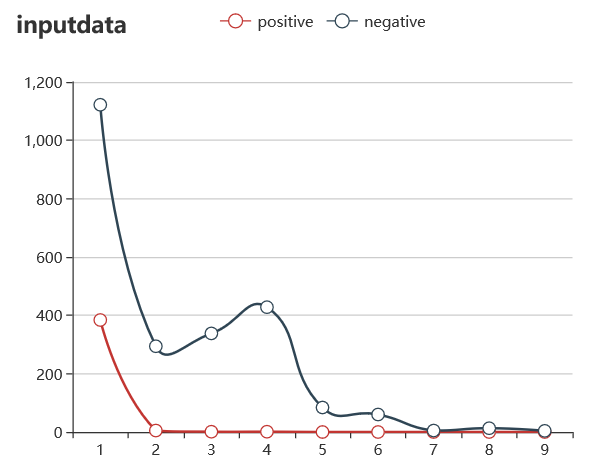


Figure 11: Initial state of input chart

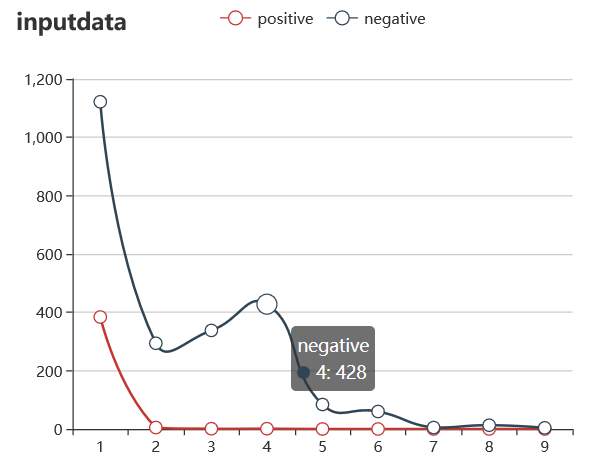


Figure 12: Move the mouse to display detailed information



Figure 13: Click the button to show or hide the curve

### Output data chart

The output graph is a relationship graph that contains more functions. The user can drag and drop the node to the desired position, click the category button above to display and hide the same type of node, use the mouse wheel to zoom the chart, and click the circle button on the upper right to reset the chart. The download button next to the reset button can download the current chart as a picture to the local.

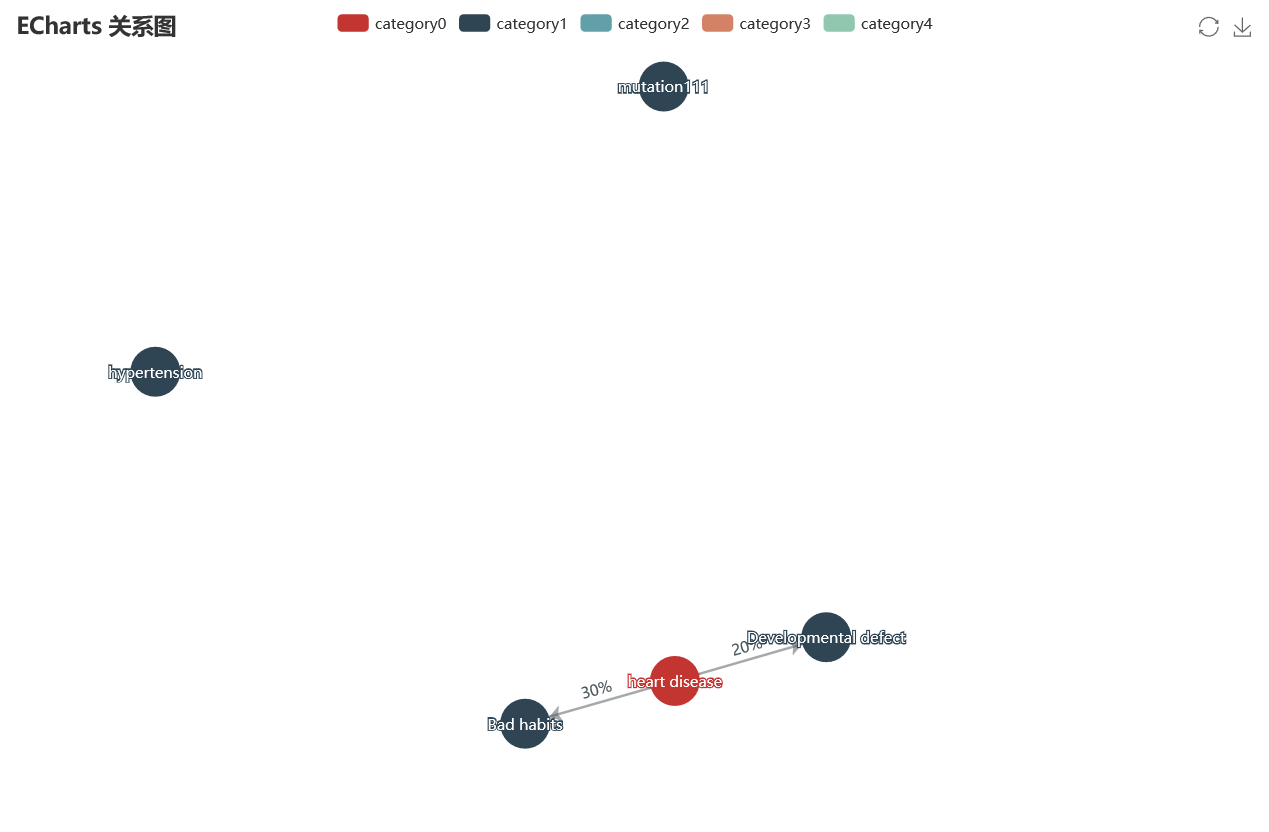


Figure 14: Initial state of output chart

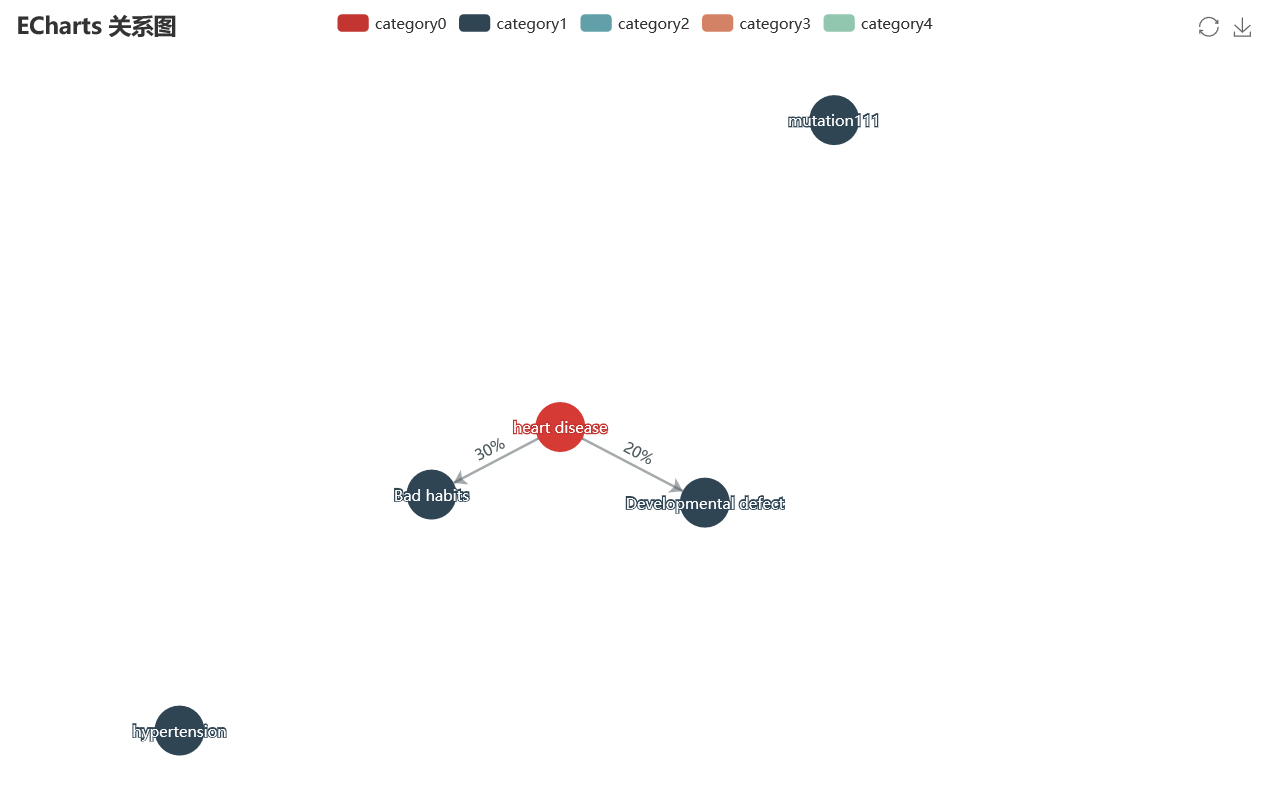


Figure 15: Drag node to move

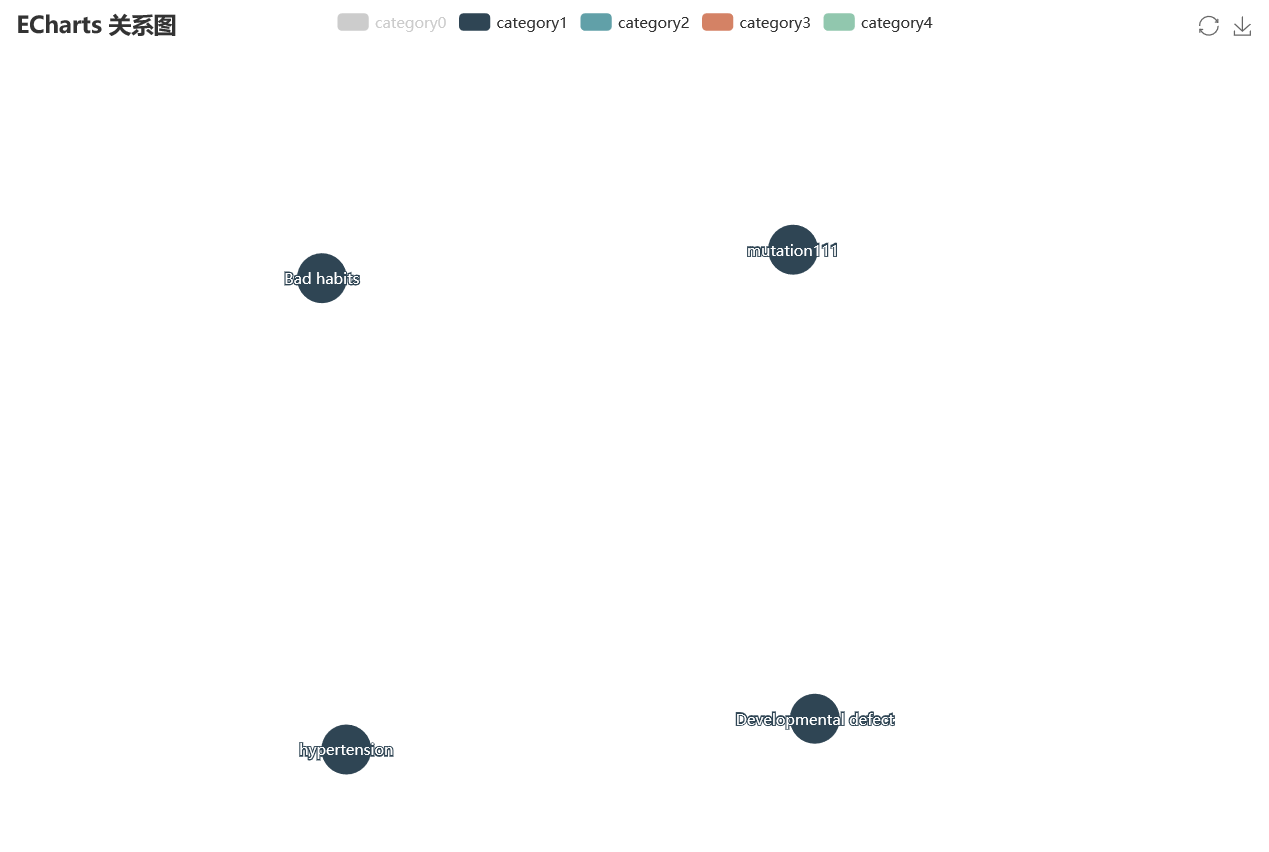


Figure 16: Click the button to show or hide nodes of the same type

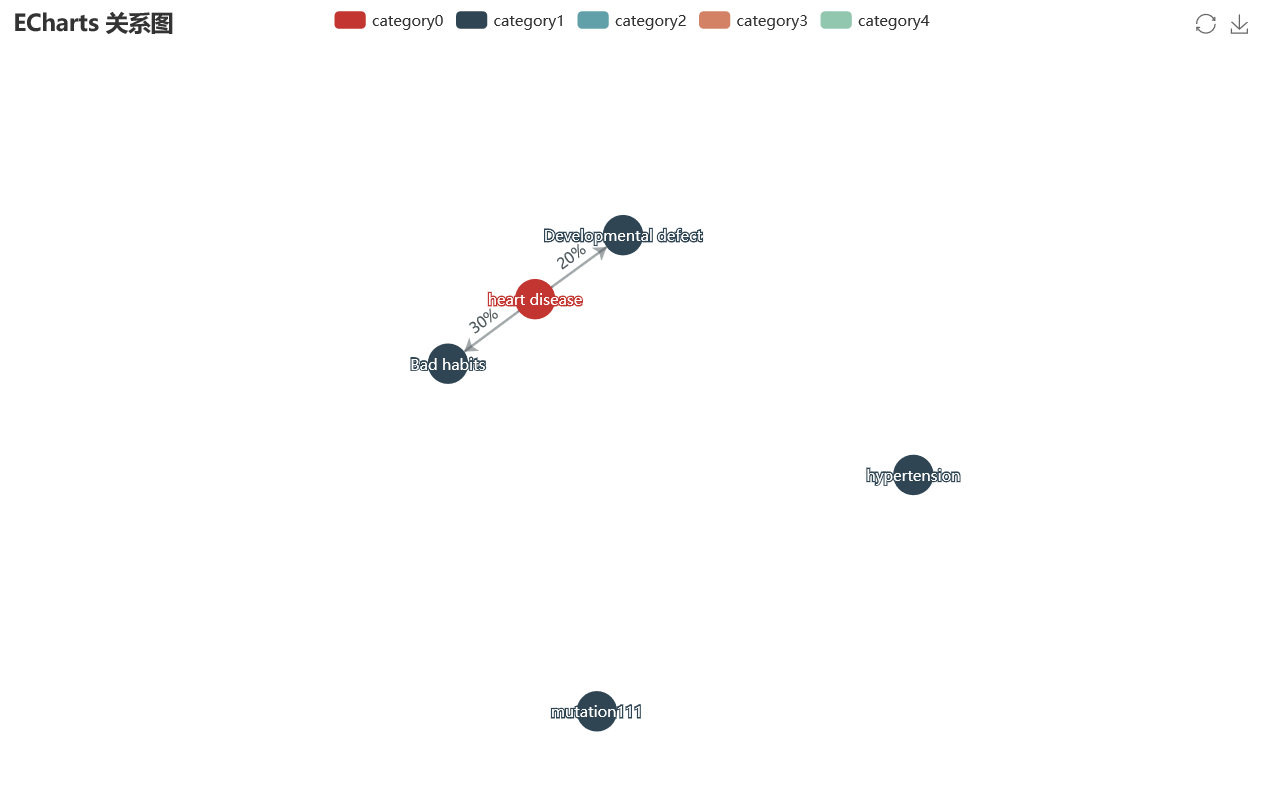


Figure 17: Use mouse scroll wheel to zoom chart

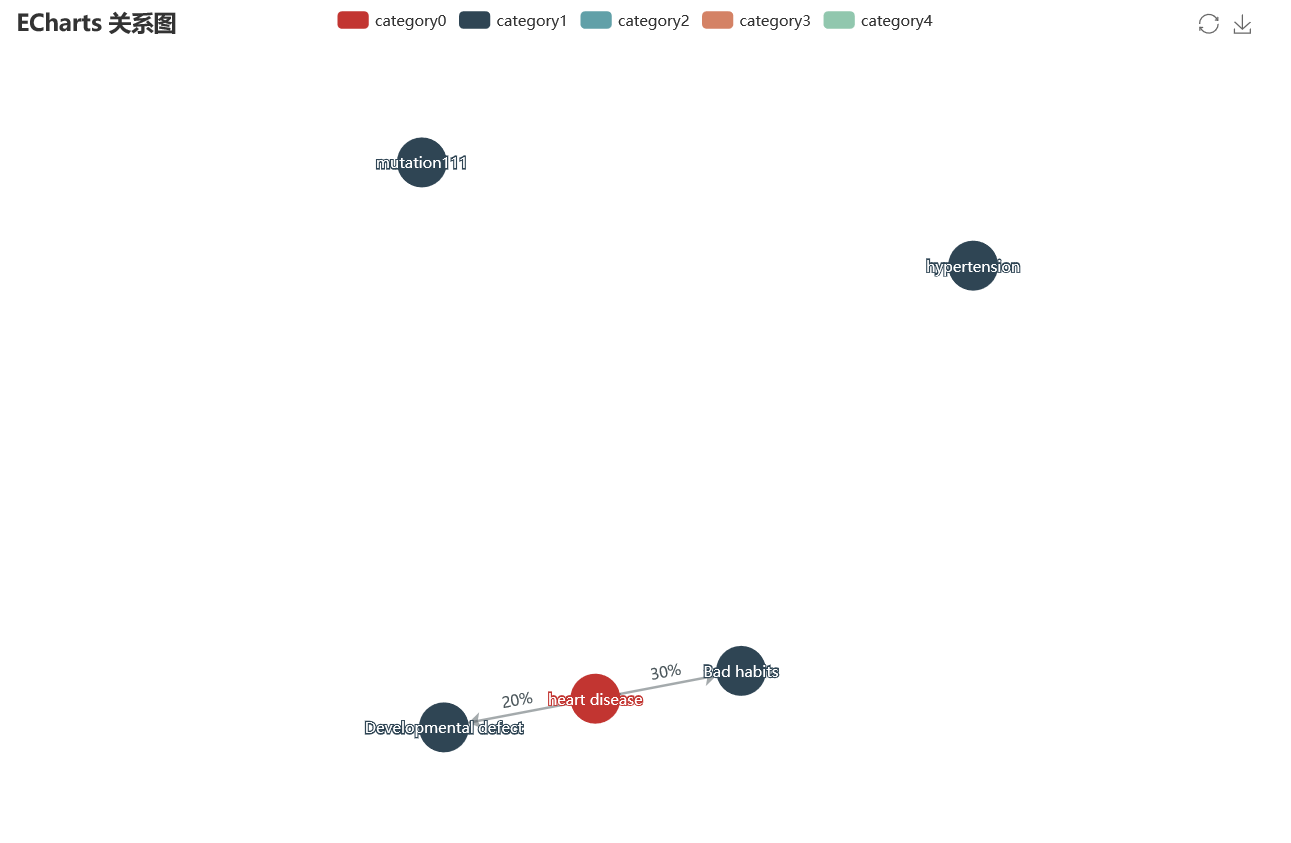


Figure 18: Click the reset button to reset

# Implementation Narrative and Description

Originally after deciding to use the echarts component, I planned to use a dynamic web page to realize the visual operation directly. However, echarts has very strict requirements on the code format. Using JavaScript alone cannot read the file data, so I decided to use another language to generate an html file. Because I have experience in generating web pages using Java, I chose Java.

Echarts needs two kinds of data: data and links, so I chose to read two data files separately. Use csv format because it is more commonly used and convenient.

A user window was written to enter the file address in order to look more professional.

# System Quality

Test the chart effect by changing the read data.

## Input data chart

The test data used here are from tox21-features.csv and LungCancer\_cancerYesNo.csv.

This shows all the data curves in the dataset. I arranged the graph in the html file in the format of 6\*3, but considering the size limitation and clarity, I changed the screenshot to the form of 3\*3.

### Test data 1: tox21-features.csv

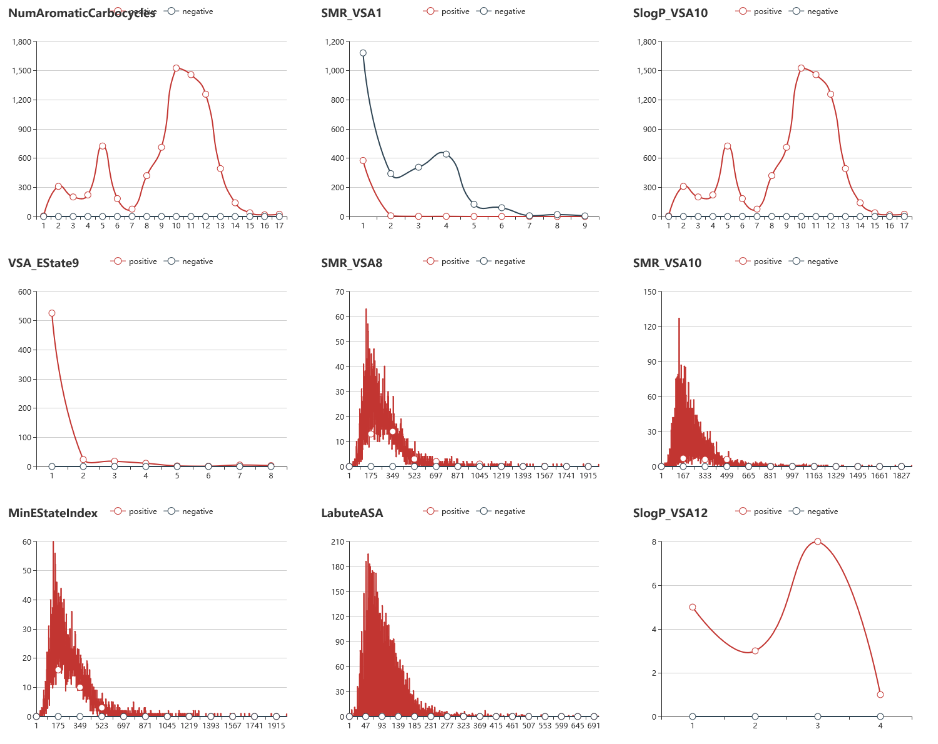
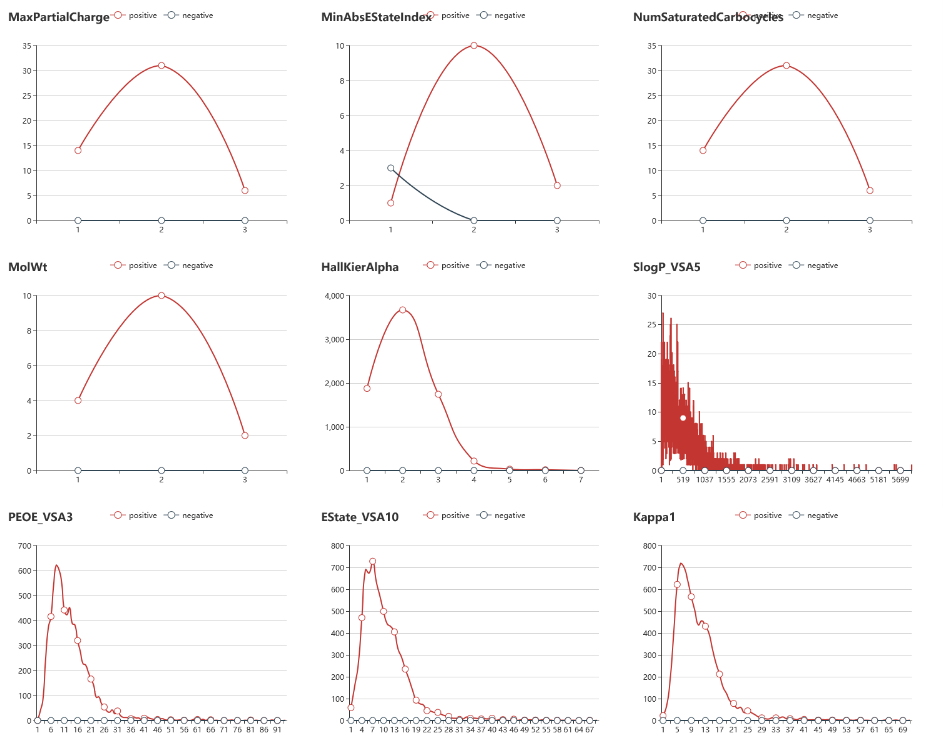
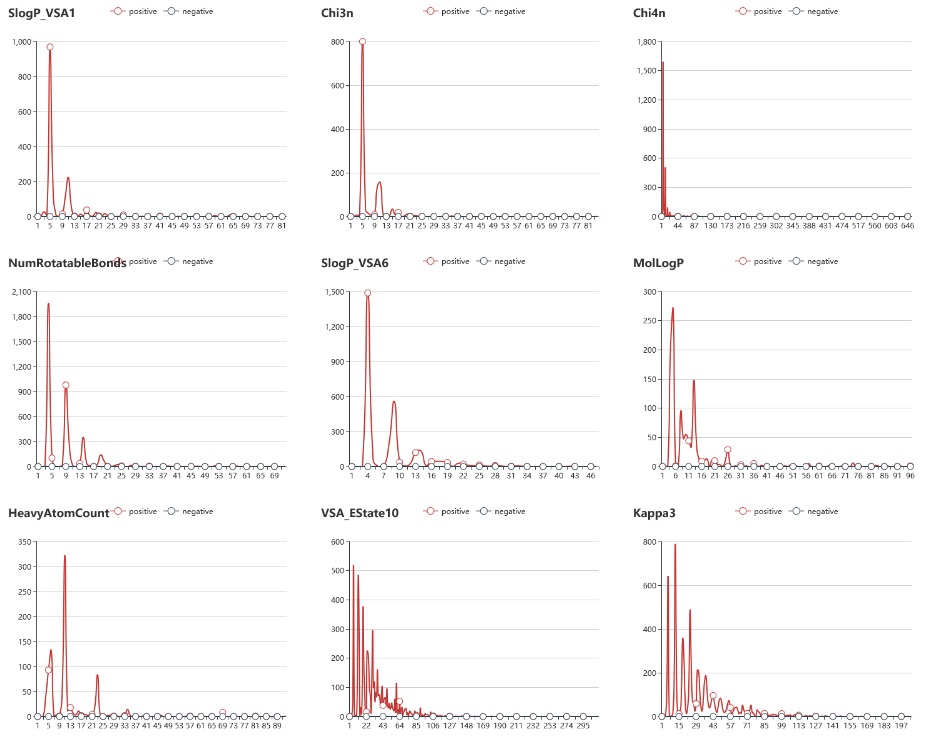
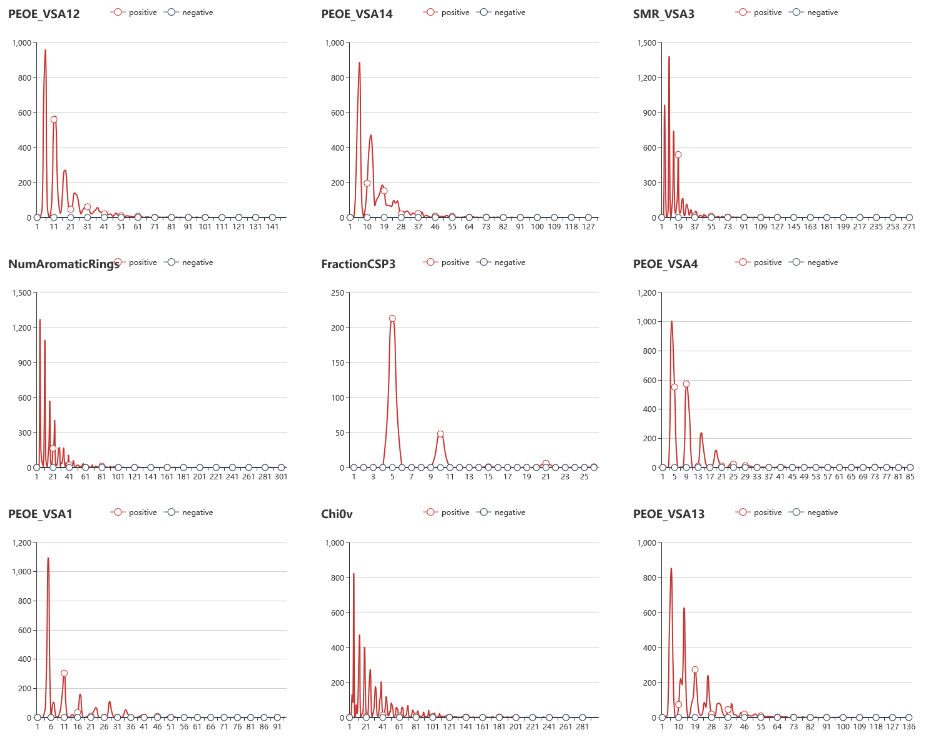
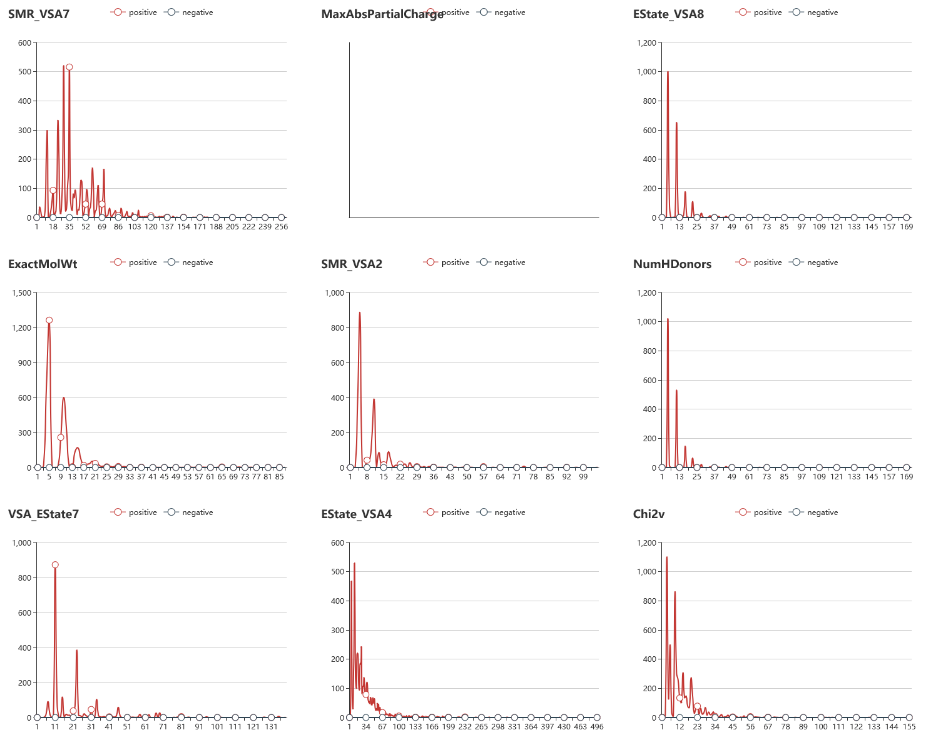
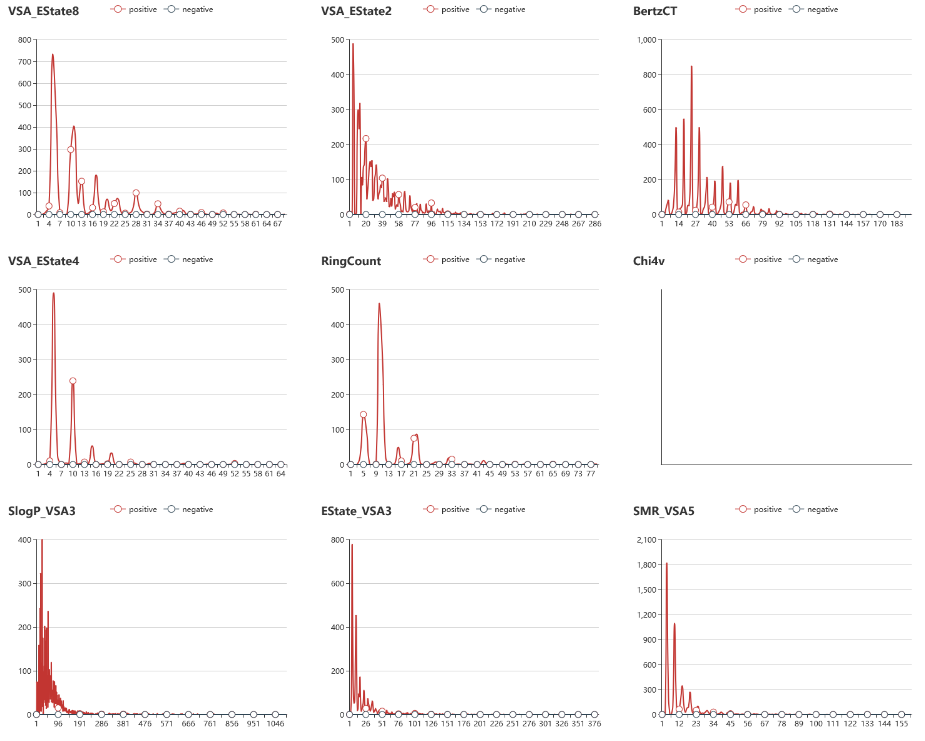
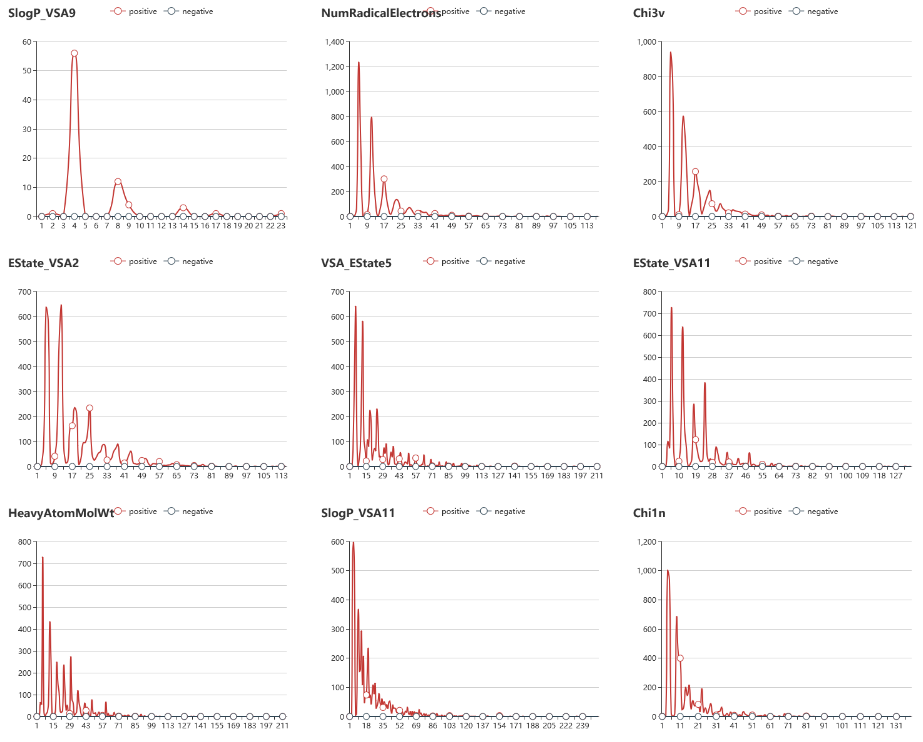
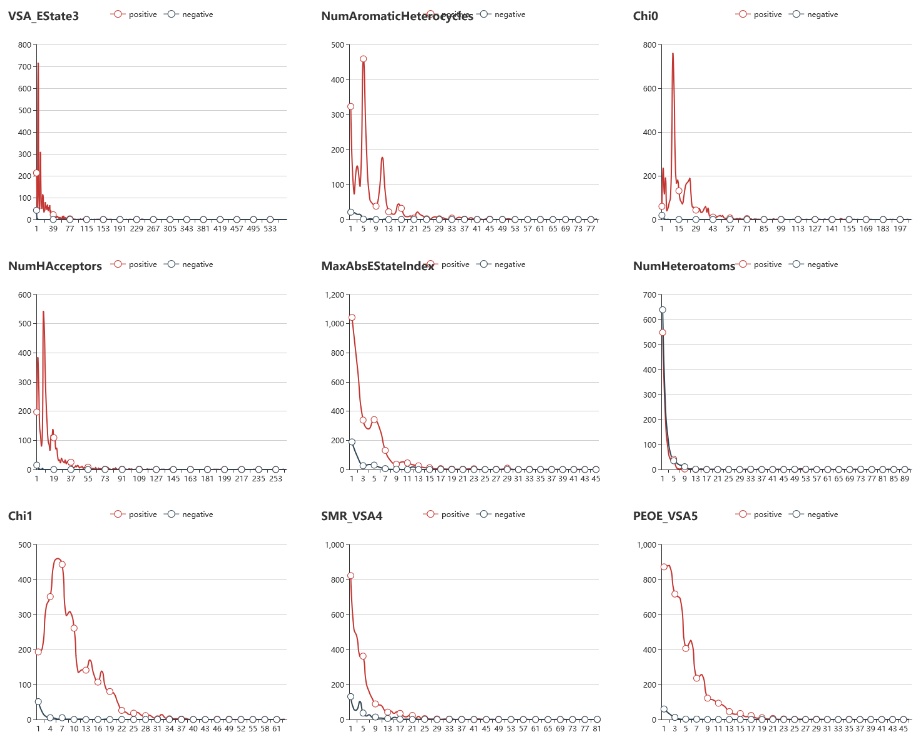
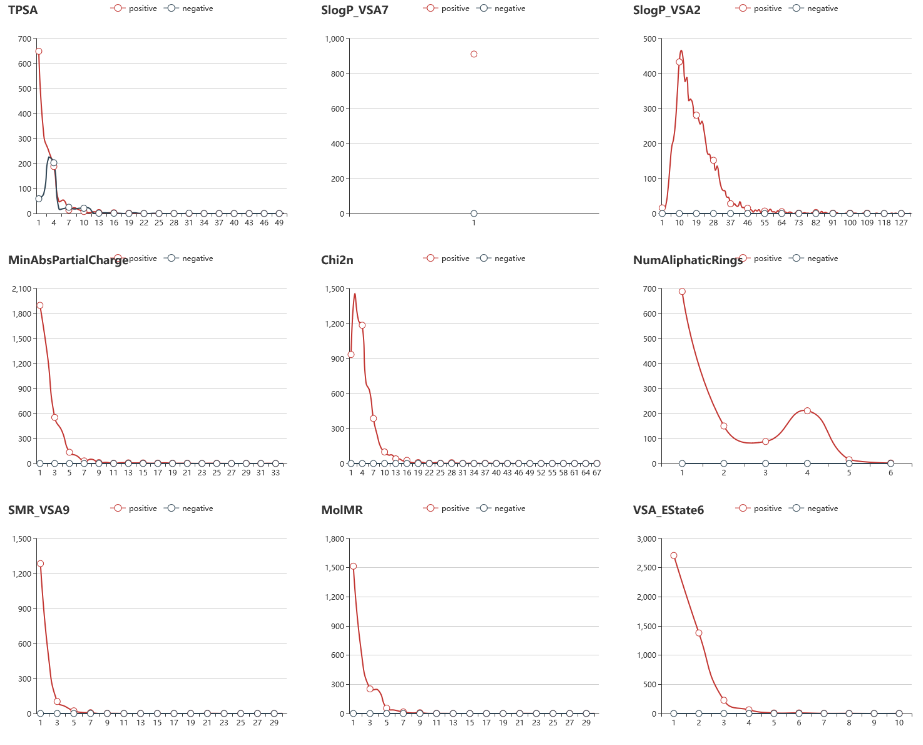
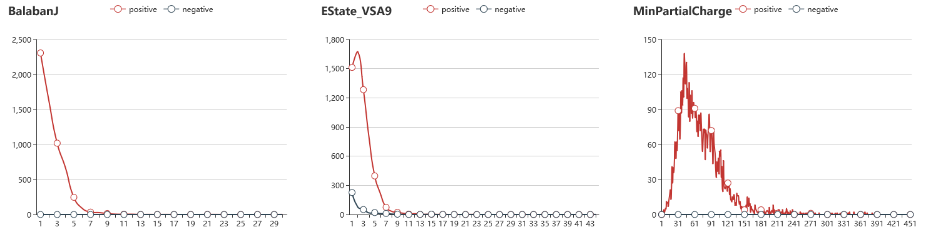
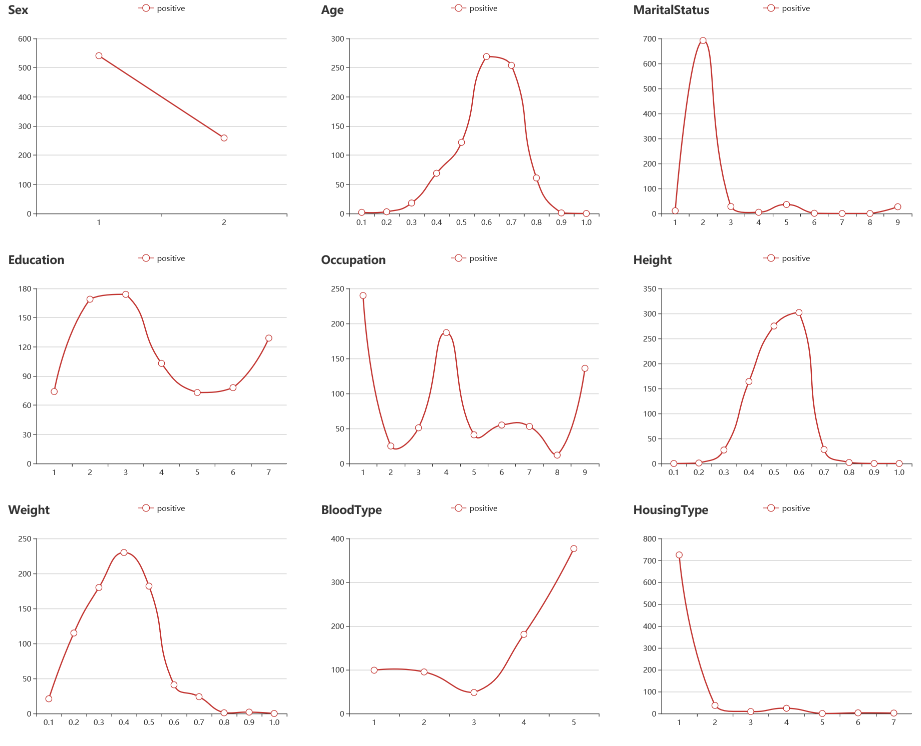
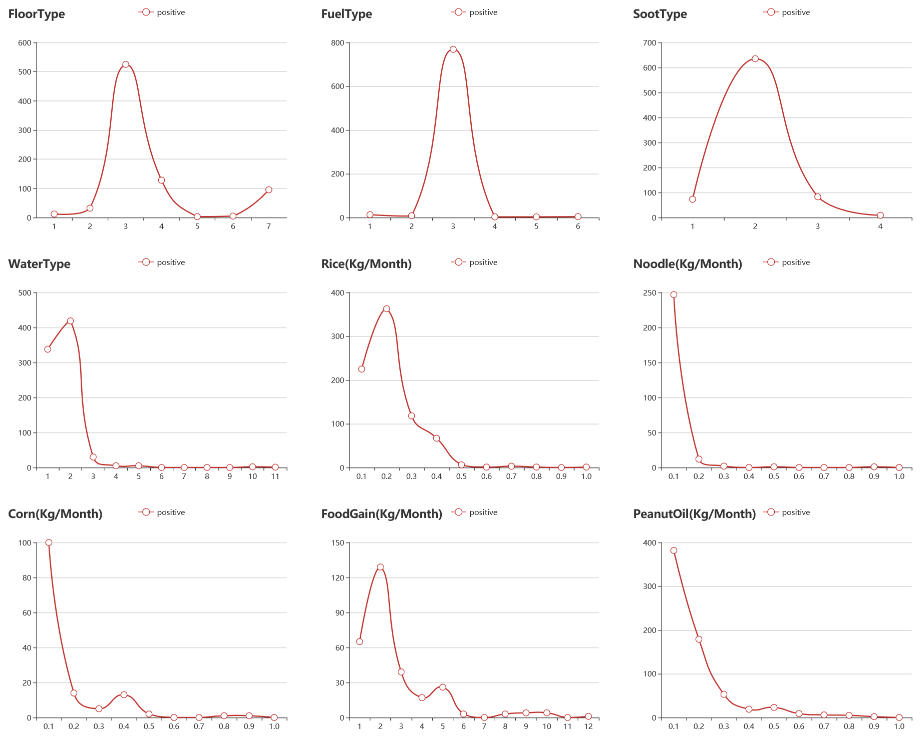
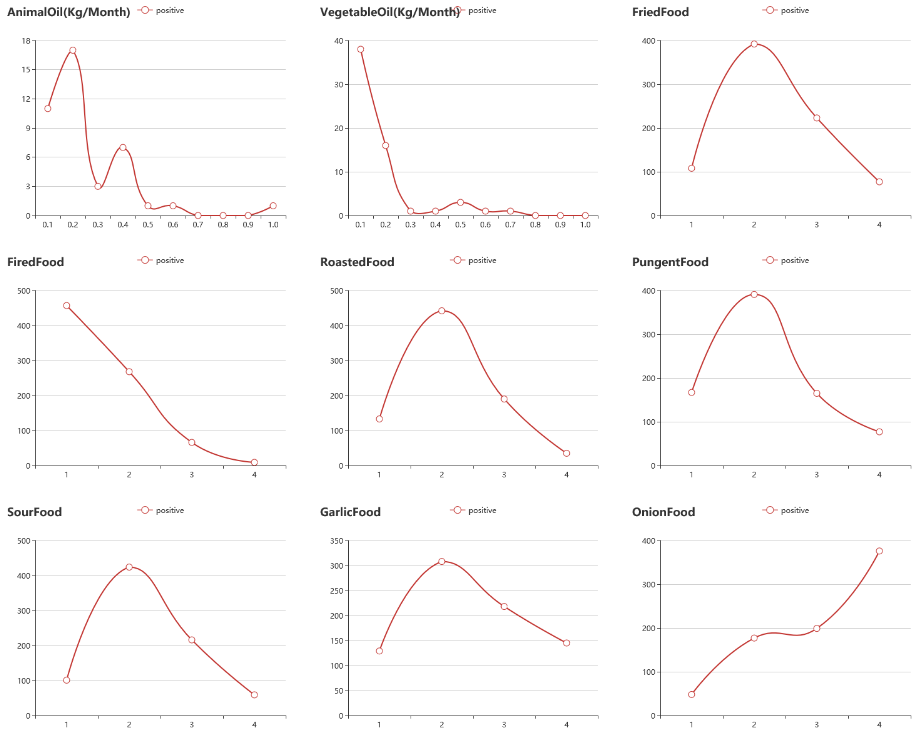
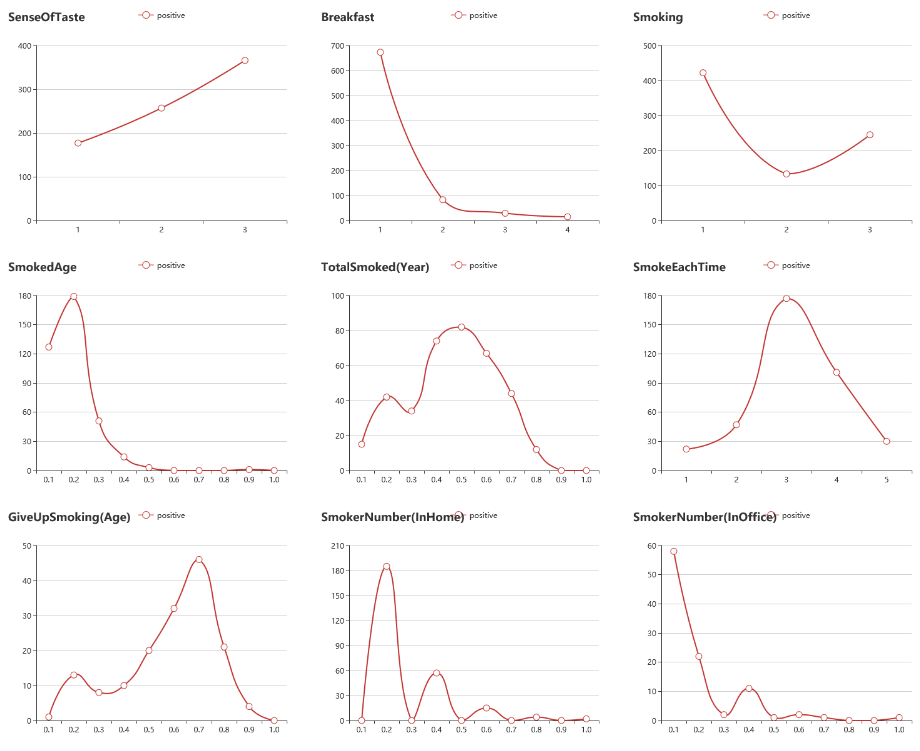
          

Figure 19:input chart test data 1: tox21-features.csv

### Test data 2: LungCancer\_cancerYesNo.csv



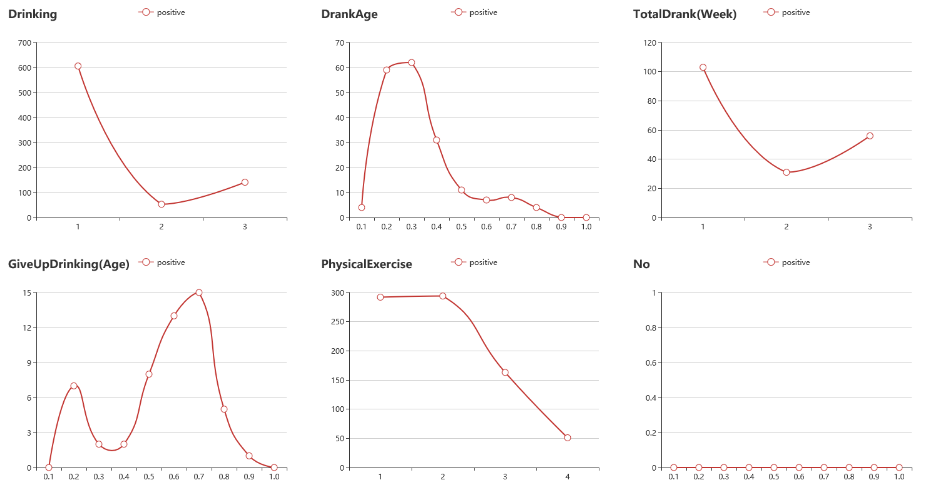


Figure 20: input chart test data 2: LungCancer\_cancerYesNo.csv

## Output data chart

The data obtained from the modeling group cannot be used directly, so I wrote a Java program to process the data and generated a usable data file.

Format-compliant data contains 70 nodes. I divided them into two parts and used the upper half and the whole data as two sets of data to test the output chart.

### Test data 1: The upper half of the data

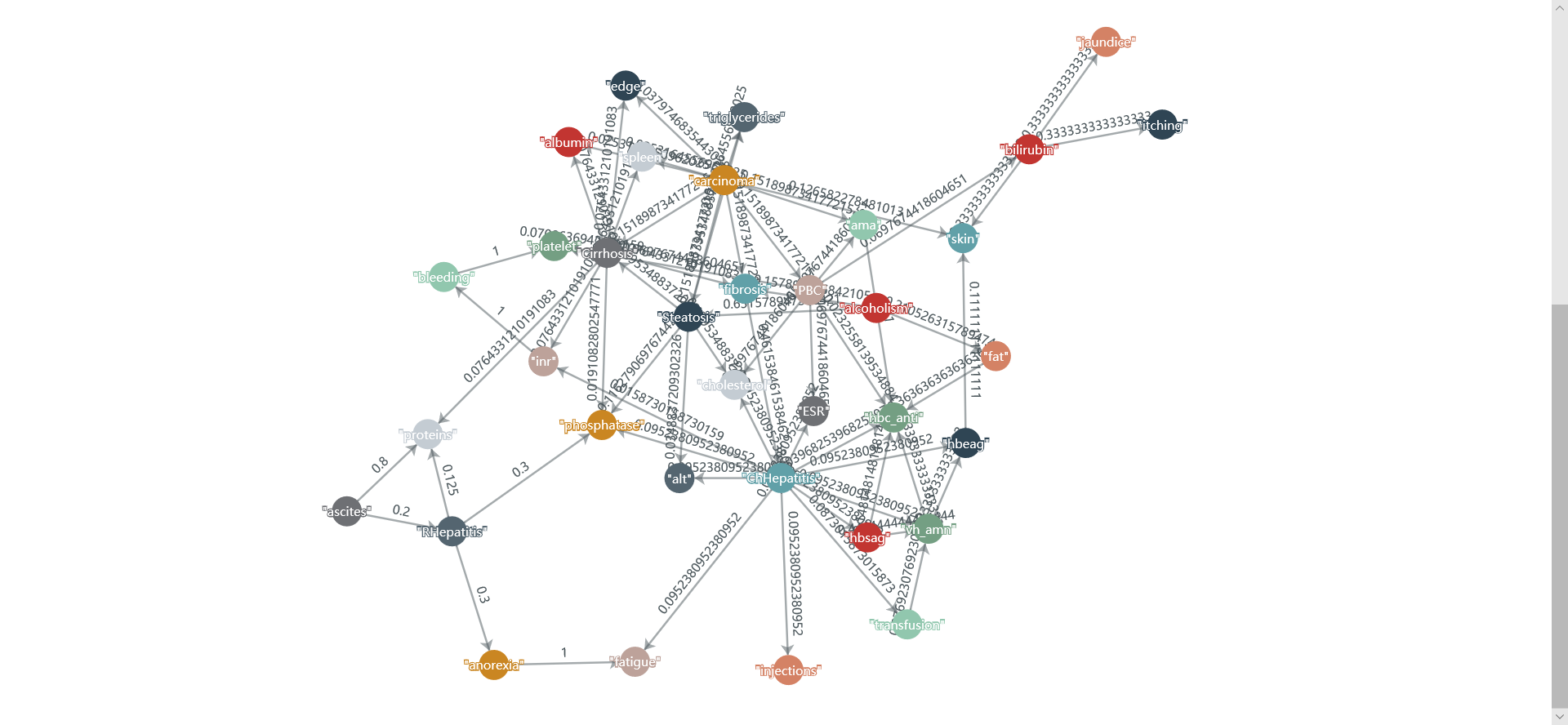


Figure 21:output chart test data 1: The upper half of the data



Figure 22: details of output chart test 1

### Test data 2: The whole data

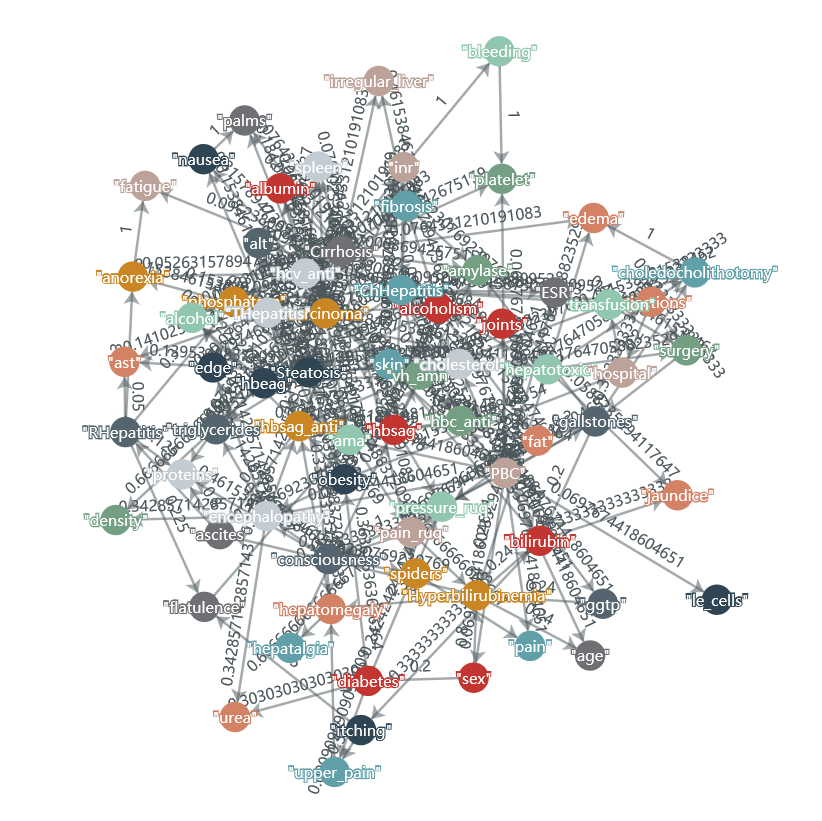


Figure 23: output chart test data 2: The whole data

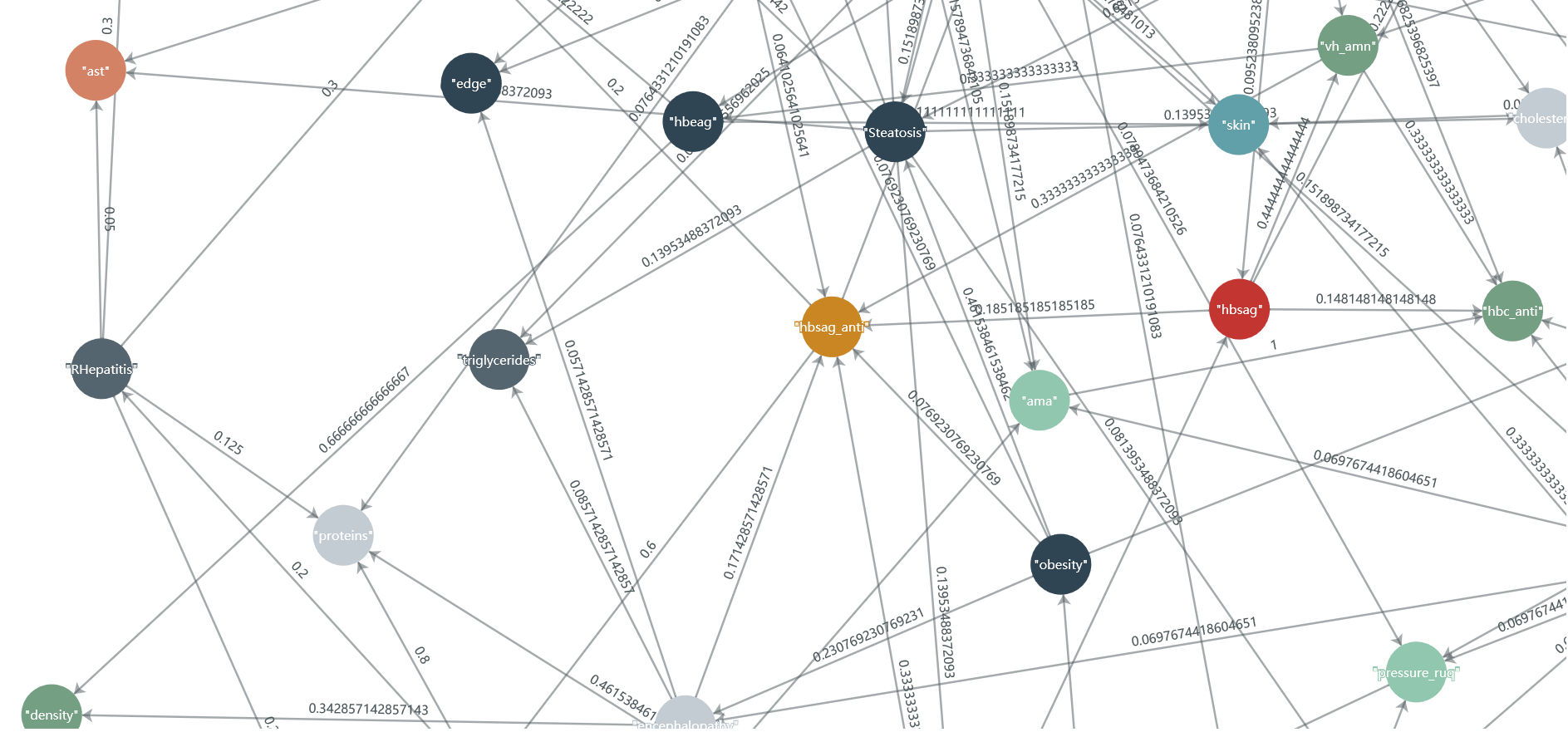


Figure 24:details of output chart test 2

# Ethics and Professional Conduct

To determine some ethical issues in the project.

## Data

The test data used in the project are all legally obtained on the Internet, without any disclosure of personal privacy, and are statistical data without specific personal information.

## Code

All the codes used are open source projects, and some teaching blogs were referenced during the writing process. Thanks to the author who wrote the teaching blog.

## Contribute

The ultimate goal of the project is to make it easy to use. This project will help doctors or other users to conduct research more conveniently and use the interactive interface to get the information they want.

## Precautions

When entering the file address, please enter it manually, do not copy the address to avoid errors.

# SUMMARY & CONCLUSION

From the charts above, we can easily find that some nodes have affected many other nodes, such as hbsag\_anti, THepatities, and Steatosis, all of which have at least 8 relationships. These relationships need additional attention. At the same time, although some nodes are only affected by a small number of other factors, the weight of individual factors is very high. For example, the effect of platelets on bleeding is 100%.

This project is an invaluable experience, and it gave me experience in JavaScript and data visualization. Although half of the time was isolated at home due to the epidemic, the project was still completed as expected.

Some problems were encountered in the project. For example, the file address copied from the file attribute in the windows system cannot be directly used in the java program, otherwise it will be unreadable. Some problems are relatively common, and solutions can be easily found in the forum, while others are even more rare, and you need to explore the solutions yourself.

In general, the project was very successful. I am very grateful to my supervisors and my teammates for their support, and to the schools and departments for providing opportunities.

REFERENCES

[1] Ilczuk, G., Wakulicz-Deja, A.: Attribute selection and rule generation techniques for medical diagnosis systems. RSFDGrC 2005, LNCS 3642 (2005) 352–361

[2] Ilczuk, G., Mlynarski, R., Wakulicz-Deja, A., Drzewiecka, A., Kargul, W.: Rough sets techniques for medical diagnosis systems. Computers in Cardiology 2005 32 (2005) 837–840

[3] Mlynarski, R., Ilczuk, G., Wakulicz-Deja, A., Kargul, W.: Automated decision support and guideline verification in clinical practice. Computers in Cardiology 2005 32 (2005) 375–378

[4]Zhongshenofang, Zhikehao, What are the benefits of big data visualization, [Online]. Available: https://qianjia.com/zhike/ZkhWriter.html?authorid=1006

[5] echarts official website https://echarts.apache.org/zh/index.html