IOP Conference Series: Earth and Environmental Science

PAPER • OPEN ACCESS

Data Visualization Model Methods and Techniques

To cite this article: Shengyuan Bai *et al* 2019 *IOP Conf. Ser.: Earth Environ. Sci.* **252** 052063

View the article online for updates and enhancements.

This content was downloaded from IP address 114.99.54.65 on 09/10/2019 at 21:46

Data Visualization Model Methods and Techniques

Shengyuan Bai[1], Xiangyi Zhou[2], You Lyu[3], Jiali Wang[4], Chengxiang Pan[1, *]

[1]Navigation College, Dalian Maritime University, Dailian, Liaoning 116026, China
[2]Information College, Beijing Forestry University, Beijing 100080, China
[3]Information College, Liaoning University, Shenyang, Liaoning, 110136, China
[4]Information College, Liaoning University, Shenyang, Liaoning, 110136, China

*Communication author: 13236840646@163.com

Abstract. In order to meet the requirements of high-dimensional data processing in the information field, this paper aims to explore methods and techniques for visualizing general data resource clustering data. Through the visual mapping of dimensionality reduction and high-dimensional data, a visual learning model for visual influencing factors is established. The visual system model approach was tested using the IRIS dataset from the University of California Irvine database (UCL) database. The results show that the model can effectively analyze the data set, visualize the characteristics of IRIS data in real time, achieve the expected results, and point the way for other data visualization models.

1. Introduction

From an information service perspective, data visualization is understood as a way to provide the essence of the interface, regardless of the internal structure of the data visualization. What is needed for a user or information provider is a simple, easy to understand, time-saving, and efficient way to present data content and understand the meaning of the data. Regardless of the data visualization technology tool, the ultimate goal should be to meet the needs of some users, but for data visualization technology tool providers, the ultimate goal is how to effectively provide data visualization services and develop appropriate visualization tools or platforms. [1]

Model is the basic method for human beings to understand and describe things through abstraction and filtering [2], which is convenient for humans to further explore the inherent laws and potential values of things.

The visual model is highly abstract, clear and regular, and powerful, so it is a powerful tool for data resource management and utilization. As an innovation in the development and utilization of data resources, data mining visualization methods and techniques as a composite concept stem from the combination of data mining technology and visualization methods. Data mining can help employees find information that is of interest to the data domain faster, or display some novel conclusions [3]. Data visualization technology can visualize and simplify complex data, making it easy for researchers to study the inherent laws of data.

Visualization techniques currently which based on the NumPy, Pandas, Matplotlib, and Seaborn data analysis libraries in Python are widely used and can be used for various dataset analysis. Moreover, the visualization technology has the advantages of easy implementation and good interactivity.

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

2. Data visualization features and classification

Faced with intricate high-dimensional data, people often complain that we don't understand the meaning of the data. Transforming complex data into acceptable forms by using scientific data visualization techniques. [4] Therefore, data visualization technology has the following characteristics:

(1) Interactivity: Easily manage data;

(2) Multidimensional: information that acquires multiple dimensions (attributes, features) of the data, but in the form of one dimension in front of the person;

(3) Visibility: The data is finally presented in the form of graphics, images, curves and animations, and its patterns and relationships can be visualized;

(4) Ease of use: Data can be quickly analyzed and mined.

With the rapid development of computer technology, only single-scale data can be visualized before, and large-scale and high-dimensional data can now be visualized and dynamically interacted with each other. High-dimensional data no longer makes people feel terrible and scared.

Data visualization is a scientific and technological study on the visual representation of data. Among them, the visual representation of such data is defined as a kind of information extracted in a certain summary form, including various attributes and variables of the corresponding information unit [].

It is a concept that is constantly evolving and its boundaries are constantly expanding. Primarily refers to technically advanced technical methods that allow visualization of data through representation, modeling, and display of stereo, surface, attributes, and animation using graphics, image processing, computer vision, and user interfaces. Explanation. Compared to special technical methods such as stereo modeling, data visualization covers a much wider range of technical methods.

(i) Data collection

Data collection (sometimes abbreviated as DAQ or DAS), also known as "data acquisition" or "data collection," refers to the process of sampling the real world to produce data that can be processed by a computer. Typically, the data acquisition process includes the steps of acquiring and processing the signals and waveforms in order to obtain the desired information. Among the components of the data acquisition system are sensors for converting measurement parameters into electrical signals, which are acquired by the data acquisition hardware.

(ii) Data analysis

Data analysis refers to the process of detailed research and summary of data in order to extract useful information and form conclusions. Data analysis is closely related to data mining, but data mining tends to focus on larger data sets, less on reasoning, and often uses data that was originally collected for a different purpose. In the field of statistics, some people divide data analysis into descriptive statistical analysis, exploratory data analysis, and confirmatory data analysis; among them, exploratory data analysis focuses on discovering new features in the data, while confirmatory data analysis Focus on the verification or falsification of existing hypotheses.

(iii) Types of data analysis include:

1) Exploratory data analysis: refers to a method for analyzing data in order to form a test that is worthy of hypothesis, and is a supplement to the traditional statistical hypothesis testing method. The method is named by the famous American statistician John Tukey.

2) Qualitative data analysis: also known as "qualitative data analysis", "qualitative research" or "quality research data analysis", refers to non-numeric data (or data) such as words, photos, observations, etc. Analysis.

After 2010, the data visualization tools are mainly based on visual elements such as tables, graphs, maps, etc. The data can be filtered, drilled, data linked, jumped, highlighted and other analysis methods for dynamic analysis.

Visualization tools can provide a variety of data presentation forms, diverse graphics rendering forms, rich human-computer interaction methods, dynamic scripting engines that support business logic, and more.

Different from the general Dashboard or Reporting products, Yonghong's BI front-end is discovery-type: rich in interactive means and powerful in analysis. Users can further interact with data, filter, drill, brush, associate, transform, etc., allowing users to: grasp information, find problems, Find the answer and take action.

(iiii) Data Mining

Data mining is the process of sorting and sorting out large amounts of data. Data mining is often used by business intelligence organizations and financial analysts; however, in the scientific arena, [10] data mining is increasingly being used to extract information from the vast data sets generated by modern experimental and observational methods.

Data mining is described as "the extraordinary process of extracting implicit, previously unknown, potentially useful information from the data" and "the science of extracting useful information from large data sets or databases." Data mining related to enterprise resource planning refers to the process of statistical analysis and logical analysis of large transaction data sets, looking for a process that may contribute to the decision-making work [11].

3. Visual analysis of influencing factors

1. Data space: is a multi-dimensional information space composed of a data set consisting of n-dimensional attributes and m elements;

2. Data development: refers to the derivation and calculation of quantitative data using certain algorithms and tools;

3. Data analysis: refers to the analysis of data such as slicing, block, and rotation of multi-dimensional data, so that data can be observed from multiple angles and multiple sides;

4. Data visualization: refers to the process of representing data in a large data set as a graphical image and using data analysis and development tools to discover unknown information.

Data visualization has proposed a number of methods that can be divided into geometry-based techniques, pixel-oriented technologies, icon-based technologies, layer-based techniques, image-based technologies, and distributed technologies, depending on the principles of their visualization [12].

3.1. Data set description and make analysis program

This experiment selected the IRIS data set as a test data set. The selection of data sets is based on two aspects: on the one hand, the IRIS data set is a standard test data set, with rich research results, and has the highest identity and recognized best data field visual test set can reduce complexity results evaluation and Strengthen persuasiveness. On the other hand, simple categorical data can better visualize the technology, and the IRIS data set fully meets this requirement.

The IRIS dataset, also known as the iris flower feature dataset, is available from UCI (University of California, Irvine) machine learning database [5]. The IRIS data set contains a total of 150 data points obtained by collecting 50 data points of Setosa, Versicolor and Virginca from a given attribute. The data set can be described by the five attributes listed in Table 1.

Table 1. Iris flower data set attribute description

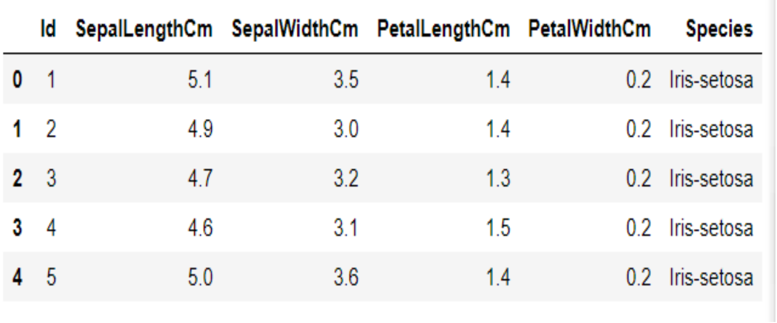
Serial number	Attribute name	Property Description
1	sepal_length	Sepals length, cm
2	petal_length	Petal length, cm
3	sepal_width	Sepal width， cm
4	petal_width	Petal width， cm
5	Class	Iris species, including setosa, versicolor, virginca three

Obviously, the table consists of four attributes of iris SepalLengthCm, SepalWidthCm, PetalLengthCm, PetalWidthCm and its iris species Species.

Specific analysis procedures: Among the multiple influencing factors, only one of them is analyzed at a time, and other factors remain unchanged. The influencing factors are analyzed by comparing the visualized iris flower data map, the text form and the trend graph of the convergence criteria.

3.2. Data set preprocessing visualization

First of all, according to the characteristics of the data set itself necessary pretreatment, and then start the feature analysis. The purpose of data preprocessing is usually denoising, eliminating redundancy, cleaning data, converting data formats, and so on. Figure 1 has been csv data format into a data structure for the array DataFrame form.



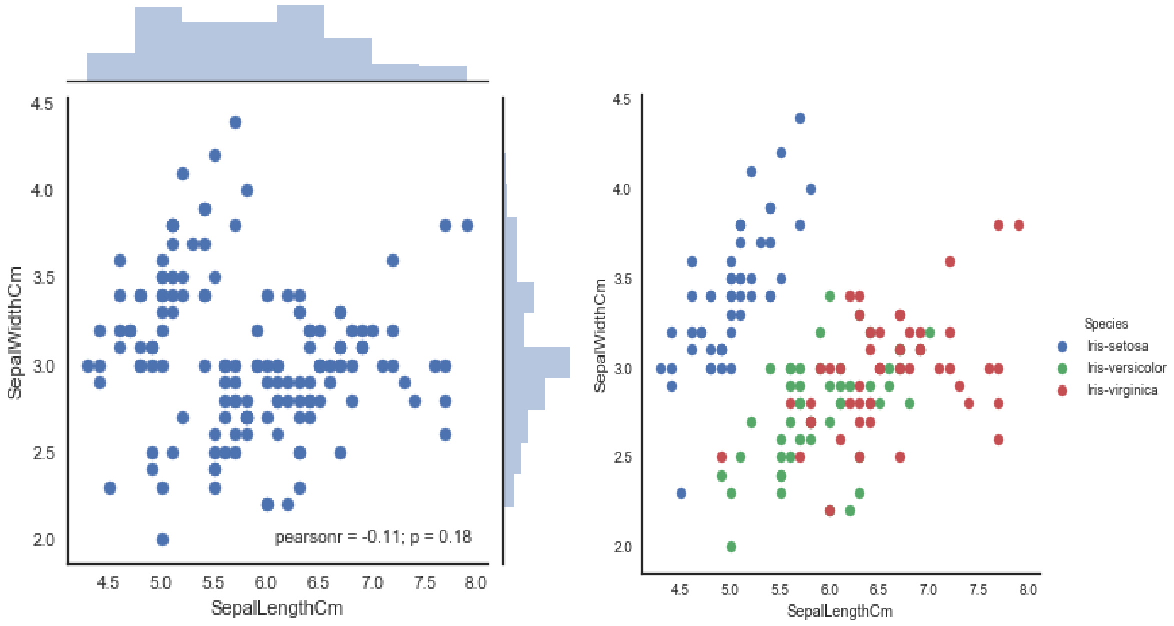
	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa

Fig.1. Iris flower dataset

3.3. Quick scatter diagram visualization method

Scatter plot visualization is the display of multidimensional data in a two-dimensional plane that can more clearly identify hidden information in the data. Scatter plot visualization is an effective way to visualize dimensionality reduction data. It expresses data sets in a concise and intuitive graphical form and is an effective way to visualize multidimensional data due to its reduced dimensionality [9].

Taking the length and width of the cymbal as the research object, the paper introduces the single influencing factor as the data visualization process of the research object.



(a) (b)

Fig 2. Scatter plot

Figure 2 (a) shows the data in the form of a scatter-point histogram, which not only shows the distribution of data but also the concentration of data. Figure 2 (b) is marked iris species, so that the type of sepals length and width distribution.

3.4. Visualization based on dimensionality reduction

The main idea of 2D dimension data visualization is to find a mapping from high-dimensional space to low-dimensional space, and to maintain some distribution and structural features of high-dimensional data space through the seed mapping technology. [6]

Since the petals are organoleptic in three dimensions, the visualization of the three-dimensional data is poorly visualized, so the data is dimensionally reduced. In this paper, two-dimensional view visualization method to describe the length and width of petals.

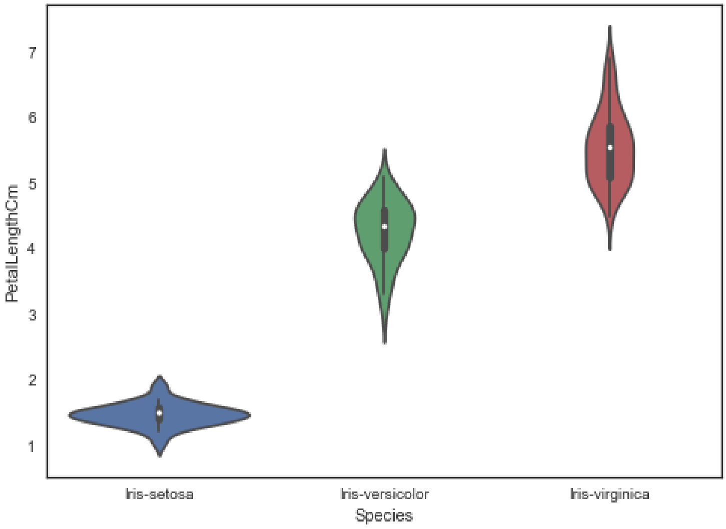


Fig 3. Petal two-dimensional representation

3.5. Visualization based on statistical distribution

Many times we want to know the statistical characteristics of the data. We can see from the central limit theorem that many data in the life will finally be normally distributed. How to visualize it? The skeplot function in the Seaborn package in Python works.

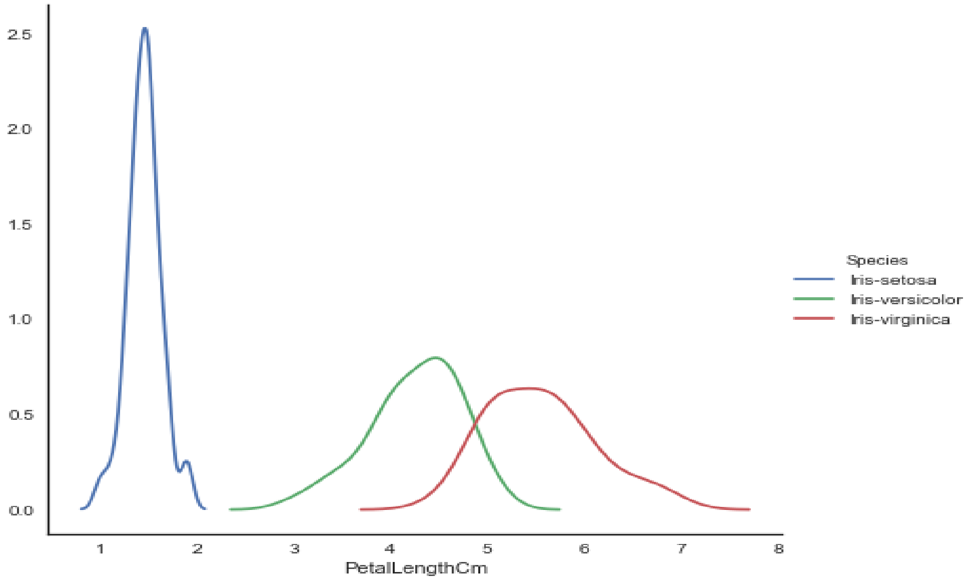
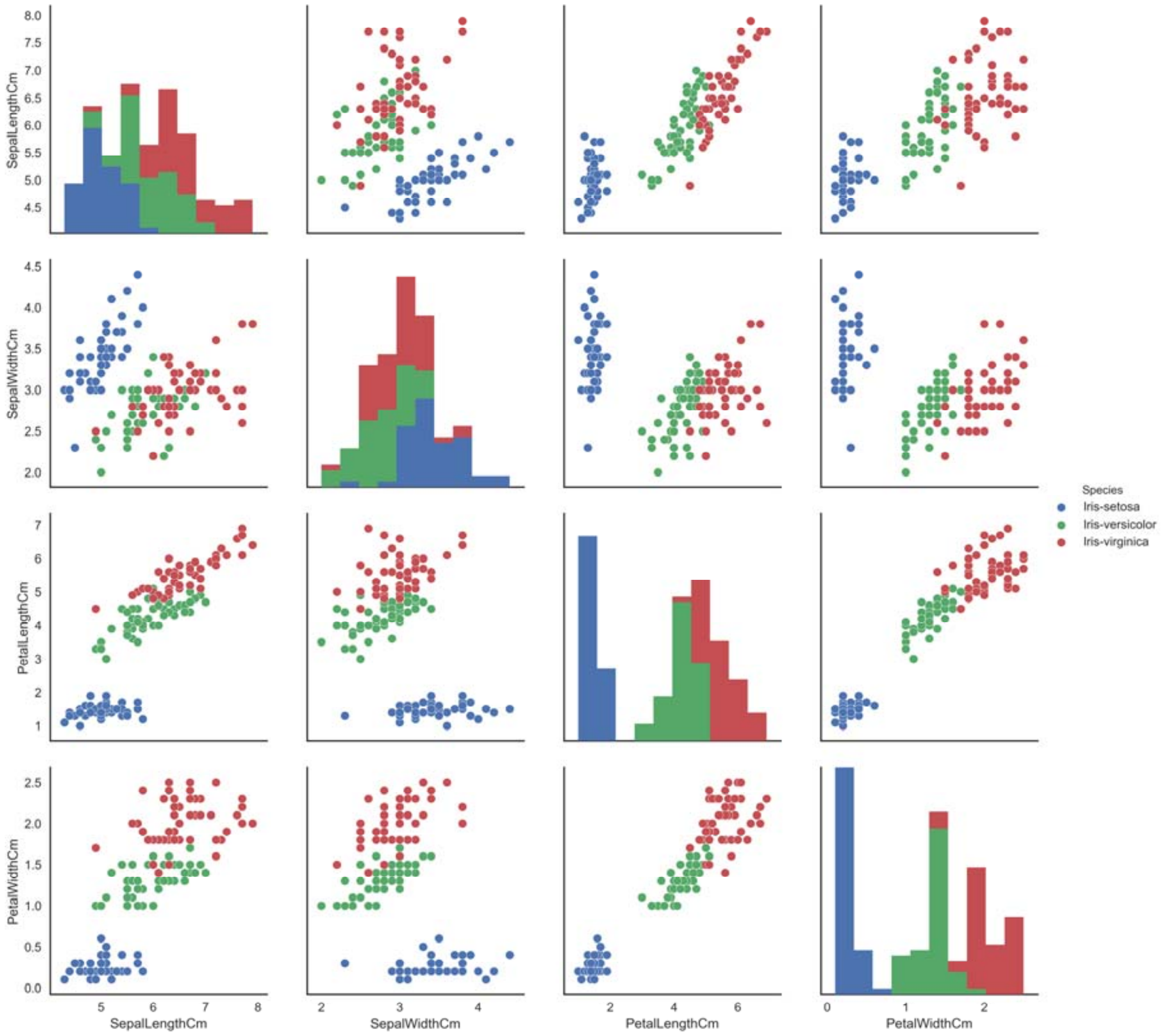


Fig 4. Data distribution chart

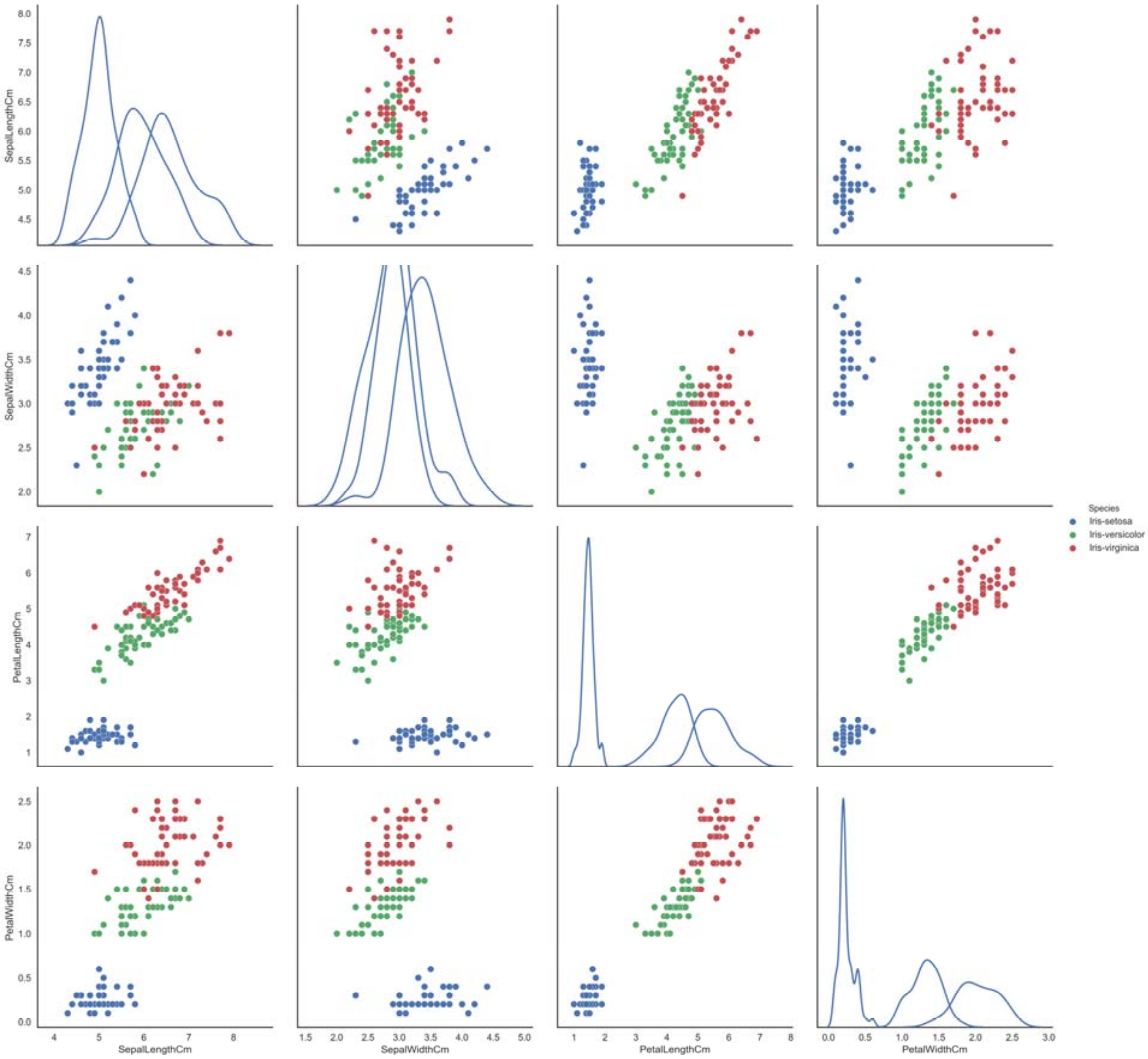
3.6. Visual comparison based on graphics

This technique is different from the icon-based visualization technique. The technique of graph-based visualization uses the graph to represent the dimension information of multi-dimensional data.

From the data set, Iris has two major characteristics. In the above, a simple description has been made on the visual analysis of a single factor. Next, we will consider the multi-factor and multivariate comparison. The Pairplot function in Seaborn takes multiple factors into account and performs a horizontal comparison at the same time.


SepalLengthCm
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
SepalWidthCm
4.5
4.0
3.5
3.0
2.5
2.0
PetalLengthCm
7
6
5
4
3
2
1
PetalWidthCm
2.5
2.0
1.5
1.0
0.5
0.0
5
6
7
8
SepalLengthCm
2
3
4
SepalWidthCm
2
4
6
PetalLengthCm
0
1
2
PetalWidthCm
Species
Iris-setosa
Iris-versicolor
Iris-virginica

(a)

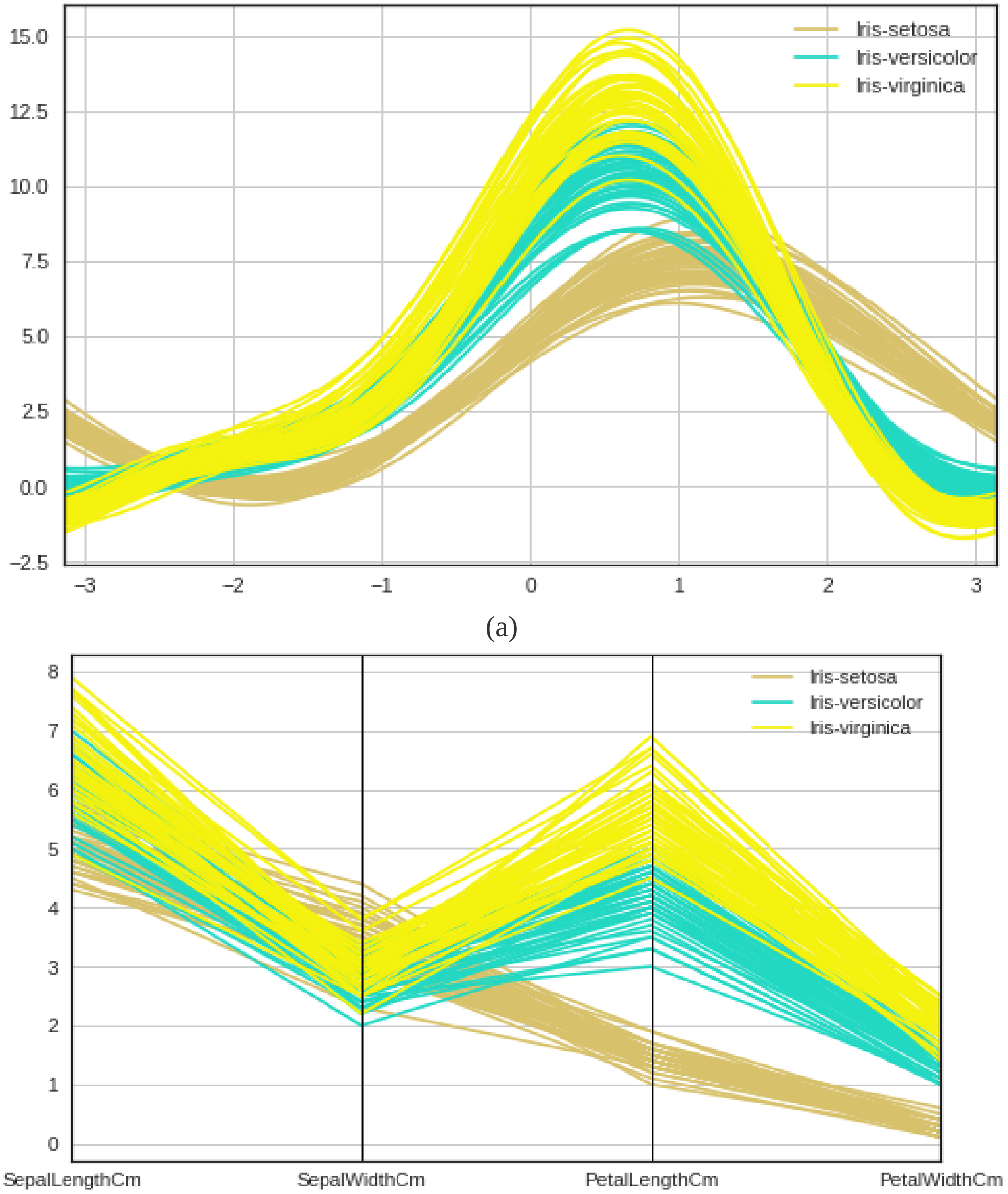


(b)

Fig 5.eature comparison chart

4. Parallel coordinate visualization method

Parallel coordinate visualization is a typical multidimensional visualization technique based on geometry. It is a visualization technique proposed by Inselberg to represent n-dimensional data in two dimensions. [7] It is widely used in the field of data mining and visualization. Its basic idea is to map n properties of n-dimensional data into two-dimensional space by n equidistant parallel axes, where each parallel axis represents the dimension of high-dimensional data and the values of these axes That is, the size of the property value, the various dimensions are connected by means of the polyline. Through the visualization method, users can directly see the distribution of high-dimensional data in two-dimensional space. [8]



(a)

Fig 6. Two-dimensional mapping chart

5. Conclusion

In this article, two factors that influence the classification of iris, sepal, petal, length, and width are visualized by using the Data Science Analysis Package and the IRIS dataset in Python. Intuitive, concise visual analysis solutions explain obscure theories in new ways, enabling researchers in new fields to understand the knowledge domain better and faster. The visual analysis program enriches the expression of scientific theory. At the same time, the success of the experiment has guiding significance for the verification and interpretation of other mature theories.

References

[1] Tan Guilong, Chen Yi. Application Research of information visualization method based on parallel coordinates [J]. Journal of Beijing Technology and Business University: Natural Science Edition, 2008, 26 (2): 7579. Roy. Soc. London, vol. A247, pp. 529-551, April 1955.

[2] Luo Jian. Research and implementation of visual data mining method [D]. Xi'an: University of

Electronic Science and technology of China, 2009.

[3] Vesanto J. SOM-based data visualization methods [M]// SOM-Based Data Visualization Methods. 1999.

[4] Sarkar D. Multivariate Data Visualization with R [J]. Springer, 2008, 25 (b02): 275-276.

[5] Van Wijk J J. Spot noise texture synthesis for data visualization [J]. Acm Siggraph Computer Graphics, 1991, 25 (4): 309-318.

[6] Bishop C M, Tipping M E. A hierarchical latent variable model for data visualization [J]. Pattern Analysis & Machine Intelligence IEEE Transactions on, 1998, 20 (3): 281-293.

[7] Xiong Zhongyang, Chen Ruotian, Zhang Yufang. An effective K-means clustering Heart initialization method [J]. Computer Application Research, 2011, 28 (11): 4188-4190.1963, pp. 271-350.

[8] Fiorini P, Inselberg. A Configuration Space Representation in Parallel Coordinates [C]. International Conference on-Robotics and Automation, CA, USA: Jet Propulsion Lab, 1989.

[9] Guan Wang. Visual Data Mining Technology Based on Parallel Coordinate Method [D].Beijing: Peking University, 2008.

[10] BOUGHRIRA A, FAY D, KHADIR M T. Kohonen map combined to the k - means algorithm for the identi -fication of day types of algerian electricityload [C].Andrienko G. Computer Information Systems and In -dustrial Management Applications. Alger: Computer Information Systems and Industrial Management Ap -plications, 2008: 78 - 83.

[11] Zhai Xu Jun. Research on Visual Data Mining Technology Based on Parallel Coordinate Method [D]. Beijing: Tsinghua University, 2004.

[12] Lei Junhu, Yang Jiahong, Zhong Jiancheng, and so on. High dimensional data visualization based on PCA and parallel coordinates [J]. computer engineering, 2011, 37 (1): 48 - 50.