Study on vessels Type Based on PCA

Tao Wang [a], Keping Guan [b]

School of Shanghai Maritime University, Shanghai 200000, China

[a]568416214@qq.com, [b]kpguan@shmtu.edu.cn

Abstract

At present, the types of ships with many vessels are not clear, including many vessels information missing from CSN. One of the missing items is the ship type, which has caused obstacles to the collection of ship information and statistical analysis, etc. , This article has made the following work to the ship type question: First, gathers the AIS data information of a certain section of the vessels; secondly, analyzes the data, extracts the sample; Third, the data cleansing, finishing and so on pretreatment work; , Extract the main components, dimension reduction, clustering. Finally, when the number of clusters is 2, the classification result is the best.

Keywords

Vessels type, PCA, dimension reduction, clustering.

1. Introduction

At present, there are many types of vessels, but many types of vessels are not clear, including many ship information missing on CSN. One of the missing ones is the type of vessel. The vessel type is not clear and the vessels information is not complete, which obstructs the statistical analysis of the flow of a certain segment.

2. Organization of the Text

2.1 Data analysis

Each vessel collects six major characteristics, namely, MMSI, Speed, Length, Width, draft, Name. The six features are first described statistically, the correlation of each feature is considered, and then the data set Several sample data points, tracking studies. Because the MMSI code and the name of the vessel are not numerical data, there is no point in making a statistical description, so delete these two columns for statistical description. Get as Figure 2-1.

	Speed	Length	Width	Draft
count	886.000000	886.000000	886.000000	886.000000
mean	14.495711	58.135440	10.446953	3.401727
std	19.563420	39.272419	6.732752	2.418611
min	2.300000	8.000000	3.000000	0.460000
25%	6.200000	52.000000	9.000000	3.020000
50%	7.100000	53.000000	9.000000	3.130000
75%	8.575000	59.000000	11.000000	3.500000
max	85.400000	1022.000000	110.000000	59.280000

Figure 2-1 Ship characterization

2.1.1 Sample selection

In order to have a better understanding of the type of vessel and to understand how the data representing the vessel will change during this analysis. It is best to choose a few sample data points and analyze them in more detail. Three data points with obvious difference were chosen as samples, and the three ships with index numbers of 11, 13 and 14 were selected by comparison. The sample information is shown in Figure 2-2.

	Speed	Length	Width	Draft
0	8.1	8	3	0.46
1	10.9	400	25	8.50
2	4.9	122	15	7.80

Figure 2-2 Ship sample information

2.1.2 Features Relevance

Consider whether one (or more) of these six features has a realistic correlation to the ship classification. That is, when a certain characteristic of the ship is established, can we determine the quantity of another characteristic. There is a simple way to detect correlations: we build a supervised learning (regression) model with a dataset with a certain feature removed and then use that model to predict which feature is removed. Score and see how the forecast is. As shown in Figure 2-3.

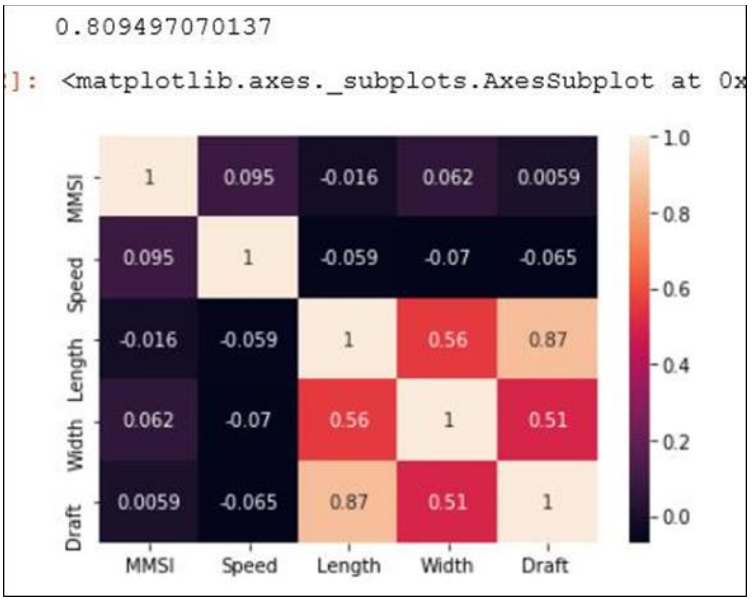

Figure 2-3 Characteristic correlation matrix

The results show that the correlation between the length of the removed feature and other features is very high and the coefficient of determination is R ^ 2 = 0.809. The correlation matrix shows that the correlation between feature Length and feature Draft is the highest Reaches 0.87. Next, the correlation between feature Length and feature Width reaches 0.56, and finally the correlation between feature Width and feature Draft reaches 0.51.

2.1.3 Visual Features Distribution

To get a better understanding of this data set, we can construct a scatter matrix for each ship feature in the data set [1]. If the predicted features are necessary to distinguish a particular ship, then this and other features may not show any relationship in the scattering matrix below. Conversely, if this feature

is not useful for identifying a particular ship, then it is clear from the scatter matrix that there is a correlation between this data feature and other features. As shown in Figure 2-4.

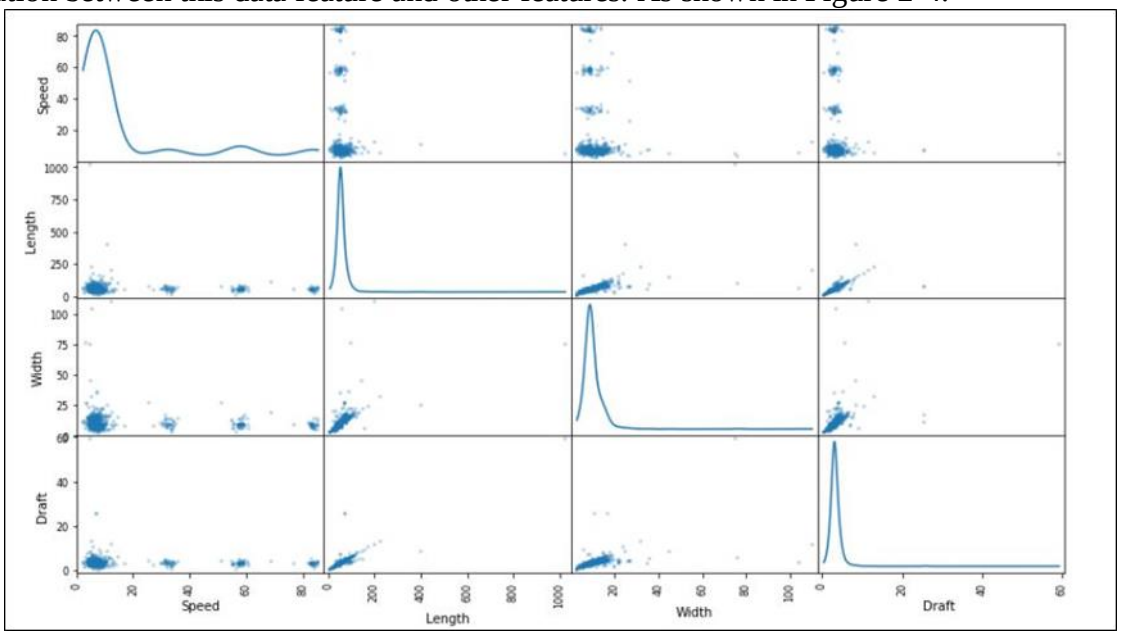

Figure 2-4 scattering matrix

2.2 Data preprocessing

Make a suitable scaling of the data and detect outliers to pre-process the data into a better form. Preprocessing data is an important part of ensuring that significant and meaningful results can be obtained from the analysis.

2.2.1 Feature zoom

If the data is not normally distributed, especially when there is a large discrepancy between the mean and the median of the data (indicating that the data is very skewed). This is usually a non-linear scaling is very suitable, this method can calculate the best to reduce the data tilt index transformation method. A simpler and most applicable method is to use natural logarithms.

After transformation as shown in Figure 2-5.

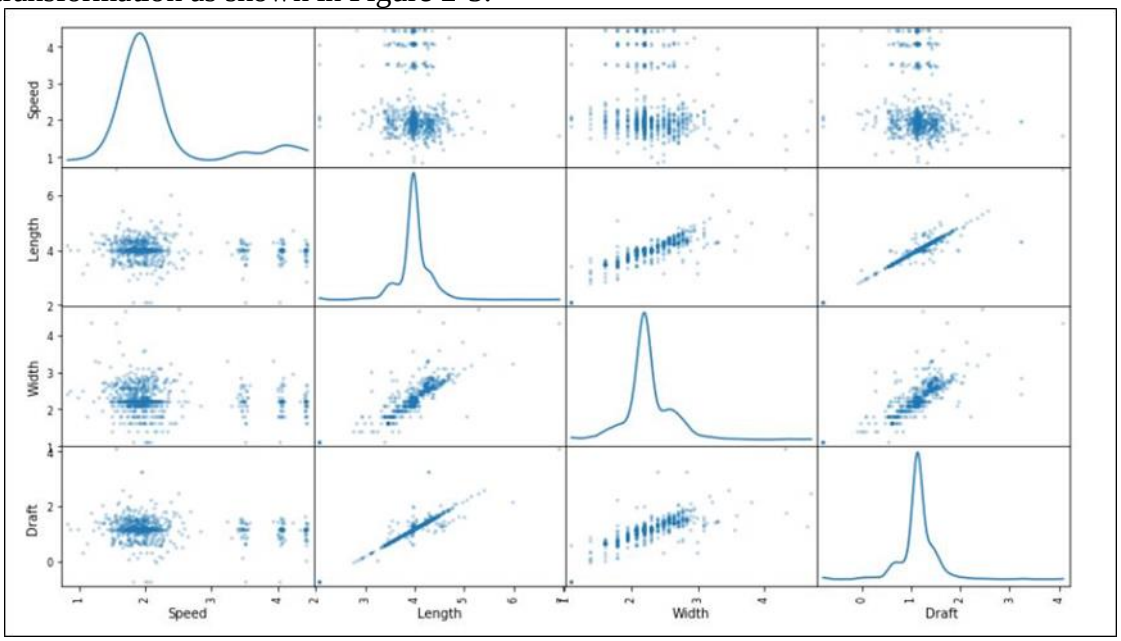

Figure 2-5 take logarithm of the scattering matrix

After using a natural logarithmic scaling, the various characteristics of the data will appear more normal distribution. For any related pair of features, their correlation still exists.

2.2.2 sample data feature scaling

After taking the logarithm of the sample data features were compressed, paving the way for the next detection of outliers. Figure 2-6 shows the logarithm of sample data.

	Speed	Length	Width	Draft
0	2.091864	2.079442	1.098612	-0.776529
1	2.388763	5.991465	3.218876	2.140066
2	1.589235	4.804021	2.708050	2.054124

Figure 2-6 Sample Data Feature Zoom

2.2.3 outlier detection

For any analysis, detecting outliers in the data during data preprocessing is a very important step. The appearance of outliers makes the results appear skewed when these values are taken into account. This article uses Tukey's method of defining outliers, ie an outlier step is defined as 1.5 times the interquartile range (IQR). A data point If a feature contains features outside the IQR of the feature, the data point is considered as an outlier. Figure 2-7 shows some of the more than one feature is detected under abnormal values of the data.

```
Counter({144: 4, 189: 4, 215: 4, 223: 4, 225: 4, 23
4: 4, 796: 4, 801: 4, 815: 4, 829: 4, 830: 4, 831:
4, 0: 3, 5: 3, 6: 3, 7: 3, 9: 3, 10: 3, 11: 3, 12:
3, 35: 3, 36: 3, 84: 3, 85: 3, 86: 3, 91: 3, 99: 3,
117: 3, 120: 3, 123: 3, 125: 3, 126: 3, 127: 3, 145
3, 206: 3, 207: 3, 210: 3, 214: 3, 219: 3, 221: 3,
38: 3, 268: 3, 275: 3, 276: 3, 319: 3, 343: 3, 352:
3, 366: 3, 368: 3, 369: 3, 397: 3, 421: 3, 453: 3,
96: 3, 602: 3, 607: 3, 609: 3, 611: 3, 627: 3, 638:
3, 681: 3, 686: 3, 695: 3, 697: 3, 701: 3, 703: 3,
44: 3, 754: 3, 758: 3, 765: 3, 768: 3, 769: 3, 773:
3, 791: 3, 793: 3, 803: 3, 804: 3, 811: 3, 819: 3,
43: 3, 849: 3, 862: 3, 865: 3, 866: 3, 868: 3, 872:
3, 884: 3, 885: 3, 1: 2, 8: 2, 16: 2, 21: 2, 83: 2,
```

Figure 2-7 abnormal data

As shown in Figure 2-7, the colon precedes the index of the outlier and the colon follows the number of detected features. For example, [144: 4] indicates that the ship data with index 144 is detected under the four characteristics Four anomalies.

2.3 Feature conversion

Use principal component analysis (PCA) to analyze the internal structure of ship data. Due to the use of PCA to calculate the dimension of maximum variance on a data set, this section will find a set of features that best describes the ship.

2.3.1 Principal Component Analysis (PCA)

Now that the data is scaled to a more normal distribution and we also remove the outlier that needs to be removed, we can now use the PCA algorithm on the preprocessed dataset to find that one of the dimensions of the data can be maximized Variance of the features. In addition to finding these dimensions, the PCA will also report on the explained variance ratio for each dimension how many variances in this data can be accounted for in this single dimension. A component (dimension) of a PCA can be seen as a new "feature" in this space, but it is made up of features in the original data. The result is shown in Figure 2-8.

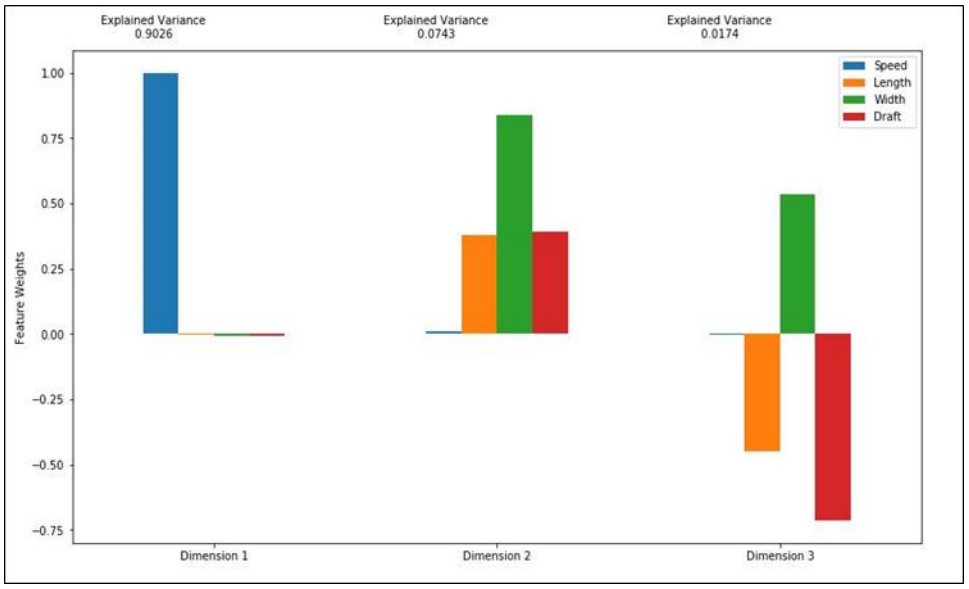

Figure 2-8 PCA analysis results

As shown in Figure 3-1, 'Dimension 1' explains the variance of 0.9026, 'Dimension 2' explains the variance of 0.0743 and 'Dimension 3' explains the variance of 0.0174.

2.3.2 Dimension reduction

When using principal component analysis, one of the main goals is to reduce the data dimensions, which actually reduces the complexity of the problem. Of course, dimensionality reduction is also costly: Fewer dimensions can represent less total variance in the data. Because the cumulative explained variance ratio is very important for how many dimensions we need to determine this problem. Figure 3-2 shows the result of dimensionality reduction of the sample data.

	Dimension 1	Dimension 2	Dimension 3
0	-0.0894	-2.4524	1.6314
1	0.1535	1.9501	-1.0736
2	-0.6362	1.0309	-0.7505

Figure 2-9 Sample dimension reduction

聚类数量	轮廓系数
2	0.816587446082
3	0.379032478292
4	0.251206194442
5	0.25396362476
6	0.232317947686
7	0.169979002648
8	0.242802217793
9	0.237355079918
10	0.188149417505

Figure 2-10 Number of Clusters - Contour coefficient

2.3.3 Clustering

Use K-Means clustering algorithms or Gaussian mixture model clustering algorithms to discover hidden customer categories in your data. We will then restore some of the key data points from the cluster and understand their meaning by converting them back to their original dimensions and scales.

For different situations, the number of clusters required for some problems may be known. However, given that the number of clusters is not a priori knowledge, we can not guarantee that the number of clusters will be optimal for this data because we do not know the structure (if any) of the data.

However, we can measure the quality of clustering by calculating the contour coefficients of points in each cluster. The contour coefficient of a data point measures its similarity to the cluster assigned to it, ranging from -1 (not similar) to 1 (similar). The average contour coefficient gives us a simple way to measure the quality of the cluster. Figure 2-10 shows the contour coefficients at different cluster numbers [2].

It can be seen from the figure that when the number of clusters is 2, the best contour coefficient is 0.8166.

2.3.4 Clustering visualization

Once you have chosen the optimal number of clusters for the algorithm obtained from the above evaluation function, you can visualize the result using code block visualization. Adjust the number of clustering clustering algorithm to compare different visualization results. The results are shown in Figure 2-11 ("X" for sample points).

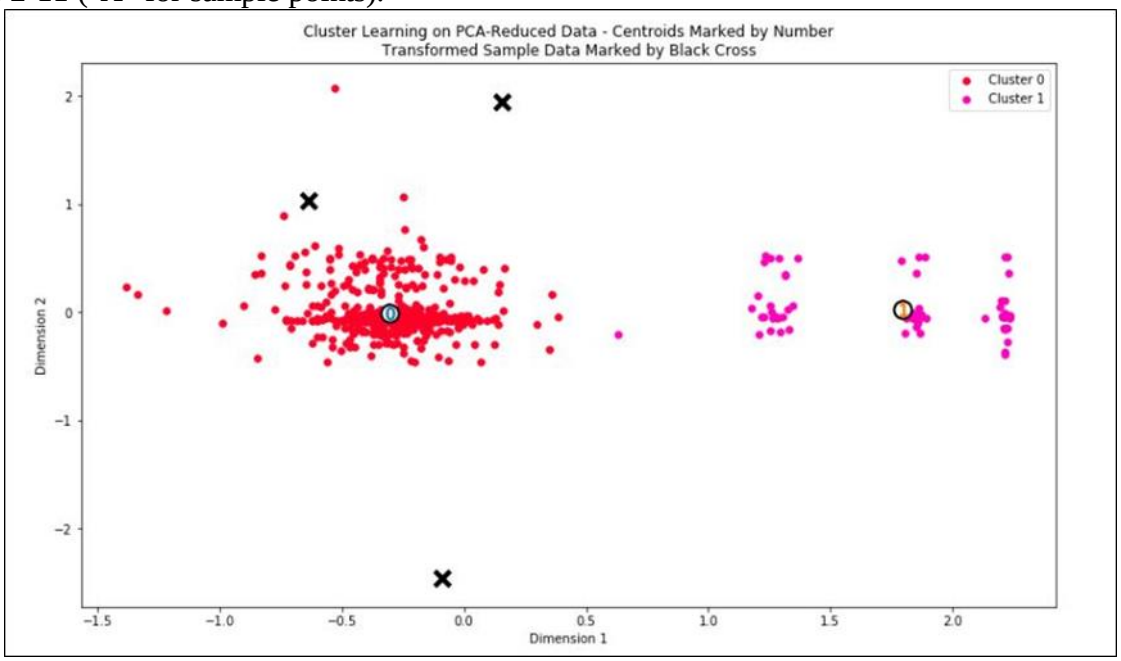

Figure 2-11.1 Results of Cluster Number 2

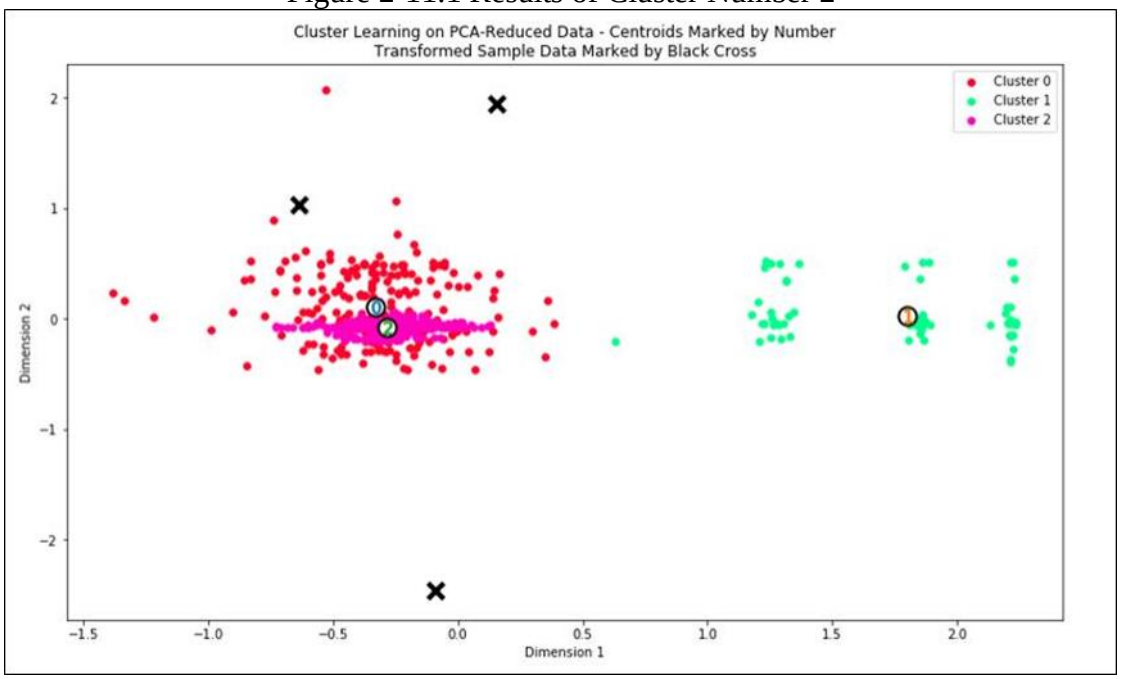

Figure 2-11.2 Results of Cluster Number 3

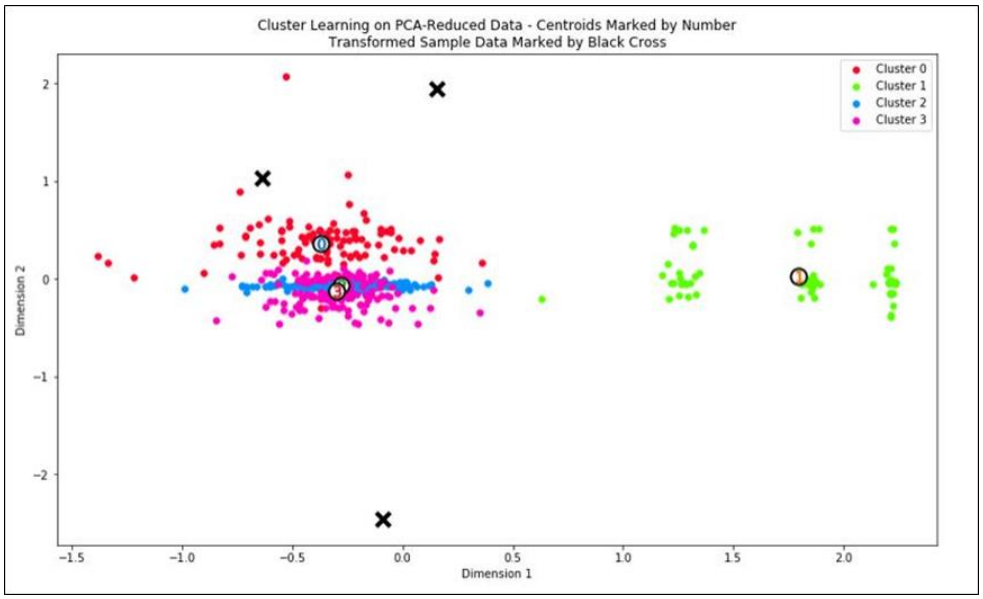

Figure 2-11.3 Results of Cluster Number 4

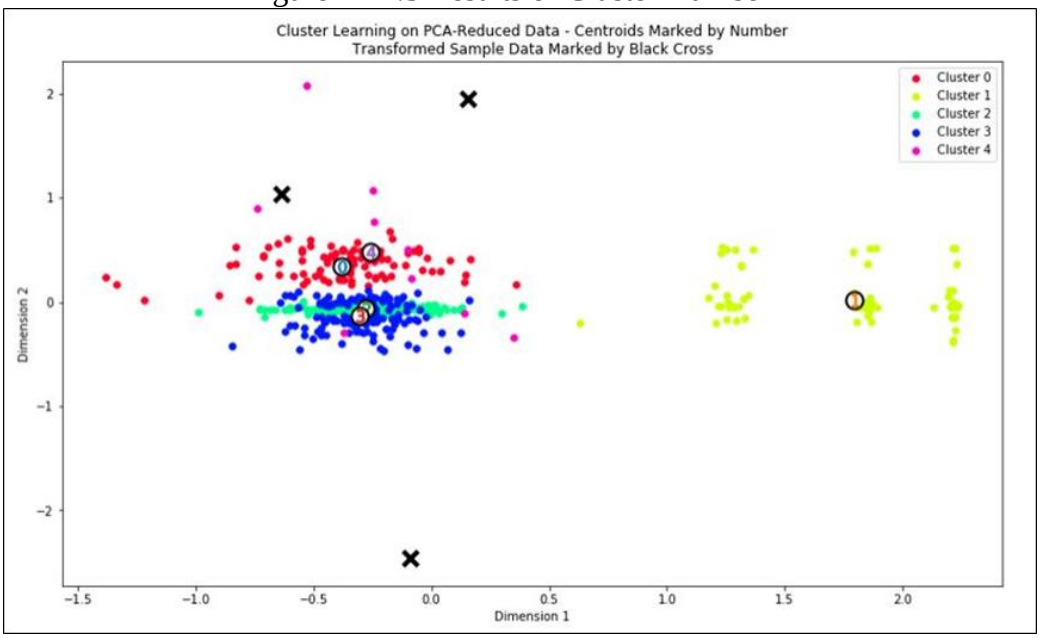

Figure 2-11.4 Results of Cluster Number 5

The above are respectively the visualization results of the number of clusters of 2, 3, 4 and 5, from which it can be seen that when the number of clusters is 2, the effect is the best.

3. Conclusion

This paper studies the problem of ship type by collecting, processing and analyzing the original data. Finally, the conclusion is that when the number of clusters is 2, the clustering effect is the best, that is, the ship types can be roughly divided into Two categories. Lay the foundation for further research on the type of ship.

References

[1] Zhu Minghan, Luo Dayong, Yi Liqun.A Generalized Principal Component Analysis Feature Extraction Method [J] .Computer Engineering and Applications, 2008 (26): 38-40 + 44.

[2] Zhang Dongmei.Cluster clustering algorithm based on contour coefficient [D] .Yanshan University, 2010.