ISSN 2616-5775 Vol. 4, Issue 6: 1-4, DOI: 10.25236/AJCIS.2021.040601

Research on Semiconductor Quality Prediction Based on Nonlinear Support Vector Machine

Qianyi Cai, Ziquan Li

Changchun University of Science and Technology, Changchun, Jilin, 130022, China

Abstract: *Using a large amount of semiconductor signal data collected by sensors, effective data can be obtained by data cleaning and data mining. With the analysis of intelligent semiconductor detection model, the effective data can find out the problems in the production process of semiconductor manufacturing process in time, so as to improve the process yield and reduce the unit production cost. Firstly, this paper completes the data preprocessing, and then through the comparison of Pearson correlation coefficients, four variables with the largest Pearson correlation coefficients are obtained, which have the greatest influence on the results, so these four variables are taken as key factors. Finally, the nonlinear support vector machine model is established, and the hypersurface model in the input space corresponds to the hyperplane model in the feature space by using Gaussian function as the kernel function, so that the model has stronger robustness and better generalization ability.*

Keywords: *Lagrange Interpolation Polynomial, Pearson Correlation Coefficient, Nonlinear Support Vector Machine*

1. Introduction

Semiconductor manufacturing is a strategic industry in China, and it is also an important part of China's science and technology field, especially in the field of chip manufacturing. With the continuous development of science and technology, a large number of sensors have been used to monitor the production of semiconductors and collect measurement signals. By analyzing the collected measurement signals, the quality of semiconductors can be detected, and the most relevant signals can be selected to determine the factors that have the greatest impact on product quality. [1] Based on these key factors, the process yield can be improved and the unit production cost can be reduced.

2. Detection Model Based on Support Vector Machine.

2.1 Data Preprocessing.

According to the definition of normal distribution, the probability beyond the mean value of 3 is $P(|x-\mu|>3) <= 0.003$, which is a very small probability event. Under normal circumstances, we can conclude that the samples whose distance exceeds the mean value of 3 basically do not exist. Therefore, when the average distance between samples is greater than 3, the samples are considered as outliers. [2]

For missing values in data, deleting data will result in discarding a large amount of hidden related information when the amount of data is small. Therefore, Lagrange Interpolation Polynomial [3] is chosen to fill the lost data to ensure the integrity of the data. Based on the idea that polynomial interpolation is to find the solutions of n-1 linear equations, Lagrange Interpolation Polynomial introduces the concept of basis function, which has certain reliability for the accuracy of interpolation in interpolation interval filled with missing values.

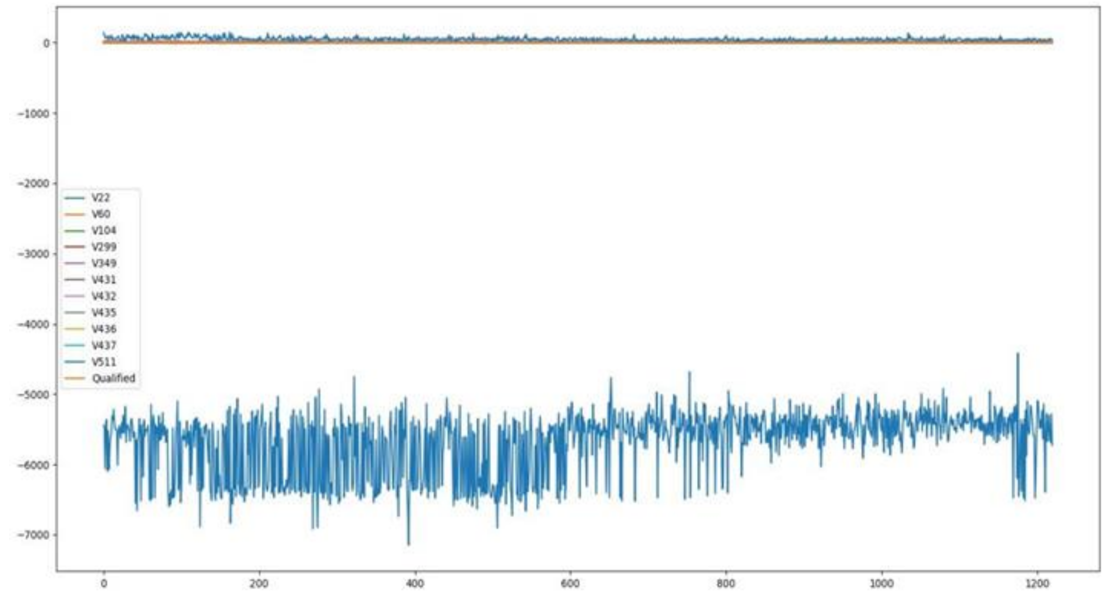

Figure 1: Data after preprocessing.

2.2 Correlation Measurement.

Invoke the Pandas library in Python, and use the corr () function (based on the mathematical principle of Pearson correlation coefficient) to get Pearson correlation coefficient among various variables [4], and draw the heat map through the Matplotlib library in Python, the effect is shown in Figure 2:

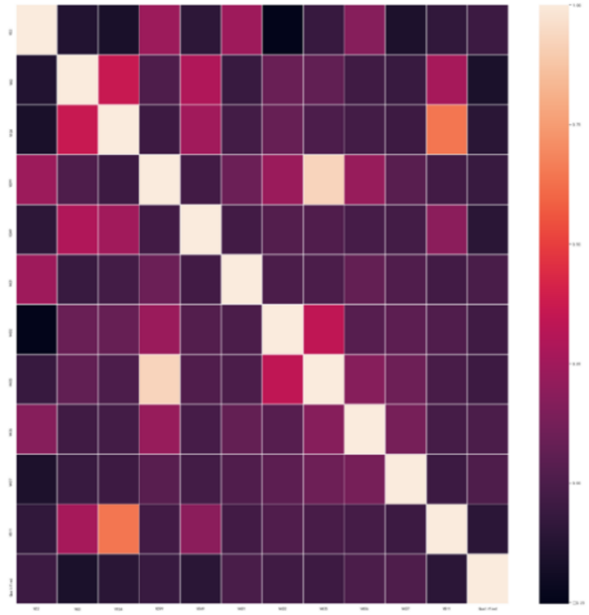

Figure 2: Pearson correlation coefficient diagram between variables.

It can be seen from the figure that the Pearson correlation coefficient between V299 and V435 is 0.92, so it can be inferred that V299 and V435 have strong correlation. The Pearson correlation coefficient of V511 and V104 is 0.64, so it can be inferred that V299 is moderately correlated with V435.

Select four key factors (V60, V104, V349, V511) that have the greatest influence on the dependent variable from the 12 variables obtained based on Pearson correlation coefficient, and draw a three-dimensional scatter chart by using Matplotlib library in Python, as shown in the following figure:

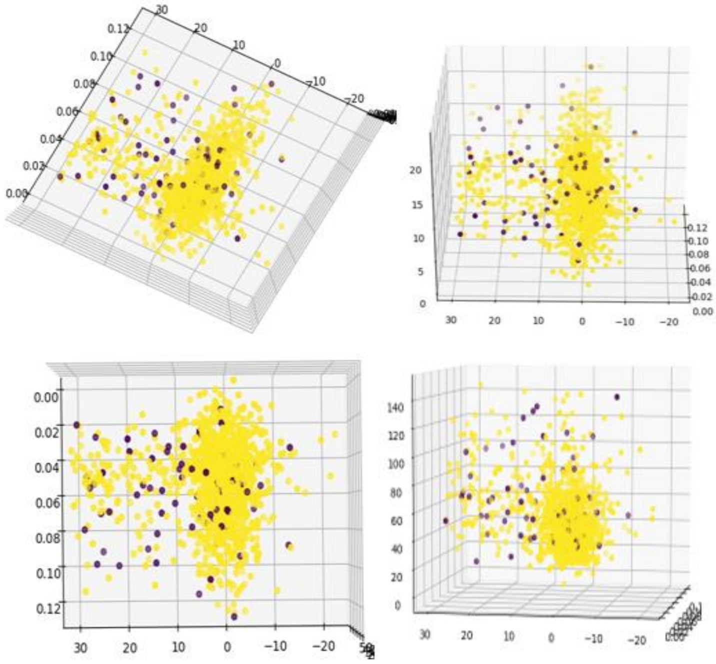

Figure 3: Influence of key variables on results.

The yellow dots in the figure are qualified dots, and the purple dots are unqualified dots. In the above figure, all of them can find a dividing surface to roughly divide qualified products and unqualified products into two parts, so V60, V104, V349 and V511 are the key influencing factors.

2.3 Nonlinear Vector Machine Model.

Considering that the data is not necessarily linearly separable, the sum function is used to find the separation hypersurface to classify the data. Mainly through a nonlinear transformation, the input space corresponds to a feature space, so that the hypersurface model in the input space corresponds to the hyperplane model in the feature space. Gaussian function is selected as kernel function for nonlinear transformation.

$$K(x,z) = \exp\left(-\frac{\| x - z \|^2}{2\sigma^2}\right)$$

The classification decision function can be obtained:

$$f(x) = \operatorname{sign}\left(\sum_{i=1}^{N_S} \alpha_i^* y_i \exp\left(-\frac{\|x - x_i\|^2}{2\sigma^2}\right) + b^*\right)$$

The use of Python to remove all unqualified product data, and each key factor constitutes a plurality of two-dimensional matrices, and draws the scatter plot as follows with the Python Drawing Tool.

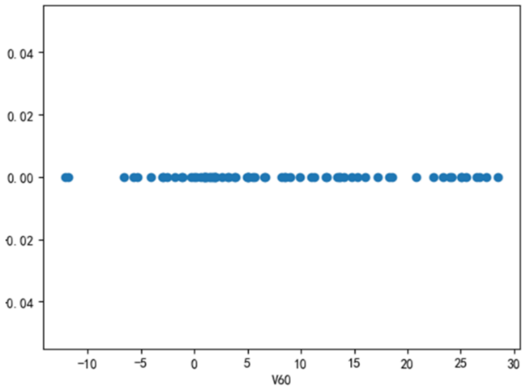

Figure 4: Distribution of all nonconforming products under variable V60.

The range of unqualified products in V60 is (-8,29).

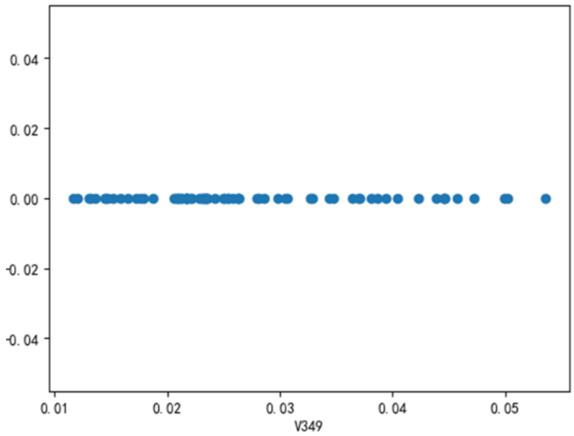

Figure 5: Distribution of all nonconforming products under variable V349.

The range of unqualified products in V349 is (0.01, 0.05).

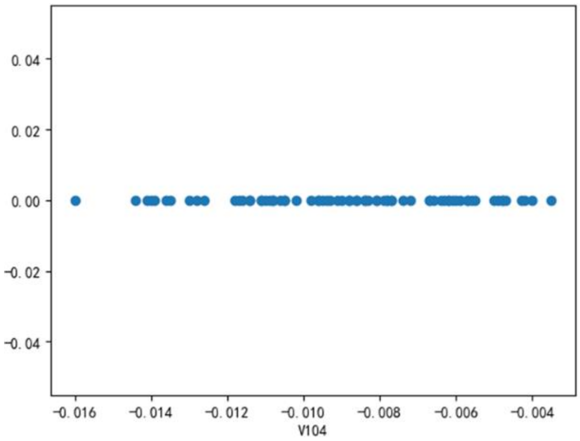

Figure 6: Distribution of all nonconforming products under variable V104.

The range of unqualified products in V104 is (-0.015, -0.003).

3. Conclusion

This is a binary prediction problem for the qualified and unqualified semiconductors. Firstly, the data is preprocessed, and then according to Pierce's correlation coefficient principle, the thermal diagram is used to show the correlation among various variables. For the key variables, the specific key variables are determined by their distribution on the scatter plot. Considering that hypersurfaces have stronger robustness and generalization ability in multidimensional space, we use nonlinear support vector machine as our model, and use Gaussian function as kernel function for nonlinear transformation of linear space, which makes the model have better classification ability.

References

[1] Jiang Qiyuan, Xie Jinxing and Alfred, Mathematical Model, Fifth Edition, Higher Education Press, 2018.
[2] Si Shoukui, Sun Zhaoliang, Mathematical Modeling Algorithm and Application, second edition editor, National Defense Science and Technology Press, 2011.
[3] Li Hang, Statistical Learning Methods, 2nd edition, Tsinghua University Publishing House, 2019.
[4] Wes.McKinney, Data Analysis by Python, 2nd edition, Mechanical Industry Press, 2018.