**Mahalanobis Distance:** Mahalanobis Distance is a statistical measure used to determine the distance between a point and a distribution. It takes into account the correlations between variables, which makes it particularly useful for high-dimensional datasets where variables may be interdependent.

In anomaly detection, Mahalanobis Distance can be used to identify unusual data points that are far from the distribution of normal data points. This is done by calculating the Mahalanobis Distance for each data point and comparing it to a threshold value. If the distance exceeds the threshold, the data point is flagged as an anomaly.

To apply Mahalanobis Distance in anomaly detection, first need to compute the covariance matrix and mean vector of the normal data points. These can be used to calculate the Mahalanobis Distance for each data point. The formula for Mahalanobis Distance is as follows:

Chart, histogram

Description automatically generatedGiven a probability distribution on and positive covariance matrix , the Mahalanobis distance of a point from is

Figure Mahalanobis Distance distribution

Once have computed the Mahalanobis Distance for each data point, we can set a threshold value based on the distribution of distances. Data points with distances greater than the threshold are flagged as anomalies.

**PCA:** Principal Component Analysis. It is a statistical technique used for reducing the dimensionality of high-dimensional data by transforming it into a lower-dimensional space while retaining as much of the original information as possible. It works by identifying the directions in the data that explain the most variance, known as the principal components. These principal components are linear combinations of the original variables, and they are ordered in terms of the amount of variance they explain in the data. The first principal component explains the most variance, followed by the second, third, and so on.

In simulation, I tried components = 2 and 3, the following is the distribution under these two conditions:

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Figure PCA components of 2 and 3

**K-means:** is a popular unsupervised machine learning algorithm used for clustering analysis. The algorithm partitions a dataset into k clusters, where k is a user-defined parameter. It works by iteratively assigning data points to the nearest centroid, which is the centre point of each cluster. The centroid is then updated to the mean of the data points assigned to it. This process is repeated until the centroids no longer change or a maximum number of iterations is reached. It can handle large datasets and is relatively easy to implement. Additionally, it can be applied to a wide range of data types and is widely used in various applications.

In anomaly detection, K-means can be used to identify clusters of normal data points and flag data points that are far from these clusters as anomalies. This approach assumes that the normal data points form dense clusters in the feature space, and anomalies are data points that fall far from these clusters.

We have two classes, one is normal another is

Chart, scatter chart, bubble chart

Description automatically generated

Figure

performance

Chart, scatter chart

Description automatically generated

Figrue Relationship volt rotate pressure vibration

Figure X shows the relationship among the four variables. We can find that the some anomaly samples are messed with normal samples together, which is very hard to classify. And then I tried use

By viewing the anomaly point I found the four variables have a certain correlation. Mahalanobis distance is a good method to identify this correlation. The result just is showed below.

Chart, scatter chart

Description automatically generated

Figrue :Relationship of Mal Distance with others

Compare with Figrue X, we can see that Mahalanobis distance is very helpful.

Next step, I calculated the threshold and do classification. The result is showed in Figure X. In the result, I found there are some continuous anomaly samples. I combine these continuous anomaly samples together as one anomaly sample which is more reasonable and calculate the result. normal 860 failure 15

Chart, line chart, histogram

Description automatically generated

Figrue :The Mal distance and threshold

Chart, timeline

Description automatically generated

Figure X: Mahalanobis classification

PCA is a very powerful tool to reduce the dimension of variables so I want to try reduce the dimension to see whether it is helpful to classification. In figure X with components of 2 and 3, the result shows the anomaly samples, it is hard to extract them from the whole samples. By using K-Mean to illustrate my suppose.

The classification result is shown in figure X. The same, combine the continuous anomaly samples together and visualize the result in Figrue X

Chart, timeline

Description automatically generated

Figure X PCA+K-mean classification

Table The classification metrics

|  |  |  |
| --- | --- | --- |
|  | Mal Distance | K-mean |
| accuracy | 98.17% | 98.06% |
| GMean | 67.03% | 30.03% |
| Recall | 33.33% | 12.50% |
| Specificity | 99.30% | 98.85% |
| Precision | 45.45% | 90.90% |
| F-score | 38.46% | 10.52% |

The result shows the Mahalanobis distance classification performs much better than PCA+K-mean. That’s mainly because the corelation ship among the four variables is high and reduce dimension is not helpful in this case. So, the Mahalanobis distance performs much better.