Characterization and identification of anomalies in geodetic data

Chujia Wang, Zhe Wu, Yunhao Liu, Ruoshi Li

(Tianjin College University of Science and Technology Beijing, Tianjin, China)

ABSTRACT

The shape of the Earth is constantly changing, undergoing certain deformations due to factors such as tidal gravity. To accurately analyze the deformation characteristics of solid tides and earthquake precursors, researchers need to set up observation points in areas vulnerable to natural disasters and human disturbances, such as caves and underground Wells. However, severe weather conditions and human activities can interfere with the deformation signal, causing it to deviate from typical features of solid tidal curves. The existence of local deformation signals also brings challenges to the study and monitoring of deformation characteristics before earthquakes. Therefore, in order to eliminate the interference of external factors, researchers need to take corresponding measures to ensure the reliability and accuracy of observation data.

For problem 1, the method of extending the data we use is linear interpolation and random generation of the data. Linear interpolation uses the linear trend between known data points to estimate the value of new data points, while randomly generated data expands the data by generating new data points with similar distribution and statistical properties. I then use the Fourier transform to analyze the frequency domain properties of the data to evaluate the quality and relevance of the data.

For problem 2, we first observe the relationship between the spectral graph and the line graph to determine the time period in which the noise occurs. Then statistical methods are used to calculate the index, root mean square error (RMSE) and signal-to-noise ratio (SNR) of the sharp fluctuation in the spectrum diagram, and score the linear weight. By scoring, we can see the intensity and interference degree of noise in the data.

For problem 3, we first preprocessed the data, and then calculated commonly used signal eigenvalues. Finally, we used machine learning algorithms such as SUVs and voting machines for classification prediction, and used the preprocessed eigenvalues as inputs for training and classification prediction of the model to achieve a high accuracy of 96.4%.

Through these methods, researchers can more accurately provide effective support for earthquake prediction and disaster prevention and control. At the same time, the continuous improvement and innovation of research methods have also brought new possibilities and opportunities for the development and progress of the earth science field.

Keywords: linear interpolation; random generation; Fourier transform; SVM vector product; voting machine

1 QUESTION RESTATEMENT

1.1 Question background recap

The shape of the Earth is not constant, but undergoes continuous changes akin to the malleability of dough. Scientists employ a diverse range of precision scientific instruments for measuring and documenting Earth's surface deformation, known as Earth deformation observation. These observation sites are often concentrated in areas susceptible to natural disasters and human disturbances, such as caves and underground wells. The solid portion of the Earth experiences certain deformations due to tidal gravity, which can be studied through observations of solid tidal curves.

However, due to the interference of various external factors, the deformation signal on the Earth often deviates from the standard solid tidal curve characteristics. In disaster-prone areas, severe weather conditions (such as severe thunderstorms or precipitation) can cause short-term surface deformation that interferes with solid tide analysis, and these modes are extracted without the need for solid tide processing.[1] In addition, human activities such as manual maintenance or artificial blasting can rapidly disrupt surface deformation, resulting in observed deformation signals that run counter to typical solid tidal curve features. The presence of local deformation signals characterized by rapid distortion, abrupt jumps, and oscillation rusher complicates the analysis of solid tides. Simultaneously, these interference-induced deformation signals also impede the observation of earthquake precursor deformations' characteristics. Preceding an earthquake event, surface deformations may exhibit specific patterns and trends that aid in understanding seismic activity regularities and potential predictions thereof; however external factors' interference hampers the detection of earthquake precursor deformation signals while posing challenges for studying and monitoring pre-earthquake deformation characteristics.

Therefore, in order to accurately analyze the deformation characteristics of solid tides and seismic precursors, it is necessary for researchers to take corresponding measures to reduce or exclude the interference of external factors and ensure the reliability and accuracy of observation data.

1.2 Problem restatement

Problem 1: We are required to reasonably expand the small amount of data in Attachments 1,2 and 4 to 30 items with a sampling rate of 1Hz and constant data length, and discuss the quality of the new data after expansion.

Problem 2: The question asks us to analyze the data in Annex 4 and develop a mathematical model or signal processing method to achieve the overall average distribution of noise in different environmental contexts and calculate the noise profile for each data.

Problem 3: We need to pre-process the signal(noise reduction, truncation, spectral conversion),establish an appropriate recognition model, which can accurately distinguish the four different records in the attachment, and make the average recognition rate above 85%(after more than 100 cycle tests).

2 PROBLEM ANALYSIS

Problem 1: We need to expand the limited amount of data in Annexes 1, 2, and 4 to include 30 entries with a sampling rate of 1 Hz. If the length of the data remains unchanged, then the expanded data should have an adjusted sample rate while maintaining the same number of samples. We can achieve this by interpolating the original data and ensuring that the expansion aligns with a sampling rate of 1 Hz. The quality of the expanded data can be evaluated by comparing it with the original dataset.

Problem 2: Noise levels in various environmental backgrounds significantly interfere with signal processing and feature analysis. For Annex 4's dataset, we can employ signal processing methods and feature analysis techniques to address noise-related issues. By utilizing filters, we can remove or reduce noise while selecting appropriate filter types for effective filtering processes. Additionally, we can develop mathematical models that describe average noise distribution based on different environmental backgrounds' characteristics. Through statistical analysis of both original data and background noise, we can obtain evaluation indicators that depict noise nature and calculate each dataset's specific noise profile for further analysis and processing.

Problem 3: In terms of signal pre-processing, several methods such as noise reduction, truncation, and spectral conversion could be considered. Noise reduction involves using filters to eliminate or minimize unwanted interference caused by noise sources. Truncation can intercept data according to signal characteristics to remove useless information. Spectral conversion converts a signal from the time domain to the frequency domain, using methods such as Fourier transform. To build a suitable recognition model, we can train it on four different records we know, using machine learning algorithms or pattern recognition algorithms for classification and recognition. After more than 100 cycle tests, the average recognition rate needs to reach more than 85% to meet the requirements.

3 MODEL ASSUMPTIONS

1.Assume that the interpolation or other methods used in the process of data expansion can accurately fill in the missing data and maintain the homogeneity of the data.

2.It is assumed that the expanded new data can maintain a similar feature distribution and change trend as the original data.

3.It is assumed that the new data quality after expansion can meet the requirement of 1 Hz sampling rate, that is, the time interval between data points remains consistent.

4.Suppose that we can develop mathematical models or signal processing methods suitable for different environments according to the noise level and environmental background of the attached data.

5.Assume that we can effectively statistically analyze the overall average distribution of noise and describe the characteristics of noise in each environment based on different evaluation indicators.

6.Assume that our model can accurately calculate the noise profile of each data and provide information about noise levels and noise characteristics.

7.Suppose that we can use advanced noise reduction technology to reduce noise interference in the signal to improve the quality of the signal.

8.It is assumed that the truncation and spectral conversion in the pre-processing process can effectively extract the key features in the signal and help distinguish the four different records.

9.Assume that we can establish an accurate recognition model and achieve an average recognition rate of more than 85% accuracy after repeated cycle testing.

10.It is assumed that the performance of the preprocessing and recognition model will not be affected by the increase of data volume or different environmental backgrounds, and can be applied to a wide range of application scenarios.

4 SYMBOL DESCRIPTION

Symbol	Definition
x	Location of data points
y	The value of the data point
dx	Spacing between data points
X	The location of the new data point
Y	The value of the new data point
X_{norm}	Normalized values
$Score$	The score calculated by linear weighting

5 MODEL BUILDING AND SOLVING

5.1 Establishment and solution of problem 1 model

5.1.1 Problem analysis

First, we use linear interpolation to extend the data. Linear interpolation estimates the value of a new data point based on the linear trend between the known data points. That is, by calculating the slope between two known data points, and then calculating the value of the new data point based on this slope and spacing.

We then choose to randomly generate new data points. We can use a random number generation algorithm to generate a new data point with similar distribution and statistical properties to an existing data point.

For the newly extended data, we use Fourier transform to analyze its frequency domain properties. The Fourier transform can convert the data in the time domain to the data in the

frequency domain, thus revealing the energy distribution of the data in different frequencies. By Fourier transforming the extended data and taking its average value, we can get the average energy distribution in the frequency domain. We calculate the correlation between the original data and the extended data by measuring the degree of linear relationship between two sets of data.

Finally, we compare the correlation between linear interpolation and randomly generated data. Compared to the quality of the extended data, the higher the correlation, the better the linear relationship between the extended data and the original data.

5.1.2 Data preprocessing

We pre-process a small amount of data in Annexes 1, 2 and 4, including steps such as outlier removal, data cleaning and de-noising, to ensure the quality of the data.

And according to the characteristics of the data, we choose the appropriate interpolation or filtering method to fill the blank part of the data. Common interpolation methods include linear interpolation, spline interpolation and Kriging interpolation. Common filtering methods include mean filtering, median filtering and Gaussian filtering. Here we choose linear interpolation and random data generation methods to expand to 30 entries.

5.1.3 Linear interpolation algorithm

We first calculate the spacing between the known data points. Assume that the positions of the known data points are x1 and x2, and the corresponding values are y1 and y2. $dx = x2 - x1$.That is, the distance between known data points.

The position of each new data point is then calculated. Suppose to expand to 30 data points, except for the known data points, there are 29 data points that need to be expanded. The 29 data points are evenly distributed between x1 and x2 to get the expanded new data point locations.

Then calculate the value of the new data point. Based on the linear trend between the known data points, as shown in Figure 1, we calculate the value of the new data points by linear interpolation. For the position X of the new data point (between x1 and x2), we calculate the interpolation using the following formula:

$$Y = y1 + (X - x1) * \frac{y2 - y1}{dx} \quad (1)$$

Where y1 and y2 are the values corresponding to the known data points, and Y is the value of the new data point to be calculated.

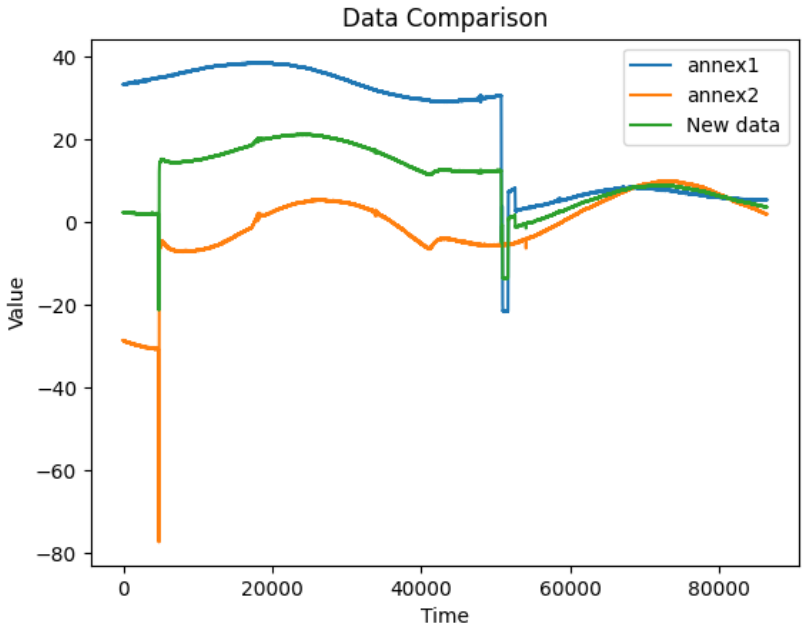


Figure 1: Results of linear interpolation of raw data

Finally, we perform a Fourier transform on each data point to convert data in the time domain to data in the frequency domain. We average all the transformation results to obtain the energy distribution over the average frequency domain, as shown in Figure 2.

```
Data1 Fourier results: [1943604.221          -0.j          -82135.35440662-773492.70089173j
   48190.5716531  -98855.40632568j ... -144048.25213714+130738.89021745j
   48190.5716531  +98855.40632568j  -82135.35440662+773492.70089173j]
Data2 Fourier results: [ -61104.66           -0.j          -44339.35690217+114650.10560842j
 -274054.27382289+291502.81883431j ...  -84126.2761942 -118751.38160897j
 -274054.27382289-291502.81883431j  -44339.35690217-114650.10560842j]
Data3 Fourier results: [ 941249.7805         -0.j          -63237.3556544 -329421.29764165j
 -112931.8510849  +96323.70625432j ... -114087.26416567  +5993.75430424j
 -112931.8510849  -96323.70625432j  -63237.3556544 +329421.29764165j]
```

Figure 2: Fourier Transform results (linear interpolation)

In order to determine the degree of linear relationship between the original data and the extended data, we use Pearson correlation coefficient as the calculation method of correlation. Due to the large amount of data, we choose to extract a small amount of data from the original data and extended data for observation and analysis. Here, data 1 and data 2 in Annex I are selected for analysis.

By calculating the correlation coefficient, we can get a value between -1 and 1. According to the value of the correlation coefficient, we can judge the strength of the linear relationship between the two data sets. If the correlation coefficient is close to 1, it indicates that there is a strong positive correlation between the two data sets. If the correlation coefficient is close to -1, it indicates that there is a strong negative correlation. If the correlation coefficient is close to 0, there is no linear relationship between the two data sets.

By observing and analyzing the correlation between data 1 and data 2 in Annex I, we can get the results shown in Table 1. Table 1 shows the scatter plot between the two data sets and

the values of the correlation coefficients. According to the value of the correlation coefficient, we can preliminarily judge that the linear relationship between the two data sets is strong.

Table 1：Correlation calculation results (linear interpolation)

Correlation between data1 and data3	0.796255898
Correlation between data2 and data3	0.254584295

5.1.4 Random number generation algorithm

Random numbers are composed of numbers with the same nature, and a single or several numbers cannot be said to be random numbers, so random numbers are generally called random numbers or random sequences.[2] We generate random numbers that fit a particular distribution based on the statistical properties, mean, and standard deviation of the known data points. A random number generation algorithm is then used to generate new data points with similar distribution and statistical properties to existing data points. Here, we use a normal distribution to generate values for new data points. The new data points generated should have distribution characteristics similar to the existing data points, as shown in Figure 3.

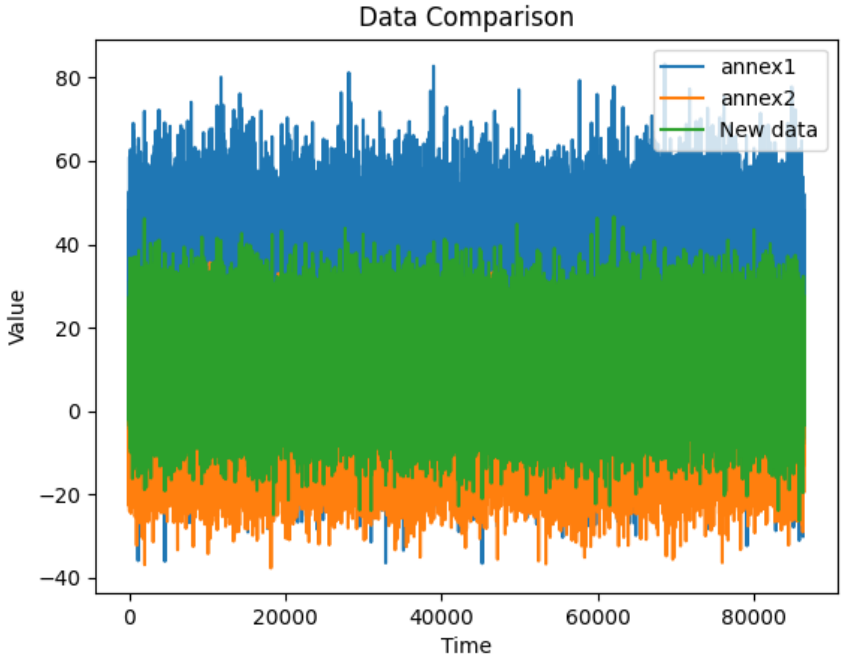


Figure 3：Results of raw data in randomly generated data

Although various algorithms apply the principle of Fourier transform is different, but there is a common advantage, that is, the operation speed is relatively fast.[3] We perform a Fourier transform on each data point in the raw data, converting the time domain data into frequency domain data. In this process, we convert the data from the time domain to the frequency domain, thus revealing the characteristics of the data at different frequencies.

Specifically, the Fourier transform breaks the signal down into a series of frequency components that can represent the energy distribution of the data at different frequencies.

In this process, we apply the Fourier transform to each data point and get the corresponding frequency domain data. Then, we average all the transformation results to get the energy distribution over the average frequency domain. The purpose of this is to obtain the energy distribution of the data over the entire frequency domain to get a more complete understanding of the characteristics of the data at different frequencies.

By observing the Fourier transform results in Figure 4, we can analyze the energy distribution of the data at different frequencies, and then understand the characteristics of the data at different frequencies.

```
Data1 Fourier results: [ 1.94606152e+06   -0.j          6.65956208e+03 +633.81296677j
  1.59910454e+03+1821.46631281j ... -3.41171269e+03-1838.85406791j
  1.59910454e+03-1821.46631281j  6.65956208e+03 -633.81296677j]
Data2 Fourier results: [-56865.44766121   -0.j          -3666.01098756-2268.84854456j
  -5975.63421102-4557.64693235j ...   -747.4657852 +3391.10768571j
  -5975.63421102+4557.64693235j  -3666.01098756+2268.84854456j]
Data3 Fourier results: [944598.03729623   -0.j           1496.7755468  -817.51778889j
  -2188.26483642-1368.09030977j ...  -2079.58923967 +776.1268089j
  -2188.26483642+1368.09030977j   1496.7755468  +817.51778889j]
```

Figure 4：Fourier Transform results (randomly generated)

To determine the degree of linear relationship between the original and extended data, we use covariance and correlation coefficients to calculate the correlation between them.[4] By calculating the covariance of the original and extended data and dividing it by their respective standard deviations, we get the correlation coefficient. The correlation coefficient ranges from -1 to 1, with closer to 1 or -1 indicating a stronger correlation.

In the calculation of correlation, we use Pearson correlation coefficient as an index to measure the strength of correlation.[5] Due to the large amount of data, if the correlation coefficient of all data is calculated directly, the data points will overlap and be difficult to observe. Therefore, data 1 and 2 in Annex I were selected as simple representative samples for observation. By calculating the correlation coefficient between these two data, we get the results shown in Table 2. Table 2 shows the correlation between these two data points.

According to the phase relationship values in Table 2, we can know the degree of linear relationship between data 1 and data 2 and data 3. The correlation coefficient close to 1 means that there is a strong positive correlation between the two data, and the correlation coefficient close to -1 means that there is a strong negative correlation. By observing the correlation results in Table 2, we can preliminarily judge that the degree of linear relationship between the original data and the extended data is strong.

Table 2：Correlation calculation results (randomly generated)

Correlation between data1 and data3	0.847192424
Correlation between data2 and data3	0.531688413

5.1.5 Model solving

Fourier transform is applied to linear interpolation algorithm and random data generation algorithm, and the results are compared with the results of correlation calculation. We made a surprising finding that when the algorithm scaled using randomly generated data, the new data quality performed best, up to 84.72%. This result reveals to us an important fact that randomly generated data algorithms have significant advantages in extending data quality.

5.2 The establishment and solution of problem two model

5.2.1 Problem analysis

First, we observe the spectrogram of the attached data and analyze the relationship between the spectrogram and the line graph. By looking at areas of the spectrum where the value fluctuates wildly, we can determine the time period when the noise appears.

For each piece of data, we use statistical methods for noise identification. Calculate indicators of sharp fluctuations in the value of the spectrum diagram, such as root mean square error (RMSE) and signal-to-noise ratio (SNR). For each data, its RMSE and SNR metrics are calculated. According to the definition of indicators, RMSE is regarded as a very large indicator and SNR is regarded as a very small indicator.

According to the calculated RMSE and SNR indicators, we performed a linear weighted score. Multiply RMSE by weight factor w1 and SNR by weight factor w2, then add the weighted RMSE and SNR together to get the final score.

5.2.2 Spectrum analysis

In order to further analyze the frequency domain characteristics of the data, we transform the data in each time window to obtain the corresponding spectral diagram. Fourier transform transforms the time-domain signal into frequency-domain representation, decomposes the data into components of different frequencies, and reveals the energy distribution of the data at different frequencies.

In this process, we divide the data into multiple time Windows and Fourier transform the data within each time window. By applying the Fourier transform to the data of each time window, we can obtain the corresponding spectral diagram, where the horizontal axis represents the frequency and the vertical axis represents the energy.

Since there are many spectral graphs, we choose to show the first four spectral graphs as examples here, and the specific results are shown in Figure 5. This is done to present the frequency domain characteristics of the data over different time periods, where each graph represents a spectral graph of the data within a time window.

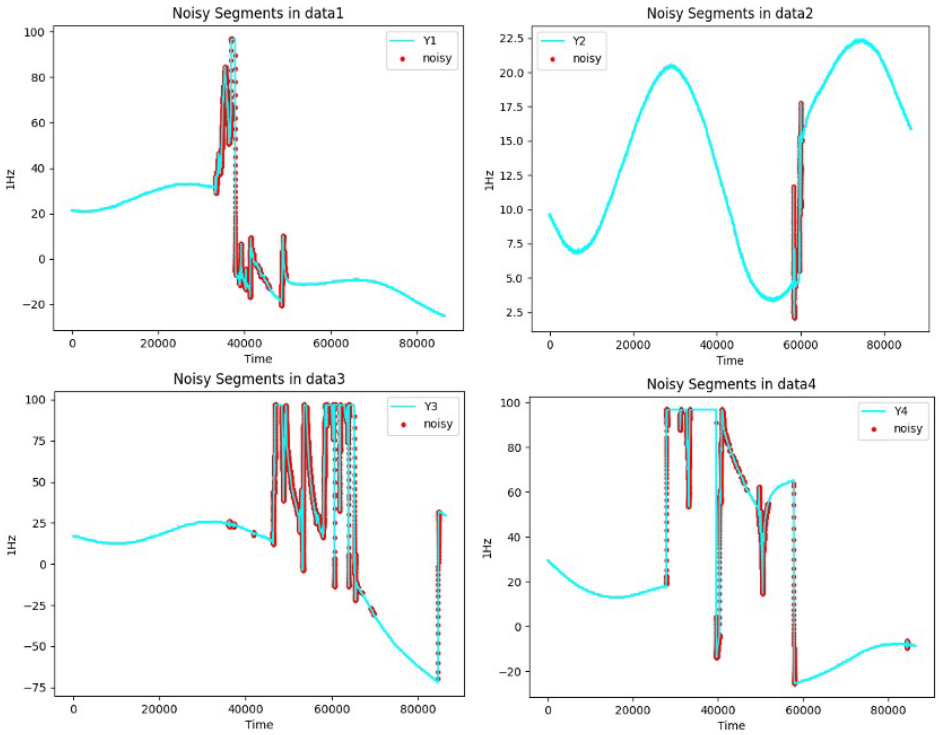


Figure 5：Spectrum diagram

By looking at the spectrum diagram in Figure 5, we can distinguish the different characteristics of noise and signal in the frequency range. Noise appears as a random distribution of energy on a spectrogram, while signals appear as having concentrated peaks and energy regions on the spectrogram.

To evaluate the noise profile of the data within each time window, we need to compare the statistical features on the spectrogram. By comparing the statistical features of different spectrographs, we can assess the presence and intensity of data noise within each time window.

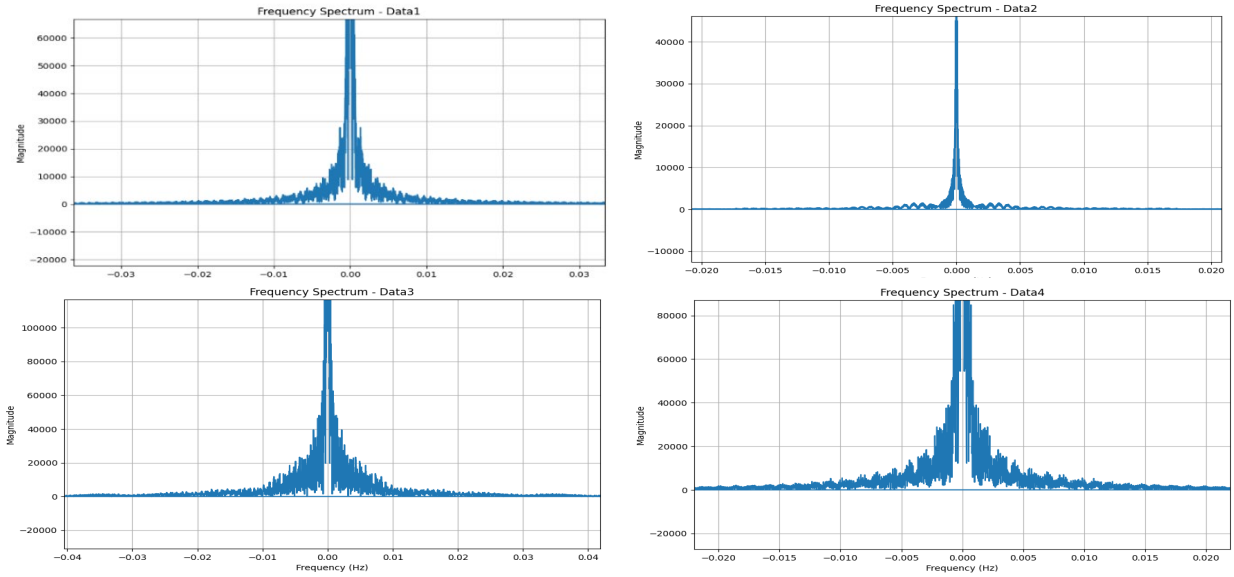


Figure 6：Noise segment

In Figure 6, we show the noise spectra of data 1-4. In order to highlight the noisy areas, we have marked the noise segments in red. By observing the performance of the noise segment on the spectrogram, we can judge the characteristics of the noise in the frequency range. According to the random distribution of noise and the average energy, we can distinguish the different performance of noise and signal on the spectrum diagram.

By evaluating the noise profile of the noise spectrum diagram in Figure 6, we can identify and analyze the noise components in the data. This helps us to more fully understand the characteristics of the data and provides an accurate basis for subsequent data processing and analysis.

5.2.3 Normalized processing

In order to evaluate the size of the noise, we use the mean square error (MSE) as a suitable evaluation index. The mean square error measures the difference between the noise and the original signal, and the greater the value, the greater the noise amplitude.[6] Using the mean square error, we can compare the amount of noise in different data attachments.

Table 3：mean square error values

Date	RMSE
date1	0.31784336
date 2	0.09986026
date 3	0.72106163
date 4	0.79058307
date 5	0.10215769
date 6	0
date 7	0.22146196
date 8	0.31576255
date 9	0.13276701
date 10	0.5001058
date 11	0.30105847
date 12	0.30763667
date 13	1

From the mean square error values in Table 3, we can observe that Annex 13 has the largest mean square error, which means that the noise amplitude in this annex is larger. In contrast, Annex 6 has the smallest mean square error, indicating a smaller noise amplitude in this annex.

In addition to the mean square error, we can also use the signal-to-noise ratio (SNR) to evaluate the proportional relationship between noise and signal. The signal-to-noise ratio reflects the relative strength of signal and noise.

Table 4: Signal-to-noise ratio

Date	SNR
date1	0.96733522
date 2	0.24683014
date 3	0.95819669
date 4	0.97343707
date 5	0.732208
date 6	0.94647987
date 7	0.75339011
date 8	0.85116933
date 9	0.74840426
date 10	1
date 11	0.89986472
date 12	0
date 13	0.9769951

By comparing the signal-to-noise ratio in Table 4, we can observe that Annex 2 has the largest signal-to-noise ratio, indicating that the noise in this annex is relatively small. On the other hand, Annex 10 has the smallest signal-to-noise ratio, indicating that the noise in this annex is relatively large.

By using the mean square error and the signal-to-noise ratio, we can quantitatively estimate the magnitude of noise in different data attachments.

We first need to clarify the importance of the root mean square error (RMSE) and signal-to-noise ratio (SNR) metrics, which determine their weight in the final score. Set weight factors for them, usually the sum of weights is 1, indicating relative importance.

For very large indicators, we use Min-Max Normalization, and for very small indicators, we use maximum-minimum normalization.

Next, we first normalize the very large indicators, using the method of Min-Max Normalization. The purpose of normalization is to map its range to between 0 and 1. The formula is:

$$Normalized\ Value = \frac{Original\ Value - Min\ Value}{Max\ Value - Min\ Value} \quad (2)$$

We then normalize the very small indicators, using the method of Max-Min Normalization. The formula is:

$$Normalized\ Value = \frac{Max\ Value - Original\ Value}{Max\ Value - Min\ Value} \quad (3)$$

Next, we calculate the evaluation score by linear weighting according to the set weighting factors. Take two indices X and Y, whose normalized values are X_{norm} and Y_{norm}. The formula for calculating the evaluation score by linear weighting is as follows:

$$Score = w_X \cdot X_{norm} + w_Y \cdot Y_{norm} \quad (4)$$

Where w_X and w_Y are the weights. Usually, the sum of weights is 1, indicating relative importance.

5.2.4 Model solving

Finally, we multiply the normalized values of RMSE and SNR according to the corresponding weight factors and add them to obtain the weighted RMSE and SNR values. These two weighted index values are then added to obtain the final score, as shown in Table 5.

Table 5: Weighted score results

0.64258929	0.28233675	0.63802002	0.64564022	0.52502568	0.63216161	0.53561673
0.53359774	0.17334520	0.52902848	0.53664867	0.41603413	0.52317007	0.42662519
0.84419842	0.48394588	0.83962916	0.84724935	0.72663481	0.83377075	0.73722587
0.87895914	0.51870660	0.87438988	0.88201007	0.76139553	0.86853147	0.77198659
0.53474645	0.17449391	0.53017719	0.53779738	0.41718284	0.52431878	0.42777390
0.48366761	0.12341507	0.47909834	0.48671854	0.36610400	0.47323994	0.37669505
0.59439859	0.23414605	0.58982933	0.59744952	0.47683498	0.58397092	0.48742604
0.64154889	0.28129635	0.63697962	0.64459981	0.52398528	0.63112121	0.53457633
0.55005111	0.18979857	0.54548185	0.55310204	0.43248750	0.53962344	0.44307856
0.73372051	0.37346797	0.72915125	0.73677144	0.61615690	0.72329284	0.62674795
0.63419685	0.27394430	0.62962758	0.63724777	0.51663323	0.62376917	0.52722429
0.63748594	0.27723340	0.63291668	0.64053687	0.51992233	0.62705827	0.53051339
0.98366761	0.62341507	0.97909834	0.98671854	0.86610400	0.97323994	0.87669505

By using the weighted scoring method, we can consider the RMSE and SNR indicators to evaluate the noise level and signal quality of different accessories more comprehensively.

5.3 The establishment and solution of problem three model

5.3.1 Problem analysis

We first preprocess the data, including noise reduction, truncation and spectral conversion. Noise reduction can remove the interference of noise to the signal, and truncation can remove the invalid part of the signal. Spectral conversion converts the signal into the frequency domain for further analysis and feature extraction.

And then we compute the eigenvalues. Mean, standard deviation, kurtosis, skewness, root-mean-square error and signal-to-noise ratio are commonly used signal characteristic values. The mean is the mean value of the signal, the standard deviation reflects the fluctuation range of the signal, the kurtosis describes the sharpness of the signal, the skewness indicates the skew of the signal, the RMSE is the root-mean-square error, reflecting the difference between the signal and the reference value, and the signal-to-noise ratio is the ratio of the signal to the noise.

The recognition effect and stability of AdaBoost ensemble learning are the best, but the algorithm efficiency is poor, and the algorithm of decision tree is poor in stability but high in efficiency.[7] From the perspective of comprehensive algorithm performance, the application prospect of random forest is great and it has certain practical value. Here we use support vector product and random forest for classification. SVM (Support Vector Machine) is a common machine learning algorithm that can be used to perform binary or multi-classification tasks. A random forest is an ensemble learning method that builds multiple decision trees and classifies them using methods such as voting or averaging. We take the pre-processed eigenvalues as inputs, train the model and make classification predictions, so that the accuracy of the final evaluation reaches more than 85%.

5.3.2 Extract feature

1. Mean: Calculate the average value of the signal and describe the overall level of the signal.

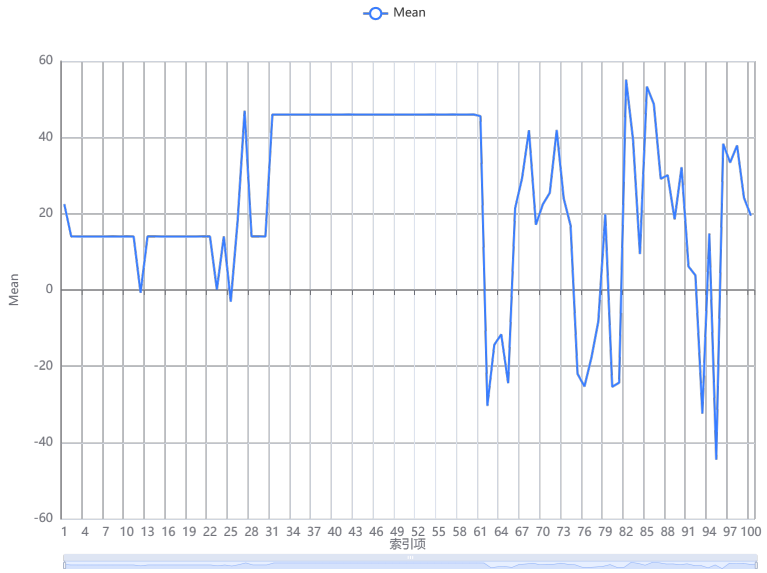


Figure 7: Line chart of the mean value

2. Standard deviation: Calculate the standard deviation of the signal, which is used to describe the change range of the signal, that is, the degree of fluctuation of the signal.

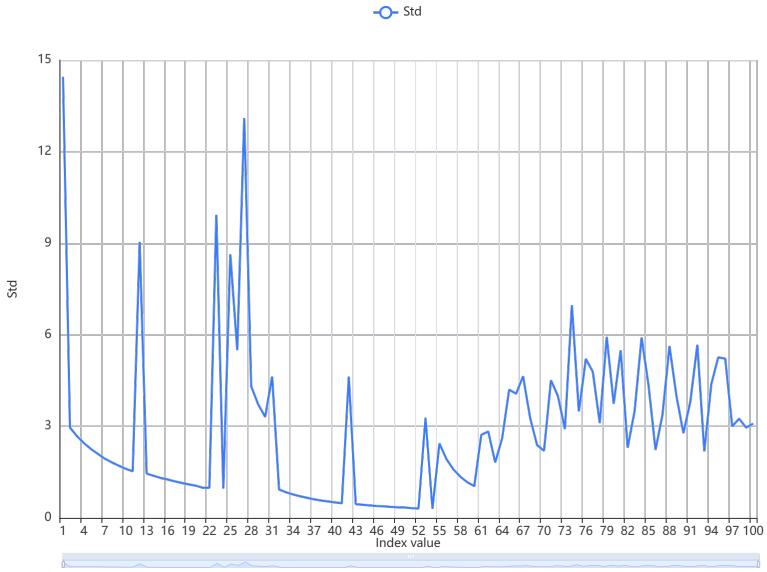


Figure 8: Line chart of standard deviation

3. Kurtosis: describes the peak degree of the signal and reflects the distribution pattern of the signal. Commonly used kurtosis indexes include fourth order kurtosis and kurtosis coefficient.

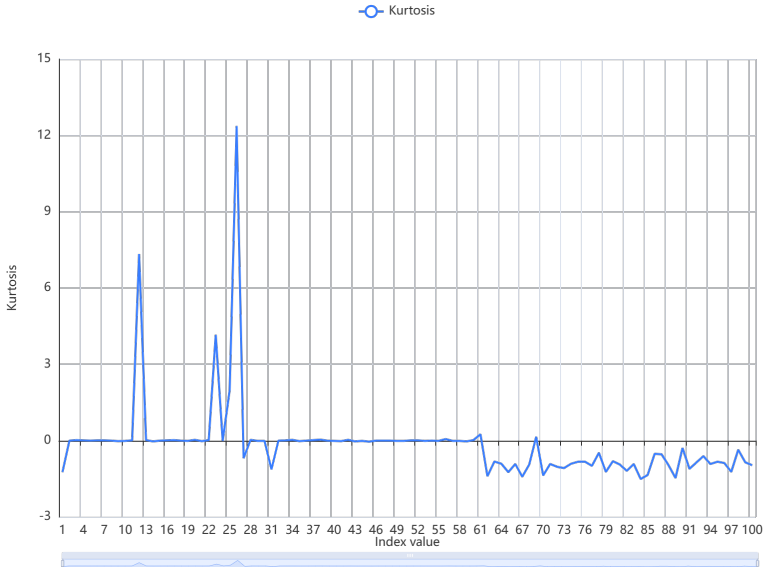


Figure 9: kurtosis line diagram

4.Skewness: describes the degree of skewness of the signal, that is, the symmetry of the signal. The commonly used skewness indexes include third-order skewness and skewness coefficient.

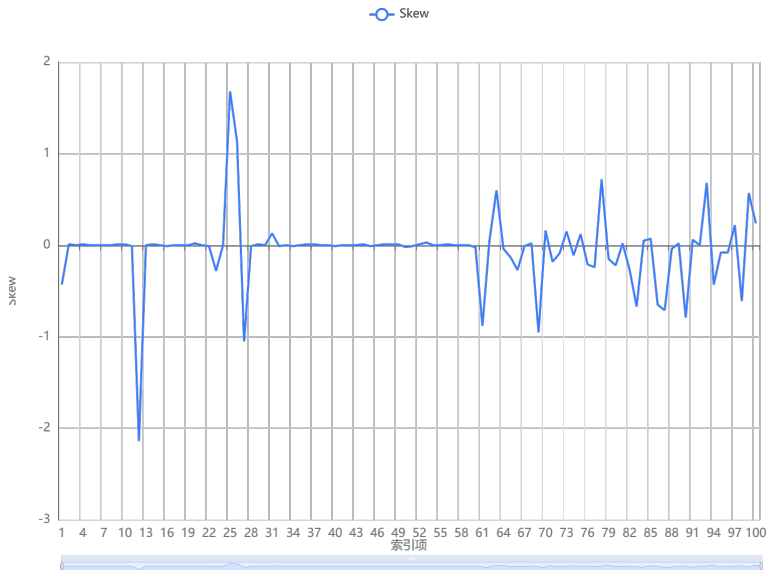


Figure 10: Deflection line diagram

5.Root-mean-square error: Calculate the difference between the signal and the reference value to measure the fitting error.

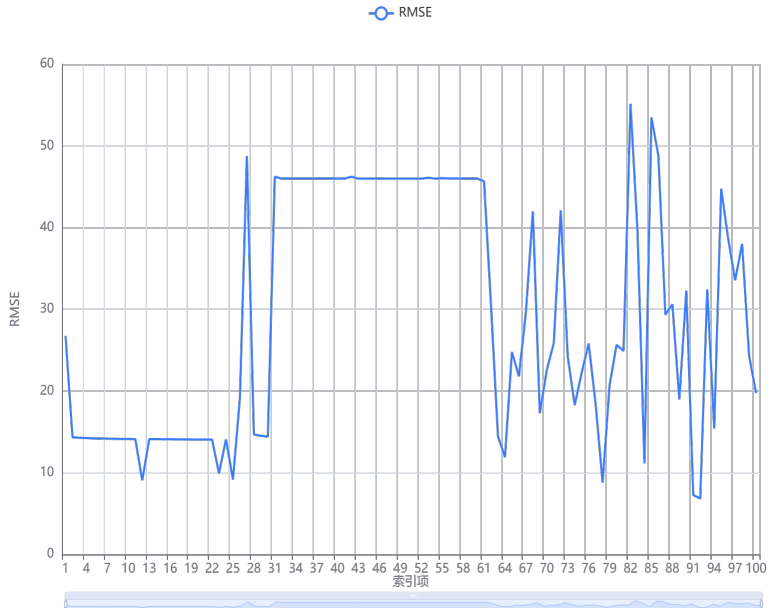


Figure 11: Line chart of root-mean-square error

6. Signal-to-noise ratio: calculate the ratio between signal and noise, reflecting the relative strength of signal and noise.

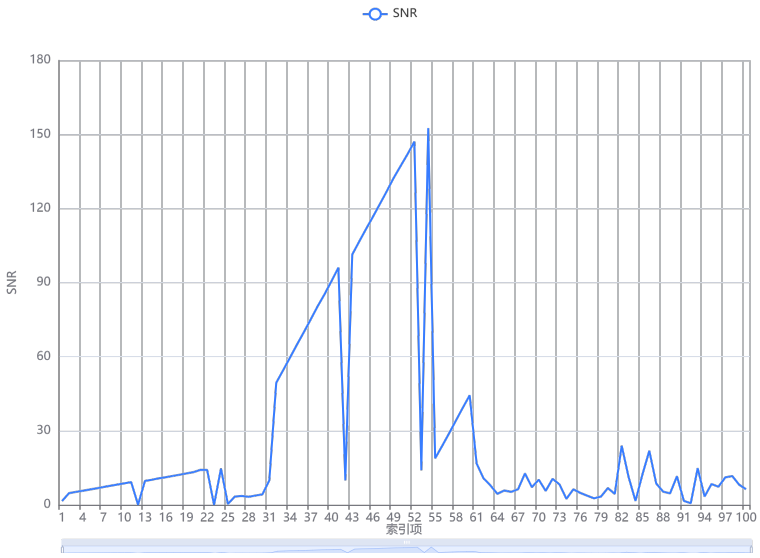


Figure 12: Line diagram of SNR

5.3.3 Model solving

In this study, we propose an ensemble learning-based classification algorithm that aims to improve classification accuracy by integrating the advantages of various machine learning models. At the heart of the algorithm is a soft voting classifier, which combines four different basic models: Support vector machines (SVM) with a linear kernel, extreme Gradient Lift (Boost), random forests, and AdaBoost. AdaBoost technology is continuously used to update the sample weight distribution to increase the weight value of difficult to distinguish samples, and then several weak classifiers are trained into a strong classifier to improve the anti-interference accuracy of the strong seismograph [8]. As shown in Figure 13, we present a mind map.

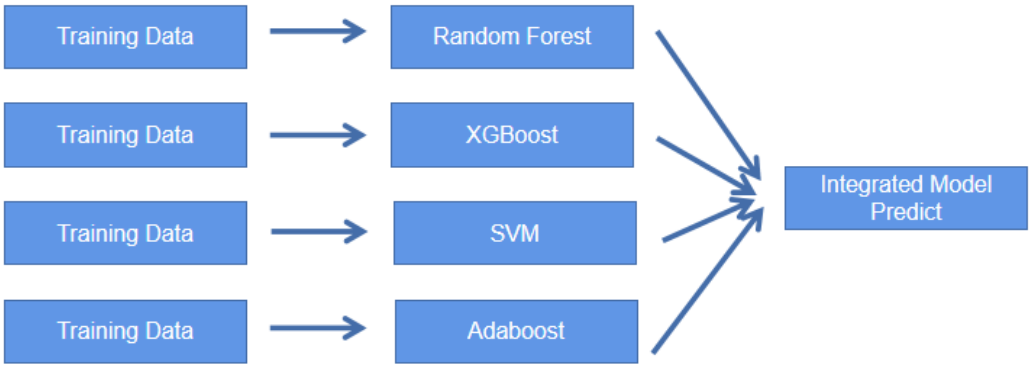


Figure 13: Mind map based on voter analysis

During the implementation of the algorithm, we first perform data splitting, randomly dividing the dataset into training and test sets to ensure that the model is trained and tested on different data subsets. In addition, each base model is initialized with default parameters, with the SVM model set to probability=True to support the soft voting strategy.

The choice of a soft voting mechanism is based on the assumption that different models may have different predictive advantages on different data samples. By calculating the average of the output probabilities of each model as the basis for the final decision, the aim is to combine each model's assessment of uncertainty, thereby improving overall predictive performance.[9] During the model training phase, we train the voting classifier using the training set and then make predictions on the test set. The performance of the model is evaluated by calculating the prediction accuracy, and these accuracies are recorded over one hundred iterations for subsequent analysis.

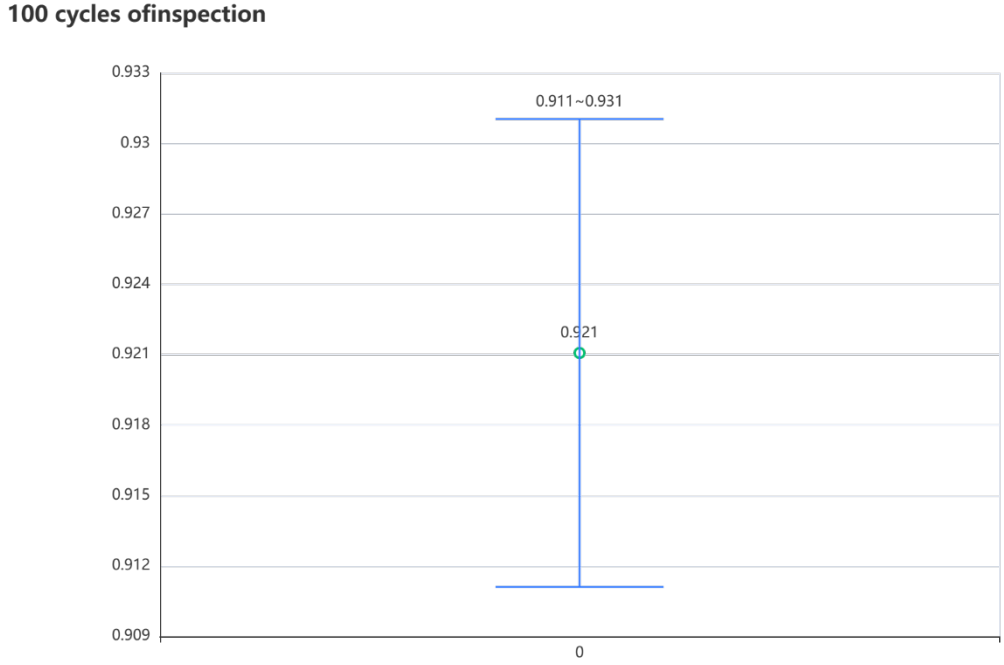


Figure 14: Results of 100 training iterations

6 EVALUATION OF THE MODEL

6.1 Advantages of the model

6.1.1 Advantages of linear interpolation

1. Linear interpolation algorithm is one of the most basic and simplest interpolation methods, which only requires us to understand and implement simple linear functions.

2. The calculation speed of linear interpolation algorithm is very fast, which is suitable for processing a large amount of data.

3. The linear interpolation algorithm has good adaptability to the data with obvious linear relationship and can provide more accurate interpolation results.

4. The linear interpolation algorithm can be easily extended to high-dimensional cases and is suitable for the interpolation of multidimensional data.

6.1.2 Advantages of random number generation

1. The random number generation algorithm can generate random numbers with preset characteristics, such as uniform distribution, normal distribution, etc., so that the generated random numbers conform to the preset probability distribution.

2. Random number generation algorithm usually has the characteristics of simple implementation and easy to use, so we can easily apply to a variety of numerical simulation, random sampling and cryptography and other fields.[10]

6.1.3 Advantages of Fourier transform

1. Fourier transform can decompose the time domain signal into a series of sine and cosine functions of different frequencies, so as to carry out spectrum analysis of the signal. This allows us to better understand the frequency components of the signal.

2. The Fourier transform is linear, so it is easy to add, subtract and multiply signals, which facilitates the design of signal processing and communication systems.

3. Fourier transform has good mathematical properties, such as translation, expansion, symmetry and other properties, which make the method widely used in signal processing, image processing and other fields.

4. Fourier transform can be used for signal filtering and denoising. By zeroing or suppressing unwanted frequency components in the frequency domain, the required signal information can be effectively extracted.

6.1.4 Advantages of vector machine model

1. SVM has good robustness to noise. Since SVM is based on the principle of spacing maximization, it is more concerned with separating samples correctly and has relatively little impact on noise.

2. SVM can map data from original space to high-dimensional feature space by introducing kernel function, which helps to classify noise in the case of linear indivisibility.

3. SVM uses the idea of support vector in optimization problems, and only some samples play a key role in the construction of decision functions. This makes the SVM less influential on noise points and more computationally efficient on large data sets.

6.2 Disadvantages of the model

6.2.1 Disadvantages of linear interpolation

1. The linear interpolation algorithm cannot capture the nonlinear relationship between the data, and the interpolation results for the nonlinear data may be inaccurate.

2. The interpolation results obtained by linear interpolation may not be smooth enough in terms of continuity, resulting in large transitions or oscillations in the interpolation results.

3. The linear interpolation algorithm has high requirements on data distribution, and the data points must be distributed relatively evenly on the definition domain, otherwise the interpolation result may be biased.

4. The linear interpolation algorithm is not robust enough for noisy data, and it is easy to be interfered by noisy data in the interpolation process.

6.2.2 Disadvantages of random number generation

1. Some common random number generation algorithms have a long period, that is, random numbers generated in a long period of time will appear periodic repetition, which may affect some applications with high randomness requirements.

2. The output of the random number generation algorithm usually depends on the seed (the random number used to initialize the algorithm), if the seed is improperly selected, it may produce different random number sequences or be susceptible to external factors.

3. Since computers operate based on deterministic operations, random number generation algorithms cannot generate real random numbers. The generated random numbers are pseudo-random sequences whose randomness is generated by mathematical models during the operation of the algorithm.

6.2.3 Disadvantages of Fourier transform

1. Fourier transform can only process periodic signals, and non-periodic signals need to be truncated and extended, which may introduce additional errors.

2. The Fourier transform cannot handle non-stationary signals, that is, the statistical properties of the signals change over time. For non-stationary signals, other time-frequency analysis methods are required.

3. Fourier transform is sensitive to noise, and the presence of noise will affect the accuracy of spectrum analysis.

6.2.4 Disadvantages of support vector machines

1. When SVM is sensitive to noise, it may overfit noise during training, resulting in wrong decision boundaries. Especially when the proportion of noise samples in the data set is high, it may be necessary to adjust parameters or increase noise filtering steps.

2. When processing highly noisy data sets, SVM may be affected by noise points, resulting in degraded performance of the classifier.

3.SVM has high computational complexity for data sets with a large number of noise points. Since the selection of support vectors is less affected by noise points, a large number of support vectors need to be calculated, resulting in a longer time for model training and prediction.

REFERENCES

[1] Yu H Y. (2005). Application Research of Digital seismic Data (Ph. D. Dissertation, Tongji University).

[2] Tan, Y. (2008). Research on random number generation algorithms. Master's thesis, Hunan Normal University.

[3] Zhang Z H, Wang H Y, Liu Z Q, Huang M, Liu F F, Yu M J & Zhao B Q. (2012). Phase unwrapping algorithm based on fast Fourier transform. Progress in Laser and Optoelectronics (12), 62-68.

[4] Fan C L. (2019). Data-driven image perception and quality assessment methods PhD thesis, University of Chinese Academy of Sciences (Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences).

[5] Li, Z. (2022). Research on Optimization methods of Automatic Machine Learning for Small Sample Numerical Table Data; Doctoral dissertation. Beijing university of science and technology).

[6] Li X Q. (2020). Machine Learning-based Radar Radiation source Sorting and recognition technology research Doctoral dissertation, National university of defense technology).

[7] Pang C, Jiang Y, Liao C W, Wu T & Ding W. (2020). A comparative study on anti-jamming methods of strong vibration monitoring environment based on machine learning. Inland Earthquakes (02), (2), 119-124.

[8] Pang Cong, Jiang Yong, Liao Chengwang, et al. *Research on Anti-interference Technology of Strong Vibration Observation Based on AdaBoost Ensemble Learning*. Sichuan Earthquake, 2020, (04): 14-18.

[9] Pang C, Jiang Y, Liao C W, Wu T, Ding W & Wang L. (2020). Research on anti-interference technology of strong vibration observation based on AdaBoost ensemble learning. Sichuan Earthquake(04),2007,(2) : 213-220. 14-18.

[10] Xue L D. (2021). Doctoral Dissertation on Blockchain Consensus Algorithm and Its Application, University of science and technology of China).