Intelligent de-noising framework and algorithms for distributed full-waveform induced polarization data

Wei-Qiang Liu, College of Geophysics, China University of Petroleum-Beijing, Beijing, 102249, China;*
Qing-Tian Lü, SinoProbe Center, Chinese Academy of Geological Sciences, Beijing, 100037, China;
Pin-Rong Lin, The Key Laboratory of Geophysical EM Probing Technologies, MNR of Institute of Geophysical and Geochemical Exploration of Chinese Academy of Geological Sciences, Langfang, 065000 China;
Ru-Jun Chen, School of Geosciences and Info-Physics, Central South University, Changsha, 410083, China

Summary

Induced polarization (IP) is an effective geophysical method for characterizing near-surface complex resistivity structures. To improve the precision and efficiency of massive-scale data processing, an intelligent signal processing technology based on machine learning is proposed. First, a database containing pure IP signals and simulated noise interference was generated. Then, a support vector machine classifier was trained to identify the noise by using the statistical characteristics of contaminated time series as inputs and the noise interference types as labels. Finally, four kinds of targeted signal processing technologies were integrated into a de-noising library to automatically separate interference. We tested the above framework and algorithms based on 10000 simulated and 5000 practical data points. The identification accuracy of the four kinds of noise was 97%~99%, and the proportion of high-quality (error < 5%) data was increased by approximately 20%~40%. The results show that artificial intelligence technology can quickly and effectively de-noise the large-scale IP data, so as to improve the quality of electrical information for deep mineral and petroleum exploration.

Introduction

Induced polarization (IP) is an effective geophysical technique for characterizing the complex resistivity structure of the shallow crust (down to -2 km) by injecting high-power current and acquiring array-induced polarization signals. Recently, many research institutions and companies have developed instruments with 2D distributed sensors, such as the Newmont distributed IP data acquisition system (NEWDAS), Quantec's 3D system, the IRIS instrument FullWaver, the Distributed Spread Spectrum IP (SSIP) system and so on (*Goldie*, 2007; *Alfouzan et al.* 2020), for use in the field. The data acquisition and storage capabilities of IP exploration have achieved revolutionary breakthroughs.

However, obtaining high-quality IP signals is still challenging in practical field surveys due to electromagnetic (EM) interference caused by natural and artificial sources. Four kinds of interference occur frequently in IP data, including trend drift interference caused by telluric currents and offset drift inside an instrument, strong discontinuous burst interference caused by machinery and equipment in mines, outliers caused by peak impulse interference, and Gaussian random noise caused by environmental disturbance (*Olsson et al*. 2016; *Barfod*, 2021). These interference types may occur at any time during an IP observation and distort the IP time series. Due to low efficiency and accuracy of the manual selection method, it is not suitable for massive-scale data. Scholars have proposed many de-noising methods to improve the IP data quality. In large-scale detection, the measured data include superposed noise interference types and pure IP signals. We need to combine various algorithms to address complex multisource noise interference. Additionally, every signal processing method inevitably loses useful signals, so it is very important to identify the type of noise interference in advance to choose the appropriate approach. The intelligent identification of noise interference in IP results is still rare in the existing literature.

The Framework and Algorithms

An intelligent de-noising framework (Figure 1) is developed.

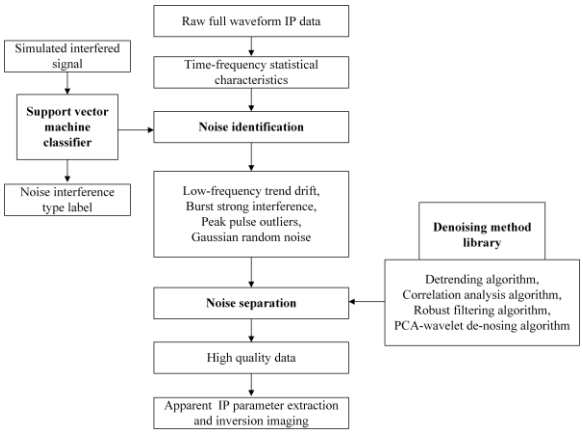

Figure 1: Intelligent anti-interference signal processing flow for a full-waveform IP time series.

We address four common noise types in multi-period full waveform IP data, including low-frequency trend drift, strong burst interference, peak impulse outliers and Gaussian random noise. We generated an IP signal and noise interference database, trained a support vector machine

10.1190/segam2021-3589725.1

model to classify and identify noise, and finally integrated four targeted signal processing technologies including a multi-period de-trending algorithm, correlation analysis algorithm, robust smoothing filtering algorithm and principal component analysis (PCA)-wavelet analysis de-noising algorithm, into a de-noising library to comprehensively address complex multisource interference.

Sample Database

In IP prospecting, the most commonly used emission current waveforms are periodic waves (such as rectangular waves, bipolar waves, or pseudorandom waves). According to the Fourier transform theory, any periodic current can be decomposed into the summation of sine and cosine series. In this paper, we mainly analyze the current of the fifth-order pseudo-random sequence. For a medium with an IP effect, the measured resistivity is a complex number that varies with the frequency and can be represented by the Cole-Cole model (*Pelton et al.* 1978). The IP voltage signal was calculated by applying the amplitude shift and phase shift of the complex resistivity to the excited current signal. We can generate the induced polarization time series using the above formula based on the given Cole-Cole model and the emission current parameters.

Additionally, four kinds of common noise interference in IP exploration were generated by mathematical simulation. The low-frequency trend drift interference was simulated by the superposition of the sine function, polynomial function, linear function and exponential function. Short-time strong burst interference was simulated by a Gaussian signal with an overall amplitude larger than that of a normal signal and duration less than that of the given time series. Peak impulse outlier interference was simulated by discretely processing Gaussian signals with amplitudes much higher than that of a normal signal. Background random noise was simulated by a Gaussian signal with an overall amplitude lower than that of a normal signal and duration the same as that of the given time series. Through random combinations of the 4 kinds of noise interference, 12 kinds of mixed noise were generated, and a total of 16 kinds of noise interference were simulated and added to the IP signal.

Support Vector Machine

The SVM have been widely applied in geophysical fields. The basic SVM model is a classic linear binary classification model. For data points with classification labels (1 or -1), the SVM algorithm is to find a boundary curve that keeps the two types of points as separate as possible, namely, that maximizes the sum of the distances from all the data points to the boundary curve. Through derivation, the SVM is finally transformed into an optimization problem; a value is found that minimizes the objective function. In practice, nonlinear multi-classification problems are common, and they can be solved by training multiple dichotomous models and mapping the samples that cannot be linearly segmented to a higher dimensional space. The formula for nonlinear SVM regression replaces the inner product of the predictors

In theory, an SVM model can be trained to classify and identify noise by taking the contaminated IP signal as the input and the corresponding noise interference type as the output. Due to the high dimensionality of IP time series, model training requires a large network structure and yields high computing costs. To improve the training efficiency, we use the second order statistical features of the original time series as inputs instead of directly using the complete time series as the input. We segmented the contaminated IP signal by period and extracted eight statistical characteristics for each data segment, including five time-domain parameters (the mean value, standard deviation, current-voltage correlation, outlier ratio and roughness) and three frequency-domain parameters (the fundamental frequency amplitude, fundamental frequency phase and major frequency energy ratio).

Integrated De-Noising Algorithms

Because noise interference in practical exploration is complex and variable, a single method cannot fully suppress all types of noise. Four signal processing techniques are improved and integrated into a de-noising method library.

First, we developed a de-trending method based on the periodicity of the IP signal. For multi-period time series without trend drift, the sampling points at the same position in different periods should be approximately equal. Therefore, we sampled various data points at the same position in different periods, interpolated them to obtain the fitting trends for the whole time period, and stacked all the fitting trends to approximate the real trend drift.

Next, we used a correlation analysis algorithm to remove strong burst interference. In IP exploration, there is a strong correlation among voltage data from various periods, and burst noise will dramatically reduce this correlation. We divided the original voltage data into several segments. For each data segment, we calculated the maximum correlation coefficient between it and other data segments. When the maximum correlation coefficient was less than 0.5, the data segment was considered to contain serious interference and was replaced by stacking other data segments.

Subsequently, we used robust smooth filtering to eliminate outliers caused by peak impulse interference. A moving-average filter was used to smooth data by replacing each data point with the average of the neighboring data points defined within a given window. For data in a smooth window, the robust mean is calculated by iterative reweighting algorithm according to the maximum likelihood criterion (*Huber*, 1964). The robust statistical method can automatically

10.1190/segam2021-3589725.1

reduce the weights of outliers and smooth time series according to the distribution of the original data.

Finally, we used the principal component analysis-wavelet analysis method to suppress Gaussian random noise. A multi-period signal can be reorganized into a matrix with one period for each column. The signal components are mainly concentrated in the principal component of the matrix, and the Gaussian noise is uniformly distributed in all components. Therefore, the signal component can be extracted by singular value decomposition and reconstruction using large eigenvalues. Wavelet analysis directly decomposes the original signal into multiple scales, with the large-scale component representing the signal and the small-scale component representing the noise. Therefore, the signal component can also be extracted by wavelet decomposition and reconstructed using the large-scale component.

Testing of Massive-Scale Simulated Data

We used a simulated massive-scale dataset to train the SVM model and used another dataset to test the accuracy of the model in noise identification and separation. We simulated 10,000 sets of pure IP signals. The emission current was a five-frequency pseudorandom wave, the sampling rate was 64 Hz, the period was 16 seconds, and a total of 20 periods of time series were acquired. We randomly generated four kinds of noise interferences and combined them. A total of 160,000 contaminated signal samples were obtained by adding various noise types to the pure IP signals. Figure 2 shows IP waveforms with different noise interference combinations. Then, we used the statistical characteristics of the noisy signals as the inputs and the noise type as the output to train the SVM model. The time for sample generation was approximately 1 hour, and the time for model training was approximately half an hour.

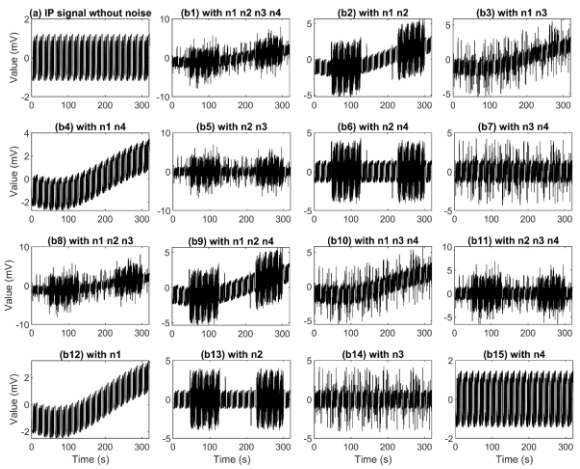

Figure 2: IP waveform with different noise interference combinations. n1: low-frequency trend interference; n2: strong burst interference; n3: peak impulse interference; and n4, background random noise interference.

We again randomly generated 10,000 sets of noisy signals for testing. The above trained SVM model was used to predict the noise type and call the corresponding signal processing techniques in the de-noising method library to eliminate noise interference. The recognition accuracies for the four types of noise interference were 99.19%, 99.89%, 99.63%, and 97.37%. Figure 3 shows the processing flow for IP data with various kinds of noise interference. The signal in Figure 3(a1) contains only Gaussian random noise, and only the PCA-wavelet method is used for noise reduction. The signal in Figure 3(b1) contains low-frequency trend drift and strong burst interference, which are processed with the de-trending algorithm and correlation analysis algorithm, respectively. The signal in Figure 3(c1) contains low-frequency trend drift, strong burst interference and peak impulsive interference, which are processed by the de-trending algorithm, correlation analysis algorithm and robust smoothing filtering algorithm, respectively. The signal in Figure 3(d1) contains all four interference types and is processed by the de-trending algorithm, correlation analysis algorithm, robust filtering algorithm and PCA-wavelet method. Figure 3 only shows the signal processing flows for the four kinds of noise combinations. The other 12 kinds of noise combinations yielded similar results. We also calculated the mean square relative error between the 10000 contaminated signals and real signals. Before noise reduction, the mean square relative error between the noisy signal and the real signal was approximately 170%. After de-noising, the mean square relative error between the signal and the real signal was reduced to approximately 8%.

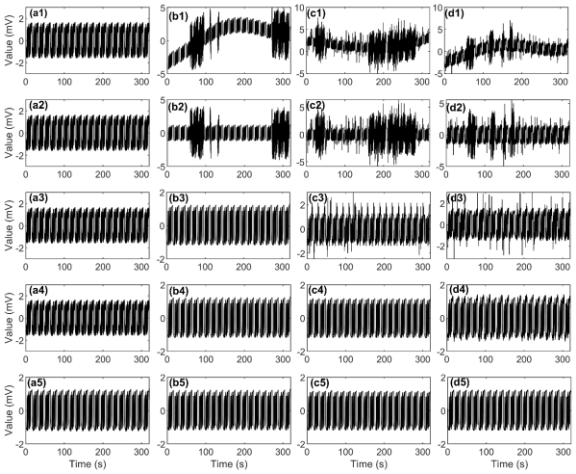

Figure 3: Processing of IP data with various noise interference combinations. (a1)~(d1): the contaminated IP

10.1190/segam2021-3589725.1

data; (a2)~(d2): IP data after de-trending; (a3)~(d3): IP data after correlation analysis processing; (a4)~(d4): IP data after robust smooth filtering; (a5)~(d5): IP data after PCA-wavelet de-noising.

Processing of Practical Data

The intelligent anti-interference processing framework and algorithm were applied to a practical dataset acquired in the ZXK lead-zinc ore concentration area in Tibet, China. To characterize the electrical structure of the shallow surface, an array-induced polarization exploration system that included 50 survey lines with 100 survey points per line was arranged using a gradient array. The survey line spacing was 40 m, and the survey point spacing was 20 m. The potential electrode spacing MN was 20 m. The current electrode spacing AB was 5000 m. For the injected current and acquired IP signal, the sampling frequency was 64 Hz, the period of one waveform was 256 s, and the measurement time was over an hour. Almost all the original full-waveform IP data were threatened by electromagnetic interference caused by artificial and natural sources. We first extracted the statistical characteristics of each time series, used the above trained SVM model to predict the noise interference type, and then called the corresponding signal processing method to reduce the noise. Finally, stacking and a Fourier transform were applied to the de-noised signal, and the Cole-Cole model parameters were estimated. After reprocessing each survey point, the false anomalies in the maps were removed.

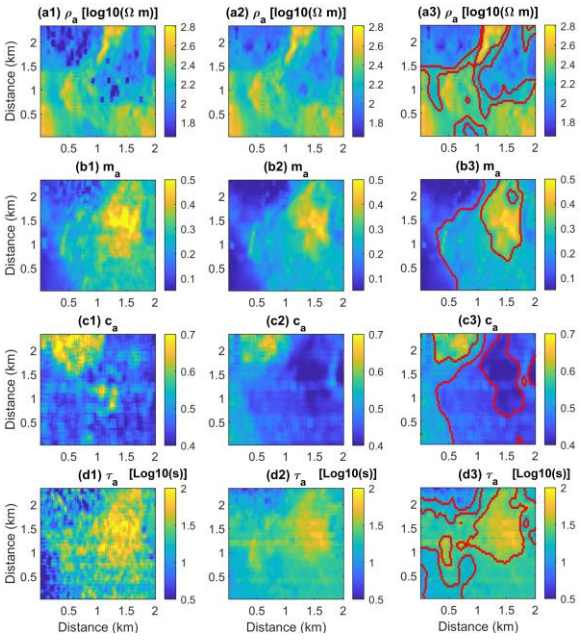

Figure 4: The plan maps of apparent Cole-Cole model parameters in the whole survey area. (a1), (b1), (c1) and (d1) are the plan maps of apparent resistivity, apparent chargeability, apparent frequency dependence and the apparent time constant without de-noising processing; (a2), (b2), (c2) and (d2) are the plan maps after de-noising; and (a3), (b3), (c3) and (d3) are the boundary demarcations of the plane maps.

The statistical errors of the above 5,000 survey points were also calculated. For apparent resistivity, through anti-interference processing, the proportion of high-quality data (error < 5%) increased from 80% to 95%. For apparent chargeability, the proportion of high-quality data (error <5%) increased from 48% to 92%. For the apparent frequency correlation coefficient, the proportion of high-quality data (error <5%) increased from 32% to 82%. For the apparent time constant, the proportion of high-quality data (error <5%) increased from 10% to 55%. It took approximately 8 hours to process these data by directly invoking all signal processing techniques without noise identification. By using the SVM to classify and identify time series, unnecessary processing was avoided, and the total time was approximately two hours.

Conclusions

A massive-scale geophysical IP signal processing method based on machine learning is developed in this paper. An SVM classifier and a de-noising method library are used for noise recognition and separation, respectively. The testing results based on 10,000 sets of simulated data show that the noise-recognition accuracy of the SVM can reach 97%~99%; de-noising based on multiple algorithm combinations can also reduce the signal error from 170% to 8%. After processing 5,000 sets of practical data from a mining area in Southwest China, the proportion of high-quality data increased by approximately 20%~40%, and a preliminary near-surface electrical structure was obtained. Both the simulated data and the practical measurement data verify the effectiveness of this method. The signal processing framework and algorithm proposed in this paper improve the accuracy, speed and automation level of massive-scale data processing.

Acknowledgments

This research was supported by the Science Foundation of China University of Petroleum, Beijing (2462020YJRC010, 2462020YXZZ005), and the Key Laboratory of Geophysical Electromagnetic Probing Technologies of Ministry of Natural Resources (No. KLGEPT201908). The code is available by contacting the corresponding author.

10.1190/segam2021-3589725.1

REFERENCES

Alfouzan, F. A., A. M. Alotaibi, L. H. Cox, and M. S. Zhdanov, 2020, Spectral induced polarization survey with distributed array system for mineral exploration: Case study in Saudi Arabia: Minerals, **10**, 769, doi: https://doi.org/10.3390/min10090769.

Goldie, M., 2007, A comparison between conventional and distributed acquisition induced polarization surveys for gold exploration in Nevada: The Leading Edge, **26**, 180–183, doi: https://doi.org/10.1190/1.2542448.

Huber, P. J., 1964, Robust estimation of a location parameter: The Annals of Mathematical Statistics, **35**, 73–101, doi: https://doi.org/10.1214/aoms/1177703732.

Olsson, P. I., G. Fiandaca, J. J. Larsen, T. Dahlin, and E. Auken, 2016, Doubling the spectrum of time-domain induced polarization by harmonic de-noising, drift correction, spike removal, tapered gating and data uncertainty estimation. Geophysical Journal International, **207**, 774–784, doi: https://doi.org/10.1093/gji/ggw260.

Pelton, W. H., S. H. Ward, P. G. Hallof, W. R. Hill, P. H. Nelson, 1978, Mineral discrimination and removal of inductive coupling with multifrequency IP: Geophysics **43**, 588–609, doi: https://doi.org/10.1190/1.1440839.