Machine Learning applied in Swell Noise classification

Almeida I.N.[1], Vargas M.[1], Neto C.S. [1], Nobre L.[1] , Silva B.[1], Bulcão A.[2], Dias B. P.[2] , Landau. L[1], Evsukoff A. [1]
(1. UFRJ/COPPE/LAMCE/Lab2M - Multidisciplinary Modeling laboratory)
(2. PETROBRAS - CENPES)

This paper was prepared for presentation during the 16th International Congress of the Brazilian Geophysical Society held in Rio de Janeiro, Brazil, 19-22 August 2019.

Abstract

Recorded seismic signals are inevitably contaminated by noise in field acquisition. Attenuating the high-amplitude noises, such as swell noise, is really a big challenge in the seismic data processing. High-amplitude swell noise is a common sort of noise in marine seismic survey. It usually affects a number of neighboring traces, and can be observed in seismic data as vertical stripes. One of the first tasks in seismic data processing is the swell noise attenuation. However this is not a trivial task, if not applied in adequate way can damage the signal and consequently affect posterior processing steps and therefore, compromise seismic interpretation. In this work, several approaches are proposed using machine learning to identify contaminated traces by this noise. Future research direction includes the attenuation of swell noise using these classified traces. The classification of the traces individually is given by the best AI model tested: it was tested seven different AI models with four different input variables. The architectures that had frequency-based features presented better results overall, especially the multi-layer perceptron.

Introduction

In the last decades, machine learning techniques have been widely used in several fields of application, ranging from image processing (Majumdar A., 2019), from audio signal process to speech recognition (Bourlard H.; Morgan N. ,1994). Among the models that stand out most today are the convolutional neural network (CNN) models (LeCun, Yann; et al., 1998), which perform better than other models of machine learning, especially in the field of computer vision. Because of these reasons, Oil and Gas industry have been investing resources in the development of these technologies in different areas, such as geoscience, in order to bring an improvement in standard traditional techniques and even offer solutions to yet not solved problems.

There are several examples of an application on literature. One of them is attenuate noise in seismic data with autoencoder network architecture (Mandelli, S. et al., 2019). There are other cases with neural networks MLP models to infer lithology, facies, porosity, and saturation of fluids (Zhang L.; Zhan C., 2017). Finally, (Waldeland A.U.; Solberg A.H.S.S., 2017) used CNN in convolution cascade to extract attributes and perform the classification of saline bodies.

Noise attenuation has a fundamental roll in seismic data processing and seismic interpretation. Typically, these fading methods are constructed with filter techniques, that are based on previous knowledge of the characteristics of the noise. In the case of Swell noise, it is known that it is a high-amplitude noise that normally contains frequencies from 2-10Hz (Elboth T.; Hermansen D., 2009), usually it affects a number of neighboring traces and can be observed in seismic data as vertical stripes.

One of the steps usually performed in seismic processing for noise attenuation is Fourier spectrum analysis. What is usually done in processing is the application of a tool based on filters. However, this approach is not effective when the noise and seismic signal occupy the same frequency range. Furthermore, the parameterization of the filtering is not intuitive, as they typically vary over time and can strongly alter the shape of the wave field in such a way as to impair subsequent steps. In this context, we believe that a trace-to-trace noise attenuation method is the best way to reduce the risk of information loss. The challenge of this new approach lies in identifying noisy traces within the seismic section. This challenge is the objective of this work.

In the literature, there are different methodologies to classify acoustic signals but mostly employ features extracted from the coefficients of the Fourier transform as inputs of the models. For example, the authors in (Piczak, K., 2015,) employ a convolutional network and use as input the spectrogram of the audio signals to train the model and perform the classification. The authors (Lim, H. et al., 2017) use the same input data methodology but the convolutional recurrent neural networks are used.

The process of identifying noisy signals can also be interpreted as an anomaly detection problem. Some examples are presented in (Prego, T. D. M.; et al., 2016) and (Oh, D.; et al., 2018). Where (Prego, T. D. M.; et al., 2016) compare a mechanism that measures the similarity between the Short Time Fourier Transform (STFT) coefficients of a reference signal and the signal of interest, while (Oh, D.; et al., 2018) employs a self-encoder network convolutional phase that uses as input the spectrogram in the time-frequency domain obtained by STFT. The anomalous data is detected by exceeding a threshold of error of the reconstruction of the signal made by the model.

Method

Data Description

The data that were used/collected for this project contains 1935 seismic traces and were split between 3 classes: strong swell noise presence, weak swell noise presence and clean trace. We will use the notation C1, C2 and C3 in order to simplify. Each of these classes contains 645 samples, those were collected from a seismic section which contains 1.328 shots, as presented in Figure 1. Each Shot contains 150 traces and each of them contains 1.793 samples with 4ms of sampling intervals, in total a recording time of 7.172s.

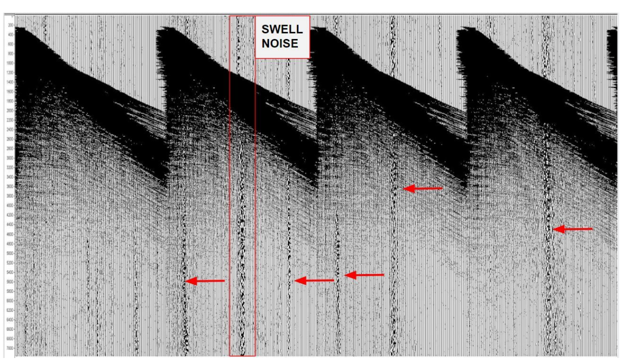

Figure 1- Swell Noise Noise examples in shots domain

Four group of features were created to train out Machine Learning models. The first group of features was created based on statistical attributes such as the mean, standard deviation, minimum, maximum, rms amplitude and absolute value. In seismic section some characteristics such as refraction, direct wave and reflections were founded/located in the region between 200ms and 3000ms. Those characteristics can affect these attributes calculation, therefore a mute was performed in the traces in the upper and lower regions, as shown in Figure 2.)

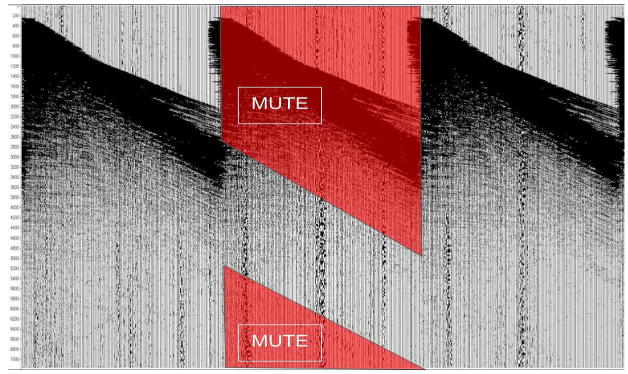

Figure 2 - Mute applied in Shots

The second group of features was created with the Fourier transform (FFT) method, that was applied/used in order to obtain the spectrogram in the amplitude-frequency domain. Once spectrogram created, a sum of the amplitudes of 2 Hz in 2Hz until reaching the range of 30 Hz was performed.

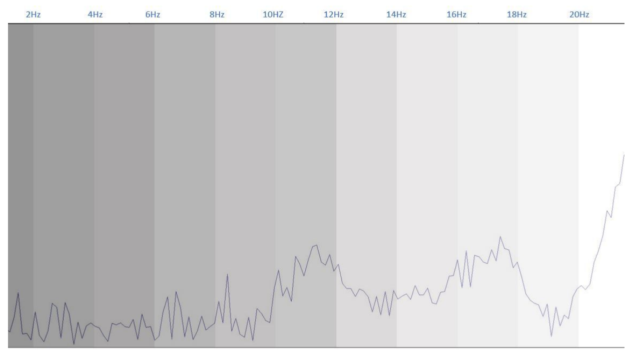

Figure 3 - Frequency range from 2 in 2Hz up to 30Hz

The third group of features is a combination of Inputs 1 and 2. Finally, the fourth feature is an spectrogram in the time-frequency domain of each of the seismic traces, according to the example shown in Figure 4 .

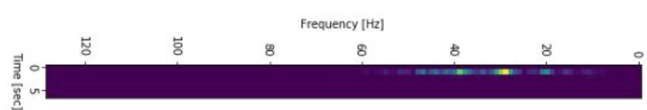

Figure 4 - Seismic trace Spectrogram

We will call each of these features as: Input1, Input2, Input3 and Input4 respectively, in order to simplify.

Machine Learning Models

In this work we have tested seven different architectures for classification of seismic traces from the database described in the previous section. These architectures are described as follow: 1) Multilayer perceptron (MLP), composed by 15 neurons in the intermediate layer and 3 neurons in the output layer, and hyperbolic tangent (tanh) and Softmax as activation functions for each layer. 2) Logistic regression. 3) SVM with linear kernel. 4) SVM with RBF kernel. 5) Decision tree. 6) Random Forest. 7) CNN model composed by 2 convolutional layers with 16 and 32 filters respectively, square kernel of dimension 3 and activation function tanh; 2 pooling layers of dimension 2; and 3 neurons in the output layer with Softmax activation function.

To perform the training of the models, the data set was divided into three subsets, named as training, validation and test, with proportions equal to 60%, 20% and 20%, respectively. The validation set was used to establish a premature training stop criterion in order to avoid overfitting. In addition, to analyze the generalization capacity of each of these architectures, the 5-folds cross-validation method was used. In both cases, as an evaluation metric the mean and the deviation of accuracy in each of the folds are used.

Results

Tables 1, 2, 3 and 4 show the numerical comparison results of the trained models for each proposed input. The best model in each case is highlighted in bold.

We can observe that the architectures that use the Input1 variable as input had lower performance than the others tested models, which leads us to infer that this input is not discriminant for this classification problem. It is verified through the confusion matrix that even opposing classes like C1 and C3 were mixed by the models.

From the tables 2, 3 and 4 we can infer that the models with inputs that make use of characteristics in the frequency domain have better performance. We highlight the results of the Random-Forest model using as input the variable Input4, which obtain 81.1% of accuracy in the test set. Analyzing the confusion matrix, it was observed that classes C1 and C3 were well discriminated by the model, however class C2 remains a problem, this is expected because the ambiguity in class definition.

Finally, we verify that the use of the statistical attributes (Input1) together with the frequency attributes (Input2), see table 3, did not lead to a considerable increase in the accuracy of the tested architectures.

Input1					
	Cross Val	Test Dataset			
	Acc Mean/ std	Acc	Conf Matrix		
			c1	c2	c3
MLP	59.2/ 0.016	54.7	38	30	61
			22	73	34
			18	10	101
Logístic Reg	50.0/ 0.011	50.9	38	24	67
			20	72	37
			34	8	87
SVM Linear	53.1/ 0.013	48.8	38	15	76
			24	58	47
			31	5	93
SVM RBF	50/ 0.011	47.5	48	17	64
			28	54	47
			42	5	82
Decision tree	55.8/ 0.020	57.1	67	23	39
			31	79	19
			34	20	75
Random Forest	**61.4/ 0.017**	**63.82**	70	26	33
			29	83	17
			32	3	94

Table 1 - Cross Validation and test sets results of Input1 for the following models: MLP, Reg. Logistic, Linear SVM, SVM RBF, Decision Tree and Random Forest.

Input2					
	Acc Mean/ std	Acc	Conf Matrix		
			c1	c2	c3
MLP	**77.5 0.144**	**77.2**	87	21	21
			19	100	10
			11	6	112
Logístic Reg	68.4 0.020	60.9	42	34	53
			17	85	27
			18	2	109
SVM Linear	73.1 0.019	69.5	74	18	37
			23	79	27
			12	1	116
SVM RBF	60.1 0.013	56.07	43	12	74
			20	71	38
			26	0	103
Decision tree	69.1 0.014	74.8 0.020	78	33	18
			30	88	11
			22	14	93
Random Forest	69.1 0.014	75.4	91	20	18
			25	92	12

			17	3	109

Table 2 - Cross Validation and test sets results of Input2 for the following models: MLP, Reg. Logistic, Linear SVM, SVM RBF, Decision Tree and Random Forest.

Input3					
	Cross Val	Test Dataset			
	Acc Mean/ std	Acc	Conf Matrix		
			c1	c2	c3
MLP	**78.0 0.021**	**78.2**	89	21	19
			21	99	9
			12	2	115
Logístic Reg	68.5 0.016	61.4	41	38	50
			15	86	28
			16	2	111
SVM Linear	73.0 0.021	68.7	71	20	38
			22	80	27
			13	1	115
SVM RBF	60.1 0.022	54.7	39	15	75
			18	72	39
			28	0	101
Decision tree	65.6 0.034	66.6	77	27	25
			35	80	14
			23	5	101
Random Forest	77.3 0.025	73.9	86	25	18
			24	92	13
			21	0	108

Table 3 - Cross Validation and test sets results of Input3 for the following models: MLP, Reg. Logistic, Linear SVM, SVM RBF, Decision Tree and Random Forest.

Input4					
	Cross Val	Test Dataset			
	Acc Mean/ std	Acc	Conf Matrix		
			c1	c2	c3
MLP	76.8 0.026	75.7	69	12	48
			19	106	4
			10	1	118
Logístic Reg	72.3 0.031	74.1	75	11	43
			14	108	7
			25	0	104
SVM Linear	75.2 0.008	75.1	80	6	43
			23	101	5
			19	0	110
SVM RBF	41.9 0.010	0.379845	41	0	88
			43	5	81
			28	0	101
Decision tree	70.0 0.020	72.8	79	28	22
			29	91	9
			11	6	112
Random Forest	**80.8 0.022**	**81.1**	105	12	12
			26	89	14
			8	1	120
CNN	78.3 0.029	79.5	82	9	38
			17	105	7
			8	0	121

Table 4 - Cross Validation and test sets results of Input4 for the following models: MLP, Reg. Logistic, Linear SVM, SVM RBF, Decision Tree and Random Forest.

Conclusions

The different architectures shown in the work presented allowed the evaluation of swell noise in seismic data. The obtained results were satisfactory although these models have presented some limitations to differentiate class C2. This problem may be related with ambiguity in the characterization of the weak Swell Noise, which adds errors in the process off seismic trace annotation. Despite the fact that this work is very preliminary, it serves as an initial reference for noise classification in the seismogram.

For future works, the method should be extended to test seismic trace classification with Swell Noise in further seismic section.

Acknowledgments

The authors would like to thank PETROBRAS and Lab2M, Department of Civil Engineering, Federal University of Rio de Janeiro for all help and support in the project.

References

Bourlard H.; Morgan N.; **Connectionist Speech Recognition: A Hybrid Approach.** 1994.

Elboth T.; Hermansen D.; **Attenuation of noise in marine seismic data**. 2009

LeCun, Yann; et al. **"Gradient-based learning applied to document recognition."**. Proc. IEEE, 1998.

Lim, H.; Park, J.; Han, Y. **Rare sound event detection using 1D convolutional recurrent neural networks**. Seoul National University, Seoul, Korea, September, 2017.

Majumdar A.; **Blind Denoising Autoencoder**. IEEE Transactions on Neural Networks and Learning Systems, January, 2019.

Mandelli, S.; Lipari, V.; Bestagini, P.; Tubaro, S. **Interpolation and Denoising of Seismic Data using Convolutional Neural Networks**. 2019.

Oh, D.; Yun, I. **Residual error based anomaly detection using auto-encoder in smd machine sound**. Korea, 24th of April, 2018.

Piczak, K. J. **Environmental sound classification with convolutional neural networks**. IEEE International Workshop On Machine Learning For Signal Processing, Boston, USA, September, 2015.

Prego, T. D. M.; de Lima, A. A.; Netto, S. L.; da Silva, E. A. **Audio anomaly detection on rotating machinery using image signal processing**. VII Latin American Symposium on Circuits and Systems (LASCAS) , Rio de Janeiro, February, 2016.

Waldeland A.U.; Solberg A.H.S.S.; **Salt Classification Using Deep Learning**. Oslo, Norway, June, 2017.

Zhang L.; Zhan C.; **Machine Learning in Rock facies Classification: An application of XGBoost**, International Geophysical Conference, Qingdao, China, April, 2017.