Seismic Discrimination between Earthquakes and Explosions Using CART Algorithm

Hyukwoo Lee, Kyunghyun Lee, and Kwanho You

***Abstract*—The distinction between earthquake and explosion signals is an essential issue in seismic signal analysis. We propose a machine learning model of decision tree (DT) applied to discriminate between earthquakes and explosions. The amplitudes of the P-wave and the S-wave are selected as feature vectors and built into the database. Classification and regression trees (CART) algorithm is used in our method, which is built through a greedy approach by the Gini impurity. The performance of the DT model using the CART algorithm is evaluated with the ROC curve. The results show the advantages of DT according to various evaluation indexes based on confusion matrix, and demonstrate that DT is efficient in seismic signal discrimination due to the nonparametric model characteristics of DT model.**

***Index Terms*—Decision tree, seismic wave, classification and regression tree, receiver operating characteristic.**

I. Introduction

As interest in earthquakes increases, more research continues to be conducted on seismic signal analysis around the world. In the field of seismic signal analysis, it is important to distinguish artificial seismic data such as quarry explosions and nuclear tests from natural earthquake data [1]. Many statistical machine learning-based methods have been proposed to make seismic signal classification more efficient and automated.

Dong [2] proposed a method to discriminate earthquakes with a nonlinear methodology using 10 ratios of different waves to distinguish velocity windows, and to verify the performance using random forests, support vector machines, and naive Bayes classifier. Orlic [3] proposed an optimal seismic discrimination method by implementing a genetic algorithm instead of using predefined features. Shang [4] applied the principal component analysis (PCA) method to 22 seismic parameters transforming the original data to a new set of lower-dimensional uncorrelated variables. Shang used the artificial neural network based on transformed variables to classify between earthquakes and quarry blasts. Zeheng [5] proposed a method of seismic wave discrimination using a combination of generative adversarial networks (GANs) and random forests.

In this paper, only the amplitudes of the P wave and the S wave of the seismic signal constitute a feature vector which significantly reduces the workload of database construction. We construct DT model based on CART algorithm with Gini impurity and introduce various parameters and receiver operating characteristic (ROC) curve to evaluate the identification performance. Decision tree is one of the most powerful method for classification and the regression for prediction that is composed with internal node, branch, and leaf node. Internal node, branch, and leaf node represent a test on an attribute, an outcome of the test, and a class label, respectively.

This paper is organized as follows. In Section II, the measurement of the P-wave and S-wave amplitude of the seismic signal is introduced and the protocol using DT is described. In Section III, we explain the decision point of the CART algorithm using Gini impurity. Section IV presents the performance of DT model using some simulations. Finally, Section V presents the conclusion.

II. Seismic Signal Discrimination

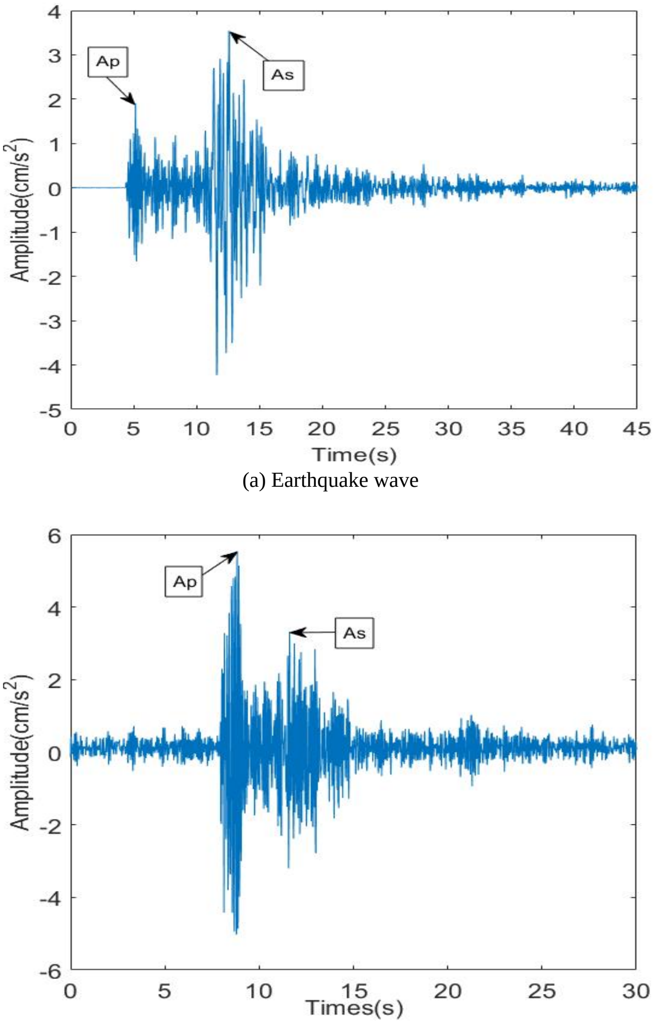
The Seismic waves are waves of energy that travel through the Earth’s layers. The seismic waves can be expressed as a graph of acceleration over time. In seismology, the peak values of P-wave and S-wave are defined as the amplitudes of P-wave (A_p) and S-wave (A_s), respectively. Earthquake wave signal and an explosion wave signal have different characteristics with the peak values of P-wave and S-wave. Fig. 1 shows an example of A_p and A_s of an earthquake wave signal and an explosion wave signal, respectively.

Many studies have shown that the ratio of P-wave to S-wave amplitude is an effective indicator of seismic and explosive distinction [6], [7]. If A_s is bigger than A_p, it is classified as earthquake wave. Likewise, if A_p is bigger than A_s, it is classified as explosion wave. There are two phases for seismic discrimination in our methods. In the offline-phase, it is the process of collecting various seismic data (A_p, A_s, label) and building a database. The configured training data can be represented as a set of $\{(x_i, y_i)\}$, $x_i \in R^2$, $y_i \in \{+1,-1\}$, $i \in \{1,2,\ldots,n\}$, where x_i is two dimensional vector with A_p and A_s values $(x_i = [A_s, A_p])$, y_i is the label with earthquake (+1) and explosion (-1), $(y_i = \pm 1)$, and n is the number of the data. Then these seismic data can be trained by a machine learning classifier (CART). In the online-phase, the new discrimination result of the measured seismic signal data is assigned to the trained machine learning classifier. Fig. 2 shows the algorithm block diagram of our proposed method.

Manuscript received August 3, 2018; revised December 25, 2018.

The authors are with the Department of Electrical and Computer Engineering, Sungkyunkwan University, Suwon, Korea (e-mail: lhw2065@skku.edu, naman2001@skku.edu, khyou@skku.edu).

doi: 10.18178/ijmmm.2019.7.1.428



(a) Earthquake wave

(b) Explosion wave

Fig. 1. Amplitudes of P-wave and S-wave.

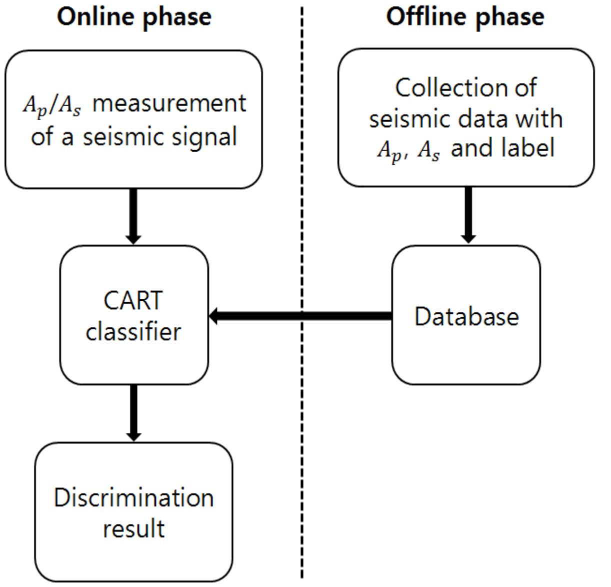

Online phase
Offline phase
A_p/A_s measurement of a seismic signal
Collection of seismic data with A_p, A_s and label
CART classifier
Database
Discrimination result

Fig. 2. Algorithm block diagram for seismic discrimination.

III. CART Algorithm Based Decision Tree Model

Decision tree is a basic non-parametric analytical method that performs classification and regression by charting decision rules into a tree structure [8]. It is a very powerful algorithm that represents the good performance on complex datasets. The structure of a decision tree consists of if-then decision rules which is composed of nodes and directed edges. In a classification problem, the internal nodes mean the features (or attributes) and leaf nodes mean the classification results (class labels). The formation process of the tree is divided into three parts (split rule, stop rule, and pruning).

CART algorithm is a binary separation decision algorithm based on statistical approach [9]. Instead of employing stopping rules, CART algorithm generates a sequence of subtrees by growing a large tree and retry again until only the root node is left. Then it uses cross-validation to estimate the misclassification cost of each subtree and chooses the one with the lowest estimated cost. In binary decision trees, there are three methods to measure the degree of data impurity (entropy, classification error, and genie index). This algorithm is determined by the value of the Gini impurity which is an index that measures how many heterogeneous elements are included in the set. This process is a step in split rule that finds the highest separation criterion of the node and separates the node if the stopping rule is not satisfied. Gini impurity is defined as equation (1).

$$Gini = \sum_{j=1}^{C} p_j(1-p_j) = 1-\sum_{j}^{C} p_j^2 = 1-\sum_{j=1}^{C}\left(\frac{n_j}{n}\right)^2 \qquad (1)$$

In a decision tree model, p_j is the proportion of class j-th class in the data set which has C classes in total, and n is the number of data in current nodes. The split points of CART algorithm can be fixed by the median of the ordered attributes value of each training data. CART algorithm selects the split that maximizes the decrease in impurity. Exceptionally, the branching is stopped when the condition of node satisfies the stop rule that the impurity degree of the node becomes zero or the size of the node reaches the limit. With the smallest Gini impurity value for each attribute, the split points become the most optimized value. Then the internal and leaf nodes of the decision tree can be obtained and finally the model is determined.

IV. Simulation and Performance Evaluation

To prove the performance of the proposed discrimination method, we use the data from the USA by the United States Geological Survey (USGS) and Incorporated Research Institutions for Seismology (IRIS) recorded in 2015. 40 seismic simulations, including 20 earthquakes and 20 explosions, were used in the DT model to train the machine learning classifier.

We use the box plot to verify the statistical difference between the earthquake and the explosion at the A_p and A_s values. Fig. 3 shows the box plot of amplitudes of earthquake and explosion, respectively. The horizontal lines inside the box stand for the median values. The top and bottom of the boxes are upper and lower quartiles [10]. The values at the end of the whiskers extended outside of the box are maximum and minimum value.

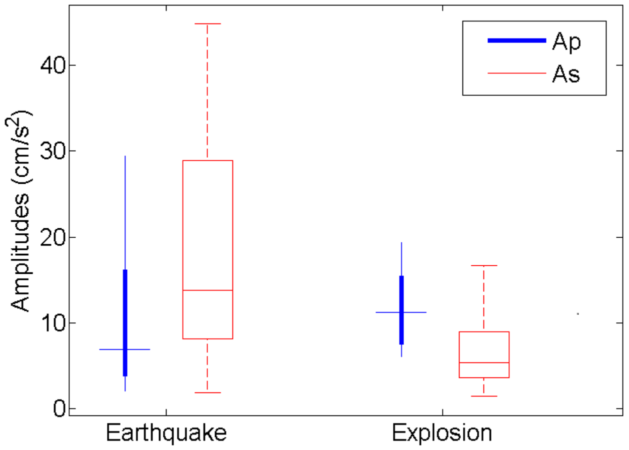
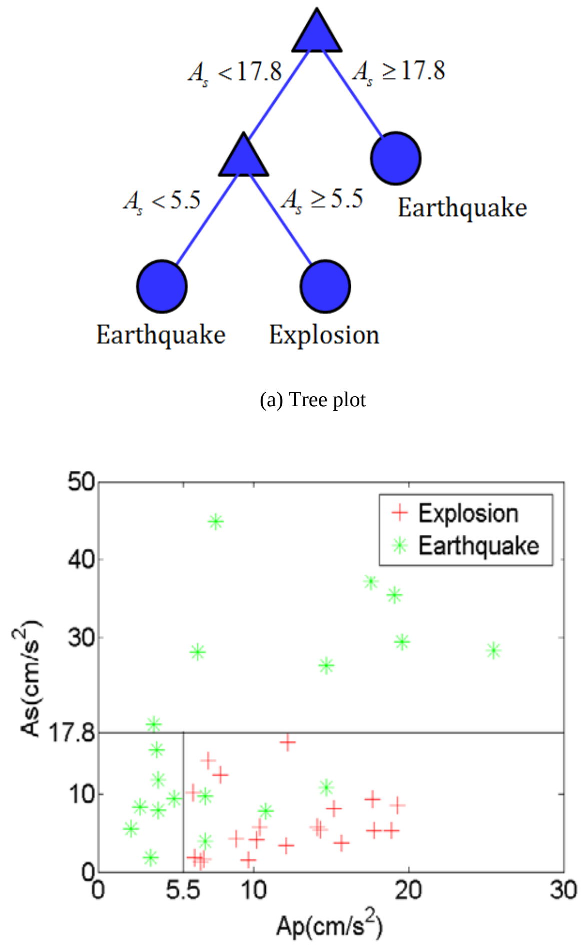


Fig. 3. Box plot of amplitudes of earthquake and explosion.

Fig. 4 shows the results of the CART algorithm of the non-parametric DT model. In the tree plot, the symbols of triangles are internal nodes that stand for features and dot symbols are leaf nodes that represent the classification results. Fig. 4(a) shows that the split point of the tree plot applies the cart algorithm. Fig. 4(b) shows the segregation of seismic data using the optimal split point results.



(a) Tree plot

(b) Feature space split by DT

Fig. 4. DT model trained by the data.

The receiver operating characteristic (ROC) curve which is a performance criterion for binary classifiers is suitable for evaluating trained models [11]. This method uses the sensitivity indicators and specificity to determine the accuracy of the test. In the ROC curve, *x* and *y*- axis mean true positive rate (TPR) and false positive rate (FPR), respectively. Specific data classified by the binary classifier can be represented by 4 kinds as shown in Table I.

TABLE I: CONFUSION MATRIX FOR BINARY CLASSIFICATION

Confusion matrix		True class	
		Positive	Negative
Hypothesized class	Positive	TP	FP
	Negative	FN	TN

The TPR defines how many correct positive results occur among all positive samples available during the test. From equation (2), TPR can be obtained.

$$TPR = \frac{TP}{TP + FN} \tag{2}$$

On the other hand, FPR defines how many incorrect positive results occur among all negative samples available during the test. From equation (3), FPR can be obtained.

$$FPR = \frac{FP}{FP + TN} \tag{3}$$

The ROC curve of our tested seismic CART classifier is shown in Fig. 5.

We use the area under the ROC curve (AUC) to verify the CART classifier [12]. This area represents the overall performance of the classifier. The value of AUC is 1 for an ideal classifier. Likewise, the performance of the classifier is low if the value of AUC approaches 0.5 that denotes the area of the reference line. The AUC value of our proposed CART algorithm is 0.9.

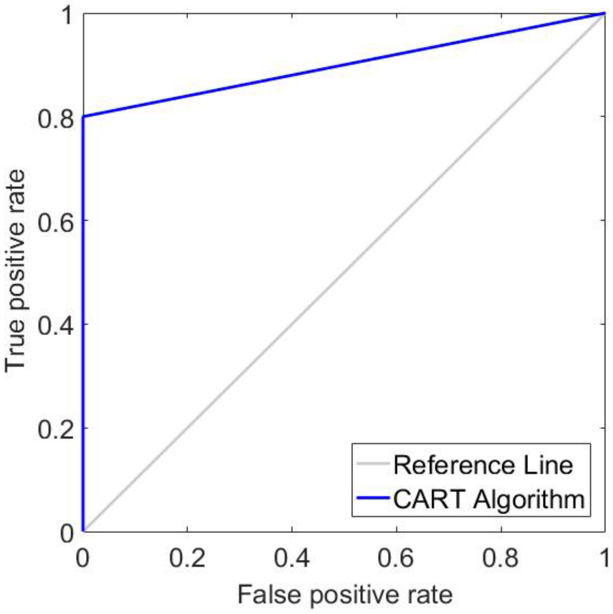


Fig. 5. ROC curve of CART algorithm.

To prove the performance of DT models, evaluation indicators of machine learning classifier performance based on confusion matrix are applied [13]. Precision, called as positive predictive value, is the fraction of relevant instances among the retrieved instances. The Precision can be obtained as $TP/(TP+FP)$, which is 1 as a result of this experiment. Meanwhile, Recall that is known as sensitivity is the fraction of relevant instances that have been retrieved over the total amount of relevant instances. The Recall can be obtained as $TP/(TP+FN)$, which is 0.8 as a result of this simulation. If compared with AUC, Precision and Recall are more sensitive to performance of the classifier than overall accuracy.

V. Conclusion

In this paper, we proposed a seismic discrimination method using CART algorithm. The seismic waveform has information on the amplitude values of the P-wave and the S-wave. We utilized the amplitude values to build a database of seismic discrimination. The DT model was designed by using the Gini impurity to find optimized split points. In order to verify the discrimination, fitness was calculated with ROC, AUC, Precision, and Recall.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NRF-2016R1A2B4011369).

References

[1] E. Yıldırım, A. Gülbağ, G. Horasan, and E. Doğan, “Discrimination of quarry blasts and earthquakes in the vicinity of Istanbul using soft computing techniques,” *Computers and Geosciences*, vol. 37, pp. 1209-1217, Sep. 2011.

[2] L. Dong, X. Li, and G. Xie, “Nonlinear methodologies for identifying seismic event and nuclear explosion using random forest, support vector machine, and naive Bayes classification,” *Abstract and Applied Analysis*, vol. 2014, pp. 1-8, Feb. 2014.

[3] N. Orlic and S. Loncaric, “Earthquake—explosion discrimination using genetic algorithm-based boosting approach,” *Computers and Geosciences*, vol. 36, pp. 179-185, Feb. 2010.

[4] X. Shang, A. Morales-Esteban, and G. Chen, “Nonlinear methodologies for identifying seismic event and nuclear explosion using random forest, support vector machine, and naive Bayes classification,” *Soil Dynamics and Earthquake Engineering*, vol. 99, pp. 142-149, Aug. 2017.

[5] L. Zefeng, M. Men-Andrin, H. Egill, Z. Zhongwen, and A. Jennifer, “Machine learning seismic wave discrimination: Application to earthquake early warning,” *Geophysical Research Letters*, vol. 45, pp. 4773-4779, May 2018.

[6] P. W. Pomeroy, W. J. Best, and T. V. McEvilly, “Test ban treaty verification with regional data,” *Bulletin of the Seismological Society of America*, vol. 72, pp. 89-129, Dec. 1982.

[7] D. R. Baumgardt and G. B. Young, “Regional seismic waveform discriminates and case-based event identification using regional arrays,” *Bulletin of the Seismological Society of America*, vol. 80, pp. 1874-1892, Dec. 1990.

[8] S. R. Safavian and D. Landgrebe, “A survey of decision tree classifier methodology,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 21, pp. 660-674, June 1991.

[9] W. Loh, “Classification and regression trees,” *Wiley Interdisciplinary Reviews-Data Mining and Knowledge Discovery*, vol. 1, pp. 14-23, Jan. 2011.

[10] K. Potter, H. Hagen, A. Kerren, and P. Dannenmann, “Methods for presenting statistical information: The box plot,” *Visualization of Large and Unstructured Data Sets*, vol. 4, pp. 97-106, 2006.

[11] C. E. Metz, “Basic principles of ROC analysis,” *Seminars in Nuclear Medicine*, vol. 8, pp. 283-298, Oct. 2006.

[12] J. Huang and C. X. Ling, “Using AUC and accuracy in evaluating learning algorithms,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, pp. 29-36, Jan. 2005.

[13] J. Davis and M. Goadrich, “The relationship between precision-recall and ROC curves,” in *Proc. of the 23rd International Conference on Machine Learning*, pp. 233-240. June 2006.

Hyukwoo Lee received his B.S. degree in electrical and computer engineering from Hankyung University in 2018. He is working as a research staff in applied optimization lab. at Sungkyunkwan University. His research interest fields are seismic wave discrimination with regression model and quadrotor tracking control using sliding-mode control.

Kyunghyun Lee received his B.S. degree in electrical engineering from Sungkyunkwan University in 2013. He is working as a research staff in Applied Optimization Lab. at Sungkyunkwan University. His current interest fields are intelligent control, geolocation problem with radio signal, measurement sensor development, estimation theory, and real-time nonlinear systems.

Kwanho You received his B.S. and M.S. degrees in electrical engineering from Korea Advanced Institute of Science and Technology (KAIST) in 1993 and 1996, and received his Ph.D. degree from the University of Minnesota in 2000, respectively. In 2001, he joined the School of Information and Communication Engineering at Sungkyunkwan University. His research interests are wireless communication, adaptive optimization methods in nonlinear process, and estimation theory.