Journal of Multidisciplinary Engineering Science and Technology (JMEST)
ISSN: 2458-9403
Vol. 7 Issue 11, November - 2020

Discrimination of Imbalanced Seismic Events between Earthquakes and Artificial Explosions using Support Vector Machine

Sangkyeum Kim, Kyunghyun Lee, Kwanho You*
Department of Electrical and Computer Engineering, Sungkyunkwan University, Korea
E-mail address: interpost94@skku.edu, naman2001@skku.edu
*Corresponding Author: khyou@skku.edu

***Abstract*—The discrimination between earthquake and artificial explosions is an important issue in seismology. Unlike an earthquake, the artificial explosion is caused by a quarry blast or a nuclear test. The surface or receiver response to earthquake and artificial explosion is different. However, the discrimination of earthquake from other impulsive explosion is not easy especially in seismically noisy environment. The discrimination accuracy is limited by the imbalanced data between earthquake and artificial explosion. This paper proposes a seismic discrimination method using a support vector machine (SVM) with the P/S amplitude ratio as a feature vector. Furthermore, to enhance the performance of discrimination using SVM, the adaptive synthetic (ADASYN) sampling algorithm is applied to reduce the error caused by imbalanced data. In the learning and test process, different kernel functions are applied on SVM to improve the performance of discrimination between earthquake and artificial explosion. The accuracy and outcome of the proposed method using ADASYN-based SVM are evaluated by the receiver operating characteristic curve and the area under the curve. The simulation results show that the proposed method leads to more precise analysis and classification for seismic event discrimination.**

Keywords—support vector machine; ADASYN; imbalanced data; artificial explosion; seismic discrimination

I. INTRODUCTION

Recently, many studies have been conducted to analyze more exactly the seismic signal. Especially, the discrimination between the earthquake and artificial explosion is an important topic in Geology. The earthquake can occur anywhere with little or no warning. Various seismological stations spaced at intervals of several kilometers are used to record and analyze all types of seismic signals around the region. The seismological stations store not only earthquake signals but also the artificial explosion signals such as quarry blasts or nuclear tests. However, seismic analysts must discriminate the seismic signals between earthquake and artificial explosion before the precision analysis is executed. [1] Such work is time-consuming and laborious. Therefore, a large number of statistical machine learning-based methods have been proposed in the seismic event discrimination to increase the discrimination accuracy. However, there is a problem of the data balance between the earthquake and the artificial explosion. In many cases, the natural earthquake happens more frequently than artificial explosion. [2]

To discriminate between earthquakes and artificial explosions, many novel approaches of research have been conducted using the machine learning method. Rabin [3] proposed a graph-based machine learning method using the diffusion maps that are recorded at several seismological stations. Nakano [4] used a convolution neural network (CNN) method to discriminate seismic signals from earthquakes and tectonic tremors. Nakano proposed seismic running spectra–CNN to improve the discrimination accuracy.

However, the artificial explosion dataset is relatively hard to obtain compared to the earthquake. It causes the imbalanced data problem which deals with uncertainty. The imbalanced dataset issue is caused by the unmatched training samples among the classes which are to be discriminated. The earthquake's dataset is typically larger than the artificial explosion's dataset. It means that the classes are divided into the majority class and the minority class, which represent the earthquake's dataset and the artificial explosion's dataset, respectively. In this paper, the imbalanced seismic data set is discriminated by the proposed method. To discriminate imbalance seismic data and reduce the seismic discrimination errors, we apply the ADASYN-based support vector machine (SVM) method. To compensate for the wide discrepancy among datasets, the ADAYSN method which is one of the over-sampling approaches is adapted to generate the synthetic data. The generated synthetic data is inserted to the minority class. The number of minority class becomes equal to the number of majority class.

This paper is organized as follows. Section II presents the feature between the earthquake and artificial explosion using the P/S amplitude ratio and the SVM based seismic discrimination block diagram. The ADAYSN-based SVM classifier is explained in Section III. Section IV discusses the discrimination performance between ordinary SVM classifiers and

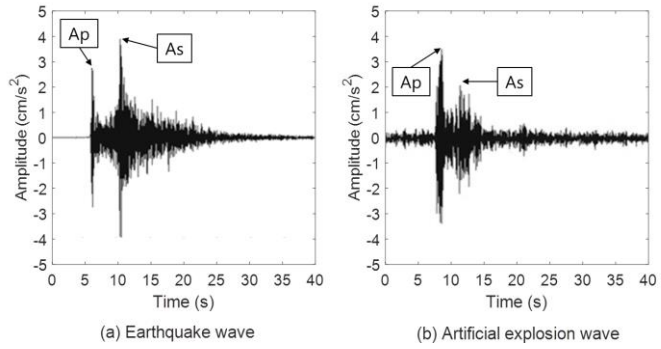

Fig. 1: *Amplitudes of P-wave and S-wave*

ADASYN-based SVM classifiers. Conclusion appears in Section V.

II. Seismic Signal Discrimination

The seismic wave signal is the signal of energy propagating through the earth's surface. The artificial explosion is similar to an earthquake, whereas it has different characteristics compared to the earthquake. It occurs under the unnatural environment such as explosions made through nuclear tests or quarry blasts. The earthquake and artificial explosion's signals have compositions of two bands, which are P-wave and S-wave with peak amplitudes of P-waves (A_P) and S-wave (A_s), respectively. Several studies have shown the P/S amplitude ratio is different between earthquake and artificial explosion. [5, 6] Generally, A_s is bigger than or equal to A_P for the case of earthquake. On the other hand, the artificial explosion has A_P bigger than A_s. Fig. 1 shows an example of the seismic signals for an earthquake and an artificial explosion with different A_P and A_s.

To discriminate between earthquake and artificial explosion using different amplitude characteristics, the discrimination process is carried out following the procedures in Fig. 2. To compose the SVM classifier, the seismic dataset which has information of A_P, A_s and labels (earthquake or artificial explosion) is used as a training dataset. The training dataset is expressed as $\{(x_i, y_i)\}, x_i \in \Re^2, y_i \in \{+1, -1\}, i \in \{1, 2, ..., n\}$, where x_i is a feature vector in two dimensions representing A_P and A_s.

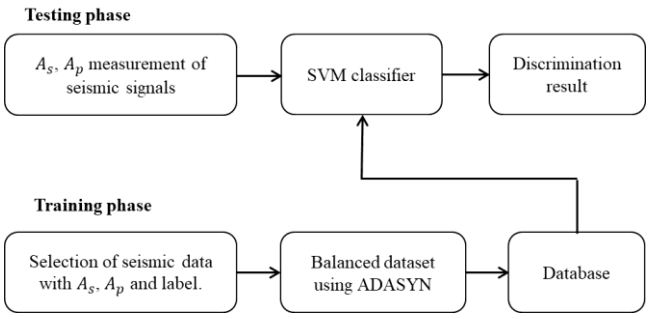

Fig. 2: *Block diagram for seismic discrimination algorithm.*

y_i is the label which means earthquake (+1) or artificial explosion (-1). n is the number of training data. The selected data is trained by SVM which is one of the machine learning based classification methods. SVM has two phases that are the training phase and the testing phase. In the training phase, the SVM classifier is optimized using the feature vectors which are A_P, A_s and labels. In the testing phase, the dataset including A_P, A_s is discriminated by the classifier. The outputs which are discriminated by SVM classifier are distinguished as the earthquake (+1) or artificial explosion (-1), respectively. Fig. 2 shows the block diagram of the proposed method.

III. ADASYN-based Support Vector Machine

The SVM is straightforward to apply for classification with its efficient performance. However, dividing the space with proper decision boundaries is critical to the SVM's performance. In order to solve the hyperplane problem, it is necessary to set up the decision rules. [7] The hyperplane of SVM is expressed as

$$w^T x_i + b = 0\,, \tag{1}$$

where w is the weight vector and b is the unregularized bias term. The decision function $f(x_i)$ is related to a hyperplane. The decision function is represented as

$$f(x_i) = \mathrm{sgn}(w^T x_i + b)\,. \tag{2}$$

For optimization, it is required to take maximum margin between two classes of data as shown in (3)

$$\begin{aligned} &\max_{w,b} \frac{2}{\|w\|}, \\ &\text{such that } \; y_i(w^T x_i + b) \geq 1. \end{aligned} \tag{3}$$

Equation (4) is the Lagrange function including the objective function and constraint equation.

$$L(w, b, \alpha_i) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i [y_i (w^T x_i + b) - 1], \tag{4}$$

where α_i is the Lagrange multiplier for support vectors. According to complementary slackness condition and stationarity of Karush-Kuhn-Tucker (KKT) condition, the necessary condition for the optimized decision rule is

$$\begin{aligned} &\alpha_i [y_i (w^T x_i + b) - 1] = 0, \\ &w = \sum_{i=1}^{n} \alpha_i y_i x_i. \end{aligned} \tag{5}$$

According to w derived by KKT condition, the SVM decision function can be represented as

$$f(x_i) = \mathrm{sgn}(w^T x_i + b) = \mathrm{sgn}\left\{ \sum_{i=1}^{n} \alpha_i y_i (x_i \cdot x) + b \right\} \tag{6}$$

The Lagrange multiplier α_i and the bias term b are the solutions of the dual optimization problem. Nonlinear transformation of the input vector $\varphi(x): R^2 \to F$ is applied to solve the non-separable classification problem. To obtain the optimal hyperplane, the decision function of SVM is rewritten as the nonlinear transformation with the kernel function $K(x_i, x_j) = \varphi(x_i) \cdot \varphi(x_j)$. In this paper, three different kernel functions such as linear kernel $(K(x_i, x_j) = x_i^T x_j)$, polynomial kernel $(K(\mathrm{x}_i, x_j) = (x_i^T x_j + c)^n, c > 0)$, and Gaussian basis function (RBF) kernel $(K(\mathrm{x}_i, x_j) = \exp(-\| x_i - x_j \|^2 / (2\sigma^2)), \sigma \neq 0)$ were applied.

Despite systematic rules and theoretical backgrounds of SVM, SVM shows a sub-optimal quality similar to the overfitting problem when the dataset is imbalanced. To keep a balance between minority and majority class, many studies have been conducted to fix imbalanced data problem. [8, 9] In various studies, the imbalanced data problem is dealt for the realization of the classification process. The data pretreatment before the classification shows high accuracy as compared to modification of the classification process. One supplementary method is to adjust the balance of the number of training dataset. The ADASYN algorithm is a method of oversampling for the minority dataset to achieve balance with the majority dataset. [10] ADASYN algorithm summarizes as follows.

ADASYN Algorithm

Input: Training dataset.
Output: Synthetic data from minority class.
Procedure:

1. From the minority class dataset, calculate the amount of the synthetic dataset: $S = (A_l - A_s) \cdot \nu$, where A_l and A_s represent the number of data of majority class and minority class, respectively. ν is used to specify the balance of synthetic data needed with $\nu \in [0,1]$
2. For each x_i which represents samples on minority class, calculate the ratio $\rho_i = \eta_i / K$ according to *K*-nearest neighbors. η_i represents the number of neighbors on the majority class.
3. Calculate the normalized ρ_i in accordance with $\hat{\rho}_i = \rho_i / \sum r_i$.
4. Compute the number of synthetic data g_i that needed to be generated for each x_i belong to $g_i = \psi \cdot \hat{\rho}_i$, where ψ is the total number of synthetic data generated in minority class.
5. Do **Loop**
   For $i < g_i$,
   a. Randomly select the data x_{zi} which is the *K*-nearest neighbors of x_i .
   b. Generate the synthetic data as following:
   $r_i = x_i + rand(0,1) \cdot (x_{zi} - x_i)$.
   End **Loop**

The synthetic data generated by the ADASYN is added to the minority class to prevent overfitting. Therefore, the minority class keeps a balance with the majority class.

IV. Simulation and Perfromance Evaluation

The simulation and experiment have 3 procedures to demonstrate the classification performance of the proposed method. As a first step, 79 earthquakes (majority class) and 17 artificial explosions (minority class) were selected from the United States Geological Survey (USGS) and the Incorporated Research Institutions for Seismology (IRIS) in 2015 to 2017. As a second step, the minority data was generated by ADASYN. The number of artificial explosion data became equal to earthquake data. The SVM classifier was trained using the balanced dataset. By applying different kernel functions, SVM classifiers were conducted using the same training data. As a third step, 40 additional data was discriminated by the trained classifier. Fig. 3 shows the seismic classification performance by using the ADASYN-based SVM classifier. As shown in Fig. 3, 4 different kernel functions which are linear kernel, 4-th order polynomial kernel, and RBF kernel $(\sigma = 0.2,1)$ are adopted as SVM classifier to obtain better performance. The *x* and *y*-axis are the peak values of P and S-wave amplitudes, respectively. The symbol of the asterisk represents the earthquake data, and the cross symbol means the artificial explosion data, respectively. Additionally, the support vectors that distinguish the optimized hyperplane are marked with circles. In the performance test, we obtained the optimal value $(\sigma = 0.2)$ through several trial and errors to avoid the overfitting problem. As shown in Fig. 4, the ADASYN based-SVM classifier discrimination performance is confirmed with RBF kernel $(\sigma = 0.2)$.

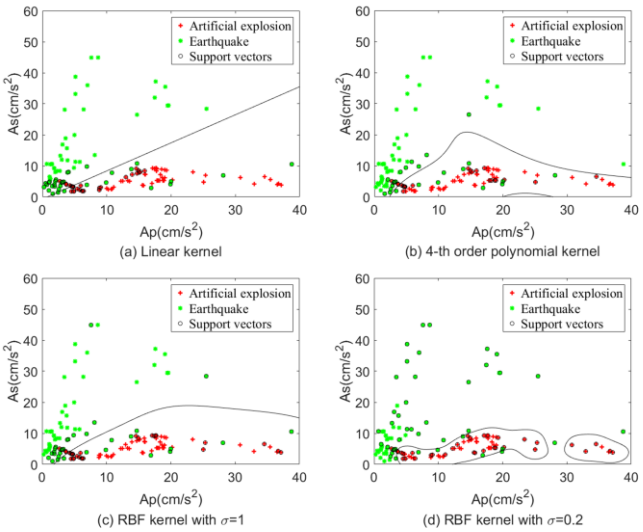

Fig 3. *ADASYN based-SVM classifier with different kernel functions.*

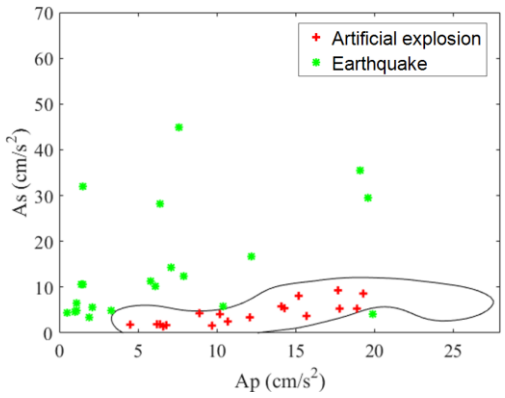

Fig 4. *Discrimination result with ADASYN-based SVM classifier.*

To prove the performance of the ADASYN-based SVM classifier, 40 additional seismic data recorded by USGS and IRIS was discriminated through the trained ADASYN-based SVM classifier with RBF kernel ($\sigma = 0.2$). Fig. 4 shows the discrimination result of the 40 additional seismic data using our proposed method.

The receiver operating characteristic (ROC) curve is a model used in measurement for binary classification. The ROC curve was used to verify the discrimination performance of the SVM classifier. The ROC curve has two axes of true positive rate (TPR) and false positive rate (FPR). [11] TPR is the proportion of what is negative is wrong in predicting the model as positive. Conversely, FPR which is the proportion of what is negative is wrong in predicting the model as positive. TPR and FPR are expressed as follows.

$$TPR = \frac{TP}{TP + FN}, \; FPR = \frac{FP}{FP + TN}. \tag{7}$$

The process of plotting the ROC curve is to calculate the TPR and FPR as the threshold of the classifier changes. The TPR is equal to the ratio of the true positive (TP) correctly classified by the system to the sum of the TP and the false negative (FN) incorrectly classified by the system. Likewise, the FPR is the ratio of the false positive (FP) to the sum of the FP and the true negative (TN). Fig. 5 shows the ROC curve of ADASYN-based SVM classifier.

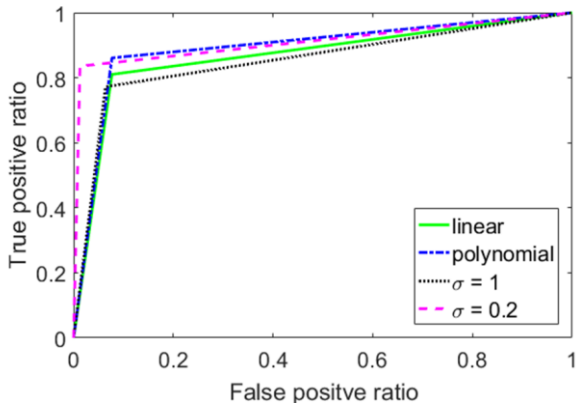

Fig 5. *Discrimination result with ADASYN-based SVM classifier.*

TABLE I. AUC VALUES OF ADASYN-BASED SVM CLASSIFIERS AND ORDINARY SVM CLASSIFIER

Kernel function	ADASYN-based SVM classifier	Ordinary SVM classifier
Linear	0.8856	0.8779
4-th polynomial	0.9111	0.8779
RBF ($\sigma = 1$)	0.9047	0.8779
RBF ($\sigma = 0.2$)	0.9555	0.9263

The area under the curve (AUC) is also used in measurement for classification. [12, 13] The AUC values of ADASYN based-SVM classifiers and ordinary SVM classifiers using linear kernel, 4-th order polynomial kernel, and RBF kernel ($\sigma = 1, 0.2$) are shown in Table 1.

V. CONCLUSTION

In this paper, the imbalanced seismic data discrimination method was proposed using the P/S amplitude ratio as the feature vectors. To compose the feature vectors, the seismic dataset from 2015 to 2017 was collected from USGS and IRIS. The ADASYN algorithm was used to keep a balance between the earthquake and artificial explosion which act as a majority class and a minority class, respectively. Utilizing the ADASYN algorithm, the adaptively synthesized dataset was generated in the minority class. Through the balanced dataset, the SVM model was designed by using different kernel functions to obtain an optimized hyperplane. Using the trained ADASYN-based SVM classifier, the discrimination between the earthquake and artificial explosion was executed. In order to verify the performance of the proposed method, the fitness evaluation was confirmed using the ROC and AUC.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NRF-2019R1A2C1002343, NRF-2020R1I1A1A01061632).

REFERENCES

[1] C. Seplaki, N. Goldman, M. Weinstein, and Y. Lin, “Before and after the 1999 Chi-Chi earthquake: traumatic events and depressive symptoms in an older population,” Soc. Sci. Med., vol. 62, pp. 3121-3132, June 2006.

[2] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, “Learning from class-imbalanced data: review of methods and applications,” Expert Syst. Appl., vol. 73, pp. 220-239, May 2017.

[3] N. Rabin, Y. Bregman, O. Lindenbaum, Y. Ben-Horin, and A. Averbuch, “Earthquake-explosion discrimination using diffusion maps,” Int. J. Geophys., pp. 1484–1492, December 2016.

[4] M. Nakano, D. Sugiyama, T. Hori, T. Kuwatani, and S. Tsuboi, “Discrimination of seismic signals from earthquakes and tectonic tremor by applying a convolutional neural network to running spectral images,” Seismol. Res. Lett., vol. 90, pp. 530-538, January 2019.

[5] S. Yilmaz, Y. Bayrak, and H. Cinar, “Discrimination of earthquakes and quarry blasts in the eastern Black Sea region of Turkey,” J. Seismol., vol. 17, pp. 721-734, November 2013.

[6] M. Mariani, H. Gonzalez-Huizar, M. Bhuiyan, and K. Tweneboah, “Using dynamic Fourier analysis to discriminate discrimination between seismic signals from natural earthquakes and mining explosions,” AIMS Geosci., vol. 3, pp. 438-449, August 2017.

[7] A. karatzoglou, D. Meyer, and K. Hornik, “Support vector machine in R,” J. Stat. Softw., vol. 15, pp. 1-24, April 2006.

[8] R. Longadge, S. S. Dongre, and L. Malik, “Class imbalance problem in data mining: review,” Int. J. Comput. Sci. Netw., vol. 2, pp. 83-87, February 2013.

[9] S. Kim, H. Kim, and Y. Namkoong, “Ordinal classification of imbalanced data with application in emergency and disaster information service,” IEEE Intell. Syst., vol. 31, March 2016.

[10] H. He and E. Garcia, “Learning from imbalanced data,” IEEE Trans. Knowl. Data Eng., vol. 21, pp. 1263-1284, June 2009.

[11] J. Hanley and B. McNeil, “The meaning and use of the area under a receiver operating characteristic (ROC) curve,” Radiology, vol. 143, pp. 29-36, April 1982.

[12] T. Saito and M. Rehmsmeier, “The precision-recall plot is more impormative than the ROC plot when evaluation binary classifiers on imbalanced dataset,” PLoS ONE, vol. 10, e0118432, March 2015.

[13] J. Huang and C. Ling, “Using AUC and accuracy in evaluating learning algorithms,” *IEEE Trans. Knowl. Data Eng.,* vol. 17, pp. 299-310, January 2015.