# Machine Learning Based Classification Algorithm for Seismic Blasting Recognition Model Research

**Rong Chen[1], Mingyuan Liu[2*], Yingchao Niu[1], Zhuolin Yu[1]**
**1. Lanzhou Vocational Technical College, Lanzhou 730070, China.**
**2. Liaoning Technical University, Fuxin 123032, China.**

***Abstract:*** Seismic signal identification is an important part of seismology and earthquake observation, but urban engineering and unnatural seismic events interfere with seismic recording and management, requiring the use of data to build relevant and reliable models for identification and exclusion. Firstly, the seismic wave signals are mapped according to the data, the signal characteristics are observed, the seismic signals are decomposed using CEEMDAN (Complete Ensemble Empirical Mode Decomposition with Adaptive Noise), the sample entropy is solved for the first 7 IMFs, and the sample entropy is utilized to construct the feature vectors as the Using the sample entropy to construct feature vectors as training features, KNN (K-NearestNeighbor), and SVM (Support Vector Machine) models were constructed to solve and evaluate the model effects, Recall and F1score, with the highest score reaching 100%. It plays a crucial role in the development of earthquake early warning technology as well as earthquake prevention technology, and has great reference value for future related research.

***Keywords:*** CEEMDAN; KNN; SVM; Seismic Blast Recognition

# 1. Introduction

Magnitude prediction is one of the important goals of earthquake prediction, relying on feature mining of historical events and seismic wave energy estimation, which helps to develop earthquake emergency response programs and reduce losses (Patterson, B. et.al.). With the development of computer technology and artificial intelligence, the application of artificial intelligence seismology was born to solve the conventional seismological problems with machine learning and neural network models, which are deeply involved in the construction of source attribute identification model and magnitude prediction (Abdalzaher, M. S. et.al.).

The purpose of the seismic source attribute recognition model is to determine whether the seismic event is natural or unnatural based on the seismic wave data, which is a classification task, so the model is built using machine learning classification algorithms such as Support Vector Machine (SVM), K Nearest Neighbors (KNN), etc., and the sample entropy is extracted as the feature input through the decomposition sequence of the data to construct the feature vector for the prediction of the solution.

# 2. Model Establishment

## 2.1 KNN modeling

KNN model is a classification or regression model based on distance metric and voting method, its core idea is to find out the K closest samples based on the distance between a new sample and a known sample, and then predict the class or value of the new sample based on the class or value of these samples (Rincon-Yanez, D. et.al.).

KNN is given a training set $T=\{(x_1,y_1),(x_2,y_2),\ldots,(x_m y_m)\}$ for prediction of classification, where $x_i=(x_1^1,x_1^2,\ldots,x_1^n)$ is the ith training sample, $i=1,2,\ldots,My_i\in\{c_1,c_2,\ldots,c_R\}$ is the category label corresponding to $X_I$, and the category to which the instance $X_J$ is to be predicted belongs to is $Y_J$, according to the selected $K$ value, choose the

Euclidean distance measure, and the Euclidean distance formula is shown in the following, iteratively traverse all the sample points in the training set to find the *K* nearest neighbors $x_q$, $q = 1,2,3, \ldots, K$, and finally decide the category $y_j$ to which instance $x_j$ belongs based on the majority voting method.

$$dist_{ed}\ (x_i, x_j) = \sqrt{\sum_{u=1}^{n} |x_i^u - x_j^u|^2} \quad (1)$$

Where $x_i$,$x_j$ are the given samples, where $i, j = 1,2,3, \ldots m$ denotes the number of samples; *n* denotes the number of features.

KNN algorithm is the simplest and crudest is to calculate the prediction point and all the points distance, and then saved and sorted, selected the category of the previous *K* samples, according to the majority voting method to determine the category of the prediction point.

## 2.2 SVM modeling

SVM model (Support Vector Machine) is a kind of classification or regression model based on interval maximization and kernel function technique, and its core idea is to find an optimal hyperplane in the feature space, so that there is a maximum geometric interval between the samples of different categories, thus improving the generalization ability.

According to the principle of SVM, the modeling is to find the optimal separating hyperplane (the hyperplane that separates the samples with the maximum interval) separating the two classes of samples. The optimal separation hyperplane can be notated as:

$$w^{\mathrm{T}}\mathrm{x} + \mathrm{b} = 0 \quad (2)$$

Points located above the optimal separating hyperplane in this way satisfy:

$$w^T\mathrm{x} + \mathrm{b} > 0 \quad (3)$$

points located below the optimal separation hyperplane are satisfied:

$$w^T\mathrm{x} + \mathrm{b} < 0 \quad (4)$$

By adjusting the weights $w$, the hyperplane of the edge can be written as:

$$\mathrm{H}_1: \quad w^T\mathrm{x} + \mathrm{b} \geq 1 \quad (5)$$

For all $y_i = +1$

$$\mathrm{H}_2: \quad w^T\mathrm{x} + \mathrm{b} \leq -1 \quad (6)$$

For all $y_i = -1$

That is, those falling in $H_1$ or above it are positive categories and those falling in $H_2$ or below it are negative categories, which are obtained by combining the above:

$$y_i(w^T x + b) \geq 1, \forall i \quad (7)$$

The training samples that fall on $H_1$ or $H_2$ are called support vectors. where $w$ is the normal vector, *b* is the offset, and $x$ is the data point.

The distance from the optimal separable hyperplane to any point on $H_1$ is $\frac{1}{||w||}$, and similarly to any point on $H_2$ is $\frac{1}{||w||}$, the maximum edge spacing is $\frac{2}{||w||}$.

For the case of nonlinear separability, the original data can be transformed to a higher dimensional space by a nonlinear mapping, and linear separability can be achieved in the new higher dimensional space (Huang, L. et.al.). This nonlinear mapping can be realized by the kernel function, which is used in this topic, and the Gaussian kernel function is formulated as follows:

$$K(x_i, x_j) = e^{-||x_i - x_j||^2/2\delta^2} \quad (8)$$

Where, $x_i$ and $x_j$ are the input sample points, $||x_i - x_j||^2$ denotes the Euclidean distance between the two sample points, and $\delta$ is the bandwidth of the Gaussian kernel function, which controls the width of the Gaussian distribution (Ren, J. et.al.). The larger the bandwidth $\delta$, the wider the range of the Gaussian kernel function, the smoother the decision boundary; the smaller the bandwidth δ, the narrower the range of the Gaussian kernel function, the more complex the decision boundary. The SVM model used in this study is the default parameter of SVC.

$$y(t) = \sum_{k=1}^{K} C_k(t) + r_K(t) \quad (9)$$

# 3. Model solving

CEEMDAN decomposition is performed on the given seismic signal data. CEEMDAN decomposition is an improved empirical modal decomposition (EMD) method, which can decompose nonlinear and nonsmooth signals into a series of intrinsic modal functions (IMFs), and effectively reduces the problems of modal aliasing and noise residuals.

The principle of CEEMDAN decomposition is to add pairs of positive and negative Gaussian white noises to the signal to be decomposed, and then perform EMD decomposition on each noise-added signal to obtain a set of IMFs, and then average the IMFs of each order to obtain the final IMFs. The mathematical formulas of CEEMDAN decomposition are as follows:

$$y(t) = \sum_{k=1}^{K} C_k(t) + r_K(t) \quad (10)$$

where $K$ is the number of IMFs, $C_k(t)$ is the kth order IMF, and $r_K(t)$ is the final residual. The seismic signal is decomposed into IMFs with a residual vector by CEEMDAN decomposition. After the decomposition, the sample entropy is solved for the first seven IMFs after the decomposition. The sample entropy is a measure of the complexity of the time series, which can reflect the irregularity and unpredictability of the time series. The larger the sample entropy is, the more complex the time series is, and the more difficult it is to be described by a simple model.

The mathematical formula for sample entropy is as follows:

$$\mathrm{SampEn}(m, r, N) = \ln \frac{A_m(r)}{B_m(r)} \quad (11)$$

Where m is the pattern dimension, which indicates how long the data segments are used to compare the similarity; r is the similarity tolerance threshold, which indicates the maximum difference between two data segments; $N$ is the data length; $A_m(r)$ is the number of pairs of data segments of length $m+1$ that satisfy the similarity condition; $B_m(r)$ is the number of pairs of data segments of length $m$ that satisfy the similarity condition.

# 4. Conclusion

The obtained entropy value is randomly divided into training set and test set according to 8:2, KNN model and SVM model are constructed for training and testing evaluation, and the evaluation indexes are Accuracy, Precision, Recall and F1score.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (12)$$

$$Precision = \frac{TP}{TP + FP} \quad (13)$$

$$Recall = \frac{TP}{TP + FN} \quad (14)$$

$$F1score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (15)$$

Where $TP$ means that the sample predicted values match the true values and are both positive, i.e., true positive, $FP$ means that the sample predicted values are positive while the true values are negative, i.e., false positive, $F$N means that the sample predicted value is negative while the true value is positive, i.e., false negative, and $TN$ means that the sample predicted value matches the true value and both are negative, i.e., true negative. The comparison results of the model on the test set in this task scenario are shown in Table 1:

Table 1Comparison results of model testing

| Model | Accuracy | Precision | Recall | F1 score |
|---|---|---|---|---|
| KNN | 1.0 | 1.0 | 1.0 | 1.0 |
| SVM | 1.0 | 1.0 | 1.0 | 1.0 |

As can be seen from Table 1, the task gets good results on all three models, with scores of up to 100% for each. By building KNN, and SVM models for classifying signals from natural earthquakes and artificial blasts, all three models perform extremely well on the task with up to 100% in Accuracy, Precision, Recall, and F1 value (F1score) scores, proving

the usability of the models.

# References

[1] Patterson, B., Leone, G., Pantoja, M., & Behrouzi, A. A. (2018). Deep learning for automated image classification of seismic damage to built infrastructure. In *Eleventh US National Conference on Earthquake Engineering*.

[2] Abdalzaher, M. S., Moustafa, S. S., Abd-Elnaby, M., & Elwekeil, M. (2021). Comparative performance assessments of machine-learning methods for artificial seismic sources discrimination. *IEEE Access*, *9*, 65524-65535.

[3] Yavuz, E., Iban, M. C., & Arpaz, E. (2023). Identifying the source types of the seismic events using discriminant functions and tree-based machine learning algorithms at Soma Region, Turkey. *Environmental Earth Sciences*, *82*(11), 1-15.

[4] Rincon-Yanez, D., De Lauro, E., Petrosino, S., Senatore, S., & Falanga, M. (2022). Identifying the Fingerprint of a Volcano in the Background Seismic Noise from Machine Learning-Based Approach. *Applied Sciences*, *12*(14), 6835.

[5] Huang, L., Li, J., Hao, H., & Li, X. (2018). Micro-seismic event detection and location in underground mines by using Convolutional Neural Networks (CNN) and deep learning. *Tunnelling and Underground Space Technology*, *81*, 265-276.

[6] Ren, J., Zhou, S., Wang, J., Yang, S., & Liu, C. (2022). Research on Identification of Natural and Unnatural Earthquake Events Based on AlexNet Convolutional Neural Network. *Wireless Communications and Mobile Computing*, *2022*.