Detecting Seismic Electric Signals by LVQ Based Clustering

Kyoko Fukuda* Mika Koganeyama* Hayaru Shouno*
Toshiyasu Nagao† Kazuki Joe*
kyochan@ics.nara-wu.ac.jp

*Nara Women's University
† Earthquake Prediction Research Center, Tokai University

Abstract

Aiming at short-term prediction of earthquakes, we have proposed the use of neural networks for analyzing telluric current data observed by the VAN method. We have already tried a telluric current data analysis method with Learning Vector Quantization. In this paper, we will show preliminary experimental results for categorization of telluric current data by its frequency for the Izu islands earthquakes in Japan.

1 Introduction

In Japan, short-term earthquake prediction is one of the most important problems because earthquakes are quite frequent and the damage caused by the great Hanshin earthquake in 1995 was severe. However an effective method for short-term earthquake prediction has not been established yet. Furthermore, it is commonly believed to be impossible to predict earthquakes effectively in the short term. However, the VAN method [1][2], which is a short-term earthquake prediction, has been in the limelight recently.

The International Frontier Research Group on Earthquakes (IRFEQ) [6] has begun investigating the VAN method in Japan. [1] In telluric current data (TCD) observed by VAN method, seismic electric signals (SESs) are often detected before the occurrence of great earthquakes. Although experts of the VAN method can recognize SESs with a careful glance, they are not mathematically modeled as time series data. In addition, since 90 percent of TCD in Japan is affected by train noise, detecting SESs in TCD itself is considered to be an extremely difficult job. We have succeeded in removing train noises from TCD by neural networks with the back propagation (BP) learning method [5]. At present, TCD is recorded every ten seconds and telemetered to IRFEQ through the public telephone system. The total amount of TCD observed in Japan in the last four years has grown too large (several terabytes) to be analyzed by hand.

Neural networks with BP is very good at recognizing patterns, but the huge amount of calculation necessary for learning makes it difficult to establish the neural network system as a practical solution. Several neural network systems with BP which solve those real-wolrd problems has adevice for reducing the calculation cost.

In a previous study, we have proposed a filter to remove train noise by dividing a series of TCD into set frames with 300 points. In this way, we succeeded in analyzing only specific time periods for an observation point. However, the straightforward extension of this neural network system may not be practical because various sets of observed data may make the resultant new learning data too complicated.

LVQ(Learning Vector Quantization) is a collection of neural network models. The advantage of LVQ to BP is a small calculation cost though the ability to recognize patterns is enough good compared with BP in some cases. In preliminary research, we have tried applying LVQ to an SES detecting system with the data of Matsushiro. In the experiments, we could not detect SESs, but we could only detect the difference of the season. The data from Matsushiro is one of the clearest data sets that is effected by train noise in Japan. Therefore, we determined that the data from Matsushiro was not good for the preliminary test of the LVQ system, and tried to test again using data observed at another point, which contained less train noise.

We focused on the TCD observed in Nijima Island. Nijima Island is close to Miyakejima Island, on which eruptions and earthquakes occurred during the summer of 2000. There is little train noise because of geographical factor in the data from Nijima Island. The SESs for the earthquakes, which occurred on the 1st and the 9th of July, were detected by experts. Furthermore, Nagao *et al.* analyzed the TCD of Nijima and reported that SESs were detected in the 10mHz band (approximately)[8]. For the reasons mentioned above, we chose the data of Nijima Island for the test of LVQ. In the rest of the paper, we report the construction of an LVQ system for detecting SESs from TCD observed in Nijima Island and the results from evaluating this system.

2 System Construction and Evaluation

It is possible to detect SESs from TCD when the characteristics of SESs are well known. Unfortunately, the characteristics of SESs have not been

[1] International Frontier Research Group on Earthquakes is under the Ministry of Education, Culture, Sports, Science and Technology in Japan.

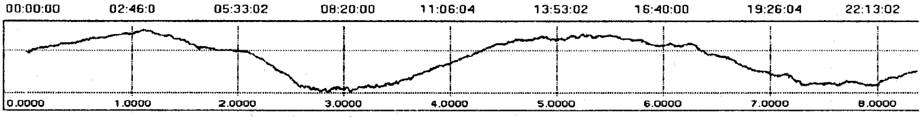

Figure 1: the TCD observed in Nijima Island

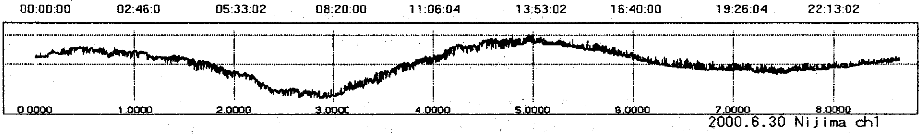

Figure 2: the TCD with SESs (2000.6.30)

modeled yet so only experts on the VAN method can find it. Fig.3 shows examples of SES data. In the figure, the vertical and the horizontal axis stands for the voltage and the frequency respectively. The SESs are indicated by underlined part of each figure.

In this study, we selected the TCD of Nijima Island, where relatively strong earthquakes occurred in 2000. There are two reasons to use this data in the analysis. The first reason is that some experts have confirmed the TCD data contain SESs that was observed a few days before the earthquake . The second reason is that the data was free of train noise. Fig.1 shows typical TCD data from Nijima. Fig.2 shows the data of the 30th of June 2000, the day before a big earthquake occurred. A number of small waves can be clearly observed. So, we planned to analyze the frequency of the data. In this study, we compose an LVQ system to classify the inputs into either normal data or anomaly one which contains SESs. We train the LVQ with the normal data, so that the abnormality is found when the SES data is given to the system. We construct the system and examine it as follows:

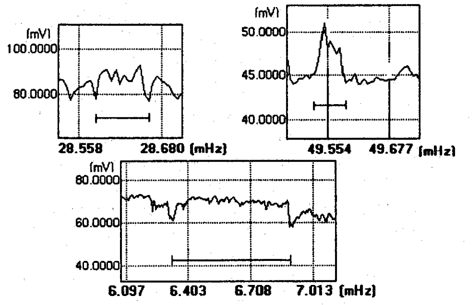

Figure 3: Examples of SESs

1. Preprocessing the data(smoothing and applying FFT)
2. Constructing the LVQ system
2-1. Initializing reference vectors
2-2. Updating rules of reference vectors
3. The recognition experiment

The detail is discussed in the following section.

2.1 Preprocessing the data

In this study, we used data from the Nijima observation point. This observation point has eight observation channels toward different directions. We choose the data of channel 1, from March to November in 2000 (240 days), because this channel contains SESs clearly before earthquake. To remove unreliable data, which were caused by problems with the observation instruments, we calculated the derivative of the data, and smoothed the sections that have large derivative values. Next, we transformed the temporal data to frequency component data by Fast Fourier Transform(FFT). To apply FFT for TCD of each day, we should regulate the size of each datum as 2^n. The size of the datum of each day is 8640 sampling points, therefore we cut the last 448 points.

After applying FFT, 8192 band frequency data was generated. We call this data the Transformed Terrulic Current Data (TTCD) in this paper. After that, we divided the TTCD into sixty-four narrowband (there are 64 bands, and each band width is 128) All the data for a specific frequency band is denoted by follows. (where f is the index of the frequency):

$$\boldsymbol{X}_0^f, \boldsymbol{X}_1^f, \cdots, \boldsymbol{X}_{239}^f,$$

f	frequency (mHz)		
0	0	~	0.775
1	0.775	~	1.556
⋮		⋮	
63	49.21	~	50.00

To make the training data, we excluded the data containing SES. The data containing SESs was recorded on the following five days: the 28th ~ the 30th of June, and the 7th ~ the 8th of July. The number of data for training is 235, and those are denoted as

$$\boldsymbol{x}_0^f, \boldsymbol{x}_1^f, \cdots, \boldsymbol{x}_{234}^f.$$

2.2 Constructing the LVQ system

In this study, to decide the initial state of reference vectors, we completed the following two steps.

1. Rough clustering of thetraining data set.
2. Initializing reference vectors with *k-means*method.

This process will now be discussed in more detail.

Rough clustering For rough clustering, we introduce new vectors denoted by $\boldsymbol{v}_0, \boldsymbol{v}_1, \cdots, \boldsymbol{v}_{m-1}$ and plan the deviding training data set into m clusters. We decide each vector to represent the typical vector of each cluster. These vectors represent the typical data of each cluster. These vectors, $v_0, v_1, \cdots, v_{m-1}$ are calculated as follows. First, we calculate the distance of the training data. $x^f_0, \cdots, x^f_{234}$ by Euclidian measure, and calculate the mean and the standard deviation.

$$d_{ij} = d(x^f_i, x^f_j) = \sqrt{\sum_{n=1}^{128}(x^f_i(n) - x^f_j(n))^2}$$

M_d : average of d_{ij} $\quad \sigma_d$: standard deviation of d_{ij}

The numbers i and j refer to the date indices, and n refers to the frequency. After getting the averages and the standard deviation, we selected twenty data randomly from the training data. We call these twenty data the candidate data. We select two data x^f_p, x^f_q from the candidate data. x^f_p, x^f_q to satisfy the following condition:

$$d_{pq} > M_d + 2\sigma_d \qquad (1)$$

If any combinations in candidate data do not satisfy the equation (1), we select another twenty data for the candidate randomly. When x^f_p, x^f_q satisfies the condition(1), we set $v_0, v_1$to x^f_p, x^f_q respectively. After the vectors $v_0, v_1, \cdots, v_s$ were selected, we selected two data points x^f_k and x^f_l from the candidate data, whose distance x^f_k, x^f_l satisfied the equation $d_{kl} > M_d$. Then, we calculated the distances between x^f_k and each $v_0, v_1, \cdots, v_s$. If each distance satisfies

$$d(x^f_k, v_i) > M_d + 2\sigma_d \quad \text{for all(i} = 1 \sim \text{s)},$$

then, x^f_k is set to v_{s+1}. On the other hand, if the distance is too short, we never select the data x^f_k as the candidate data. The criterion for disqualification due to shortness is:

$$d(x^f_k, v_i) < M_d - 2\sigma_d \quad \text{for any (i} = 1 \sim \text{s)}.$$

After that, we examine the data x^f_l in a similar manner. We repeat the selecting process as mentioned above until m data have been selected. But the datum, which is judged as too close to selected vectors, or which has been already selected, are never selected. After checking all the candidate data, if the number of vectors is less than m, another twenty candidate data are selected randomly. After the vectors $v_0, v_1, \cdots, v_{m-1}$ were selected, we calculated the distances between each training data point $x^f_0, x^f_1, \cdots, x^f_{234}$ and the selected $v_0, v_1, \cdots, v_{m-1}$ vectors. Learning data point is regarded to be in the same class as the closest vectorv_i.

Initializing reference vectors In the preceding section, the data for learning was classified into m classes. Next, we must determine the initial state of reference vectors for LVQ. We used the *k-means* method for each class to determine the typical vectors of the class, and used them as the initial reference vectors.

2.3 Training the LVQ

In this study, the LVQ training process consists of the following 2 steps.

1. Updating the reference vectors
2. Re-labeling the reference vectors

Each reference vector has a label, which exepress its category. Each category consists of plural reference vectors with the same label. After updating reference vectors, some categories can potentially be close together. In that case, we fuse them by re-labeling the reference vectors. If the fusion of categories occurred, more training and re-labeling of the reference vectors is needed. This process will iterate until the no further fusion occurs. The following describes the process in detail.

Updating reference vectors In this study, the number of iterations is not provided, but learning is finished when the total movements of all reference vectors converged to 1/10 of the average M_d.

Re-labeling reference vectors After updating the reference vectors, re-labeling is necessary. If two categories are very close, these categories needed to be fused. So we search the closest categories by calculating the average distance between the reference vectors belonging to each categories. If category B was the closest to category A, we calculated the distance and average of them M_{AB}, The condition for fusing is $M_{AB} < M_d + \sigma_d$.
If M_{AB} satisfies this condition, all the labels of category B are changed to A.

2.4 The recognition experiment

After updating and re-labeling, we evaluate the system with the TTCD (including the SES data). We calculated the distances between each datum in the TTCD and the reference vectors, and its class is the same as the class of the closest reference vector. The LVQ method classifies the input as the same group with the closest reference vector regardless of the distance between the input data and the reference vector. We assume the SES contained data will be far away from any normal data. Therefore we provide the threshold; if the distance between the data and the closest reference vector is larger than a constant value, the datum considered as an exception.

Let D_i denote the distance between the TTCD X^f_i and the closest reference vector r ($i = 0 \sim 239$). If D_i satisfied the following condition, the data will be detected as an error. Let the distance between the elements of the TTCD ($X^f_0, \cdots, X^f_{239}$) and the closest vector r be denoted d. If D_i satisfies the following condition,

$$D_i = d(X^f_i, r) > M_d + \sigma_d.$$

the datum X_i is detected as an error.

3 Experimental Results

We analyzed the average and dispersion of the TTCD for each frequency band (bandwidth = 128).

First, we calculated the average and the standard deviation of the distances for all TTCD. Fig.4 shows the average distance versus frequency. In the figure, the higher the frequency is, the smaller the average is. Fig.5 shows dispersion versus frequency, and the standard deviation is larger in higher frequency.

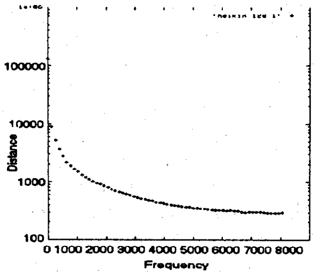

Figure 4: The average of distance

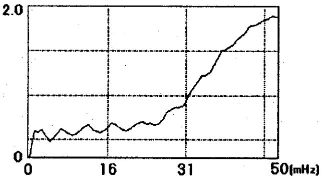

Figure 5: The dispersion versus frequency (standard deviation /average)

In the LVQ experiment with the training data, few errors are detected in the $8mHz \sim 50mHz$ band. Then, we ran the entire TTCD set through the process five times. All the SES data (28th, 29th, 30th in June, and 7th in July) were detected successfully. The data from the 8th of July was hard to be detected because few SES components were present in the data. However, it can be detected in the frequency band near $12.5mHz$. The data of the 29th of June is also detectable in the band $12.5mHz \sim 32mHz$. Fig.6 show which band had had errors that detected the SESs. The input data is Jun 28th, 2000. Another 4 days data are also detected. The y-axis showed the number as an error in the five experiments. Fig.7 show the number of errors for each band.

4 Summary

We proposed a prototype of an automatic system for short-term earthquake prediction using the VAN method. The goal of this research is to detect SESs from TCD that proceed earthquakes. It has been said that only experts could find the SES because the characteristics of SESs have not been modeled yet. However, our system could detect all the SESs in the TTCD we used. In this study, we chose the TTCD of Nijima Island observation point near Miyakejima Island where the eruption was occurred with earthquakes in 2000. Experts guessed that SESs are contained in the TTCD. We transformed the temporal data to the frequency data by FFT, and analyzed them. The datum which does not includes SESs were used for training. After learning, we applied all of the data including SESs to the system, and succeeded in detecting SESs at specific bandwidths. This result is supported by other research at IRFEQ. In this research, SESs are detected on the specific band, even though the characteristics of SES are not well known. Therefore, this report shows very effective results.

References

[1] Uyeda,S.: *Introduction to the VAN method of earthquake prediction, Critical Review of VAN (ed.Sir James Lighthill)*, World Scientific, pp.3-28 (1996).

[2] Nagao,T., Uyeshima,M., Uyeda,S.: *An independent check of VAN's scriteria for signal recognition*, Geophys. Res. Lett. 23, pp.1441-1444 (1996).

[3] Kohonen,T. : *Learning Vector Quantization for Pattern Recognition*, TKK-F-A601, Helsinki U (1986).

[4] Yokota,M., Katagiri,S., McDermott,E.: *learning in an LVQ-Based Phoneme Recognition System*, IEICE, SP88-104, pp.65-72 (1988) (in Japanese).

[5] Koganeyama,M., Nagao,T., Joe,K.: *Removing Train Noise from Telluric Current Data by Neural Networks for Automatic Short-term Earthquake Prediction in Japan*, PDPTA2000, Vol.II. pp.659–665 (2000).

[6] Int'l Frontier Research Group On Earthquakes : *http://yochi.iord.u-tokaki.ac.jp/eprc*

[7] Fukuda,K., Koganeyama,M., Nagao,T., Joe,K.: *Terrulic Current Data Analysis by Learning Vector Quantization*, IPSJ SIGMPS, 108-MPS-32, pp.25-28 (2000).

[8] Nagao,T., Hattori,K., Hayakawa,M., Uyeda,S. : *Ground based intensive observation by RIKEN and NASDA Frontier Programs on Earthquakes*, Workshop on Natural Disaster Monitoring by Satellite 2001, Paris (in press).

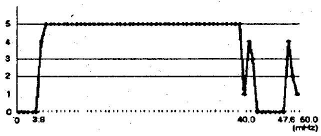

Figure 6: the result of the LVQ error. The input data is Jun 28th, 2000

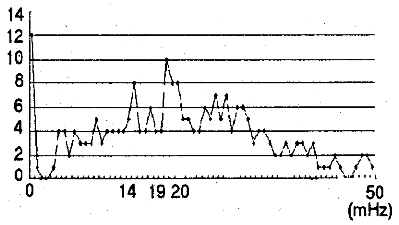

Figure 7: the number of data as error, in each frequency band