Identification of ElectroEncephaloGraph signals using sampling technique and K - nearest neighbor

H Hindarto¹*, A Muntasa², A Efiyanti¹

¹, ¹Universitas Muhammadiyah Sidoarjo, Indonesia
²Universitas Trunojoyo Madura, Indonesia
hindarto@umsida.ac.id*

Abstract. Electroencephalograph is a device that can help humans observe and analyze the results of electrical waves produced by neurons in the brain. The results of reading the tool are called Electroencephalogram (EEG), besides being able to help diagnose physician for medical therapy, it is also developed for the Brain-Computer Interfacing (BCI) application. BCI is a method that allows humans to be able to control an external system without direct contact with the system. Research on communication between humans to control external equipment has been widely investigated, including research on brain activity to control a cursor on a computer screen. This study focuses on feature extraction for ElectroencephaloGraph (EEG) signals using the sampling technique. K - Nearest Neighbor is used as a classification of EEG signals to determine whether the cursor moves up or down. The data used are EEG data originating from the 2003 BCI competition (BCI 2003 Competition). Decision making is done to classify the cursor movement up and down the cursor movement. The research data uses 250 EEG signal file training data and 50 from EEG signal file testing data, so that the whole becomes 300 EEG signal data files. The best results with K = 3 values obtained for the classification of EEG signals using K-NN are 76% of the signal data tested.

1. Introduction

The activity of recording electricity in the brain for electro physiological monitoring is called Electroencephalograph (EEG). Electrode placement is located along the scalp. The EEG measures voltage fluctuations that result from ion currents in brain neurons. In the clinical context, EEG refers to the recording of spontaneous electrical activity of the brain over a period of time, as recorded from several electrodes placed on the scalp [1]. Diagnostic applications generally focus on the potential associated with EEG spectral content.

Placing electrodes on the scalp follows a predetermined system, namely the 10-20 system. The right and good electrode placement is the main requirement to get good and reliable EEG recordings. Besides that the cleanliness of the scalp, the condition of the electrode, the EEG machine and the compliance of the subjects during recording are also very influential to get good results. Hans Berger stated that the human brain has continuous electrical activity that can be recorded. Brain activity can be possible to send commands to electronic equipment with the help of the Brain Computer Interface (BCI) [2]. BCI can be used for mental activities such as thinking moving hands, thinking moving arms and others. This work will produce different EEG signals [3]. The results of these different EEG signals will produce different actions for information for BCI. An evaluation of EEG signals in conducting movement activities in recent years was conducted by researchers [4]. However, to improve the performance of EEG signals in detecting the characteristics of these signals is a big bustle for researchers. For the
The process of taking feature extraction and classification is an important role in recognizing EEG signals. Many are used in finding features for EEG signals, including using the Fast Fourier Transform (FFT) method \[5\][6]. Taking feature extraction with the Fast Fourier Transform by taking the value from the spectral using the Welch method. The weakness of this FFT method is that it only takes frequency information while time information is not taken. Research on how to combine information from frequency and time has been investigated in this study \[7\]. One other study is to take the Autoregressive (AR) method \[8\]. In addition, in taking the features used by using the AR or multivariate autoregressive (MVAR) model \[9\]. Wang et al also analyzed the combined time-frequency for EEG signal components \[10\]. The method to find the features of an EEG signal is to use Wavelet transforms by utilizing the wavelet transform coefficient \[11\]. In research using wavelet transform mechanisms for the production of EEG signals is rather complicated, making it difficult to obtain accurate transcendent information. Other research by taking signal features \[12\]. Using the Simple Random Sampling (SRS) method for extracting EEG signal features of epilepsy patients has been carried out in this study and the results of the LS-SVM classification obtained were 80.31% for training data and 80.05% for testing data \[13\].

Based on the above background, this research is designed as follows, the second part explains the EEG signal materials and methods used in the search for EEG features, the third part explains the results of feature extraction and the classification process of EEG signals, and the fourth section describes the conclusions of this study.

2. Materials And Methods

2.1. material

This study takes data from the 2003 BCI competition uploaded by Dr. Birbaumer and his team at Tuebingen University, Germany (Blankertz 2004) \[17\]. The recording process is taken from a healthy subject by attaching 6 electrode sensors that are attached to the scalp. Recorded using 256 Hz sampling frequency and recording time in a period of 3.5 seconds. The process of recording the subject is asked to imagine the movement of the curve upwards and move the cursor down on a monitor screen. Subjects in the recording process receive visual feedback from SCP. This experiment consisted of 268 training data and 293 trial data.

![Figure 1. Position of EEG electrodes](image)

Figure 1 shows the position of the electrode attached to the head. There are 6 sensors used, namely Cz for channel 1, A2-Cz for channel 2, FC3 for channel 3, CP3 for channel 4, FC4 for channel 5 and PC4 for channel 6 \[17\].

![Figure 2. Timing pattern](image)
This research takes one channel from the 6 channels installed, namely channel 3, each channel has two different classes. The data used for each class were 688 data for training and 293 data for trial data. The data used in this study are 250 training data and 50 trial data.

2.2. Sampling technique
In this study extraction, the characteristics used are the Sampling Technique method. The EEG signal is divided into four sub-EEG signals. Of the four sub-EEG signals determined the maximum value, the minimum value, the average value and the standard deviation value to be extracted from the characteristics of the EEG signal as shown in figure 3.

![Figure 3. a. Signal sample, b. sub signal samples and c. Feature Selection](image)

There are 300 EEG signal data files taken in this study. One EEG signal file data has 896 data points. One EEG signal is divided into four sub-signals. So that each sub-signal has 214 points of data. The EEG sub-signal data is searched for the minimum value, maximum value, average value and standard deviation value. There are 20 points of data obtained from 4 x 4 sub-signals.

2.3 K-Nearest Neighbor Classification
K-NN is a simple machine learning algorithm. This is only based on the idea that an object that is 'close' to each other will also have similar characteristics. This means that if we know the characteristics of one object, we can also predict other objects based on their closest neighbors. K-NN is an advanced improvisation of the Nearest Neighbor classification technique. This is based on the idea that each new example can be classified by the majority vote of a neighbor k, where k is a positive integer, and usually with a small number [14]. The K-NN classification algorithm predicts the sample test category according to the training sample k which is the closest neighbor to the test sample, and enters into the category that has the largest probability category [15].

In pattern recognition, the KNN algorithm is a method used to classify objects based on the closest training example in the feature space. KNN is a type of insurance based learning, or lazy learning where this function is only approached locally and all calculations are deferred to classification [16]. The K-NN classification method has several stages, the first being the k value which is the number of closest neighbors that will determine which new query goes to which class is determined. The second stage, k the nearest neighbor is searched by calculating the distance of the query point with the training point. The third stage, after knowing the distance of each training point with the query point, then see the smallest value. The fourth stage takes the smallest value, then see the class. The class that is the most is the class of the new query. Near or far points with neighbors can be calculated using the Euclidean distance. Euclidean distance is represented as follows:

\[ j(a, b) = \sqrt{\sum_{k=1}^{k=n} (a_k - b_k)^2} \]  

(1)
4

\( J(a, b) \) is the distance between point \( a \) which is the point known to its class and \( b \) is a new point. The distance between the new point and the training points is calculated and taken by the nearest point. New points are predicted to enter the class with the most classifications of these points.

3. Results And Discussion

Results of experiments conducted by researchers by taking data from the 2003 BCI competition data set Ia. First of the 6 channel electrodes / channels are taken just one channel in the sampling process. One signal from a set of signals that are joined in one channel is divided into 4 sub-signals, then from each sub-signal the search for the maximum value, minimum value, average value and standard deviation values are searched. This value will be used as a feature extraction for the identification process.

Figure 4. Results of the Sampling Technique process on EEG signals

Figure 4 is an EEG recording and the division process into four sub-signals. The four sub-signals are taken the maximum value, the minimum value, the average value and the standard deviation value for each sub-signal.
Figure 5. Comparison of class 0 and class 1 to the maximum value, minimum value, average value and standard deviation value of each sub signal

In Figure 5 shows that for values of maximum values, minimum values, average values and standard deviation values for each sub signal. Each signal for the cursor movement up and down the cursor has a difference value. With differences in values that are not the same, it shows that the level of classification by taking the maximum value, minimum value, average value and standard deviation value is quite good. Classification using K-NN is implemented using the maximum value feature, minimum value, average value and standard deviation value from the sampling technique as input. In this study, the training set numbered 250 EEG signal data and data of test is 50 EEG signal data. In table 1 is the distribution of sample classes in the training and data of test collection. Data obtained from different subjects in the training process is a way to improve the ability of K-NN. To train K-NN, use training data sets, while to verify the accuracy and effectiveness of K-NN test data to detect cursor movements up and down.

Table 1. Sampling of class distribution on training data and data of test

| Class          | Training set | Test set |
|----------------|--------------|----------|
| Up cursor      | 125          | 50       |
| Down cursor    | 125          |          |

The test will be carried out with a variation of the K value in the KNN system and variations in training data and test data used. The accuracy in question is the accuracy of the classification results carried out by the program with the results of manual classification. The level of accuracy can be formulated with equation (2).

\[
Level\ of\ accuracy = \frac{\text{The number of EEG signals is correctly classified}}{\text{Total amount of test data}} \times 100\% \quad (2)
\]

In this test variations in the number of K are performed on the KNN function. k is the number of closest neighbors. The K values tested are one, three, five, seven, and nine. Odd values are chosen to avoid similarities in proximity to two different classes. Because KNN will classify based on the most class voting. From this test the average accuracy value for each k value is determined. Classification Testing with k Value = 1
The classification results with $K = 1$ are shown in Table 2. The test results with $K = 1$ obtained an average accuracy of KNN classification of 54.4%. With the lowest accuracy 48% and the highest accuracy 64%.

| Test | Amount of training data | Amount of test data | Right | wrong | accuracy |
|------|-------------------------|---------------------|-------|-------|----------|
| 1    | 50                      | 50                  | 13    | 12    | 52%      |
| 2    | 50                      | 50                  | 13    | 11    | 52%      |
| 3    | 50                      | 50                  | 14    | 11    | 56%      |
| 4    | 50                      | 50                  | 16    | 9     | 64%      |
| 5    | 50                      | 50                  | 12    | 13    | 48%      |
|      |                         |                     |       |       | **54.4%**|

Classification Test with $k = 3$

The classification results with $K = 3$ values are shown in Table 3. The test results with $K = 3$ obtained an average accuracy of 64%. The highest level of accuracy is found in the fifth test of 76%. The lowest level of accuracy is found in the third test of 56%.

| Test | Amount of training data | Amount of test data | Right | wrong | accuracy |
|------|-------------------------|---------------------|-------|-------|----------|
| 1    | 50                      | 50                  | 14    | 11    | 56%      |
| 2    | 50                      | 50                  | 14    | 11    | 56%      |
| 3    | 50                      | 50                  | 16    | 9     | 64%      |
| 4    | 50                      | 50                  | 19    | 6     | 76%      |
| 5    | 50                      | 50                  | 17    | 8     | 68%      |
|      |                         |                     |       |       | **64%**  |

Classification Test with $k = 5$

The classification results with $K = 3$ values are shown in Table 4. The test results with $K = 5$ obtained an average accuracy of 54.4%. The highest level of accuracy is found in the fifth test with 64%. The lowest level of accuracy is found in the first and second test of 44%.

| Test | Amount of training data | Amount of test data | Right | wrong | accuracy |
|------|-------------------------|---------------------|-------|-------|----------|
| 1    | 50                      | 50                  | 11    | 14    | 44%      |
| 2    | 50                      | 50                  | 11    | 14    | 44%      |
| 3    | 50                      | 50                  | 15    | 10    | 60%      |
| 4    | 50                      | 50                  | 16    | 9     | 64%      |
| 5    | 50                      | 50                  | 15    | 10    | 60%      |
|      |                         |                     |       |       | **54.4%**|

Classification Test with $k = 7$

The classification results with $k = 7$ are shown in Table 5. Testing with $k = 7$ produces an average accuracy of 60.8%. The highest accuracy results are found in the fifth test with a value of 68% accuracy. The lowest accuracy of 52% is in the third test.
Table 5. Results of testing k = 7 classification

| Test | Amount of training data | Amount of test data | Right | wrong | accuracy |
|------|-------------------------|---------------------|-------|-------|----------|
| 1    | 50                      | 50                  | 15    | 14    | 60%      |
| 2    | 50                      | 50                  | 15    | 14    | 60%      |
| 3    | 50                      | 50                  | 13    | 10    | 52%      |
| 4    | 50                      | 50                  | 16    | 9     | 64%      |
| 5    | 50                      | 50                  | 17    | 10    | 68%      |
|      |                         |                     |       |       | Average 60.8% |

Classification Test with k = 9

The results of classification testing with k = 9 values are shown in Table 6. Classification results with a value of k = 9 indicate an average accuracy value of ten testing times of 62.4%. Of the five tests there were three tests which produced 64% accuracy, namely in the second, fourth and fifth tests. The lowest test value of 60% is found in the first and third tests.

Table 6. Results of testing k = 9 classification

| Test | Amount of training data | Amount of test data | Right | wrong | accuracy |
|------|-------------------------|---------------------|-------|-------|----------|
| 1    | 50                      | 50                  | 15    | 14    | 60%      |
| 2    | 50                      | 50                  | 16    | 14    | 64%      |
| 3    | 50                      | 50                  | 15    | 10    | 60%      |
| 4    | 50                      | 50                  | 16    | 9     | 64%      |
| 5    | 50                      | 50                  | 16    | 10    | 64%      |
|      |                         |                     |       |       | Average 62.4% |

Comparison of Average Accuracy Values

All the tests with variations in the K value the highest accuracy value was found in the value of k = 3 with an average level of accuracy of 64%. The lowest accuracy value is found in tests with k = 1 and k = 5 with an average level of accuracy of 54.4%. Overall the value of accuracy has a value close to 60%. Classification of KNN with k = 1 is very susceptible to noise this results in a low level of accuracy. With the above description, the value of k = 3 is the most optimal k value.

4. Conclusion

This study can be concluded that the proposed KNN classification method can classify EEG signals with an accuracy of 76% at the best neighbour of value k = 3. The feature extracted for EEG signal classification uses the sampling technique.

5. Acknowledgement

The author thanks the Chairman of the Universitas Muhammadiyah Sidoarjo for providing time for research that the researchers are working on and the Directorate of Research and Community Service, Directorate General of Research Strengthening, and Development of the Republic of Indonesia Ministry of Research, Technology and Higher Education which supports funding for this research.

References

[1] Lux H D 1981 Generalized Tonic-Clonic Activity 24–8
[2] Wolpaw J R, Birbaumer N, Heetderks W J, McFarland D J, Peckham P H, Schalk G, Donchin E, Quatrano L A, Robinson C J and Vaughan T M 2000 Brain-computer interface technology: A review of the first international meeting IEEE Trans. Rehabil. Eng. 8 164–73
[3] Termal K, Menggunakan I and Neighbor K 2018 No Title
[4] Millán J 2003 Adaptive brain interfaces for communication and control Proc. 10th Int. Conf. ... 113 767–91
[5] ABDUL-BAREY RAOUF SULEIMAN T A-H F Features Extraction Techniques of EEG Signals for BCI Applications Toka Abdulhameed Fatehi Abdul-Ahmed Fatehi
[6] Jaswal R 2016 Brain Wave Classification and Feature Extraction of EEG Signal by Using FFT on Lab View Int. Res. J. Eng. Technol. 1208–12
[7] Mensh B D, Werfel J and Seung H S 2004 BCI competition 2003 - Data set Ia: Combining gamma-band power with slow cortical potentials to improve single-trial classification of electrophysiological signals IEEE Trans. Biomed. Eng. 51 1052–6
[8] McFarland D. J, Miner L. A, Vaughan T. M and Wolpaw J. R 2000 Mu and Beta Rhythm Topographies During Motor Imagery and Actual Movements Brain Topogr. 12 177–86
[9] Stolz E 1995 Multivariate autoregressive models for the classification of spontaneous electroencephalogram during mental tasks
[10] Wang T, Deng J and He B 2004 Classifying EEG-based motor imagery tasks by means of time-frequency synthesized spatial patterns Clin. Neurophysiol. 115 2744–53
[11] Adeli H, Zhou Z and Dadmehr N 2003 Analysis of EEG records in an epileptic patient using wavelet transform J. Neurosci. Methods 123 69–87
[12] Der M D Van 1995 Signal sampling techniques for data acquisition in process control Signal Sampling Techniques for Data Acquisition in Process Control Marten Derk van der Laan
[13] Li Y and Wen P 2009 Classification of EEG Signals Using Sampling Techniques and Least Square Support Vector Machines Classification of EEG Signals Using Sampling Techniques and Least Square Support Vector Machines
[14] Koutroumbas K and Kalouptsidis N 2002 Nearest neighbor pattern classification neural networks 2911–5
[15] Suguna N and Thanushkodi K 2010 An Improved k-Nearest Neighbor Classification Using Genetic Algorithm 7 7–10
[16] Imandoust S B and Bolandaftar M 2017 Application of K-Nearest Neighbor ( KNN ) Approach for Predicting Economic Events : Theoretical Background