SAMPLE SIZE DETERMINATION FOR CLASSIFICATION OF EEG SIGNALS USING POWER ANALYSIS IN MACHINE LEARNING APPROACH

V. Indira\textsuperscript{1}, R. Vasanthakumari\textsuperscript{2}, V. Sugumaran\textsuperscript{3}

\textsuperscript{1}Research Scholar, Depttment of Mathematics, Karpagam University, Coimbatore, India. indira_manik@yahoo.co.in
\textsuperscript{2}Principal, Kasthurba college for women, Puducherry-605008. vasuntharal@yahoo.com
\textsuperscript{3}Associate professor, School of Mechanical and Building Sciences, VIT university, Chennai campus, Vandalur-keelambakkam road, Chennai-600048 v_sugu@yahoo.com

ABSTRACT

Recent advances in computer hardware and signal processing have made possible the use of electroencephalogram (EEG) signals or “brain waves” for communication between humans and computers. Also, the signals that are extracted from brain through EEG are required to improve brain computer interface (BCI). EEG signal analysis and classification is one of the prominent researches in the field of BCI. In recent days, machine learning approach is used for classification of EEG signals. This method consists of a chain of activities namely, data acquisition, feature extraction, feature selection, and feature classification. In order to extract the features representing the EEG signals, the spectral analysis of the EEG signals was performed by using the three model-based methods (Burg autoregressive – AR, least squares modified Yule–Walker autoregressive methods. It is easy to collect EEG signals from healthy volunteers whereas it is too difficult to collect a sizable number of EEG recordings of those who are having one particular problem. Hence it is essential to know the minimum number of samples required to train the classifier. The objective of this paper is to determine the minimum number of signals required for training the classifier in machine learning approach to EEG signal analysis so as to get the best classification accuracy. A data set consists of 500 EEG signals with five different types was considered. A method for determination of
minimum sample size using power analysis has been proposed and the results are validated using J48 classifier (A java implementation of C4.5 algorithm).

Keywords: Auto regressive method, EEG signals; Machine learning approach; Power analysis; Sample size;

1. Introduction

Study of brain computer interface (BCI) has mainly involved in recording of electroencephalographic signals. The electroencephalogram (EEG), a highly complex signal is one of the most common sources of information used to study brain function and neurological diseases and has great potential for the diagnosis and treatment of mental diseases and abnormalities. In clinical contexts, EEG refers to the recording of the brain’s spontaneous electrical activity over a short period of time, usually 20-40 minutes, as recorded from multiple electrodes placed on the scalp. Also EEG signals are widely used in intensive care units for brain function monitoring. Hence the classification of EEG signals has become very essential. Larger and larger techniques are being used to analyze the EEG signal and extract useful information out of it. Machine learning approach may be considered as a candidate to analyze EEG signal. The dataset described in [1] which is publicly available was used for the present study. This data set consists of five sets denoted by (A, B, C, D and E), each containing 100 single-channel EEG signals. Each signal has been selected after visual inspection for artifacts and has passed a weak stationarity criterion. Sets A and B have been taken from surface EEG recordings of five healthy volunteers with eyes open and closed. Signals in two sets have been measured in seizure-free intervals from five patients in the epileptogenic zone (D) and from hippocampal formation of the opposite hemisphere of the brain (C). Set E contains seizure activity, selected from all recording sites exhibiting ictal activity. Sets A and B have been recorded extra cranially, whereas sets C, D and E have been recorded intracranially. Here only a short description of the data set is presented and refer to [1] for further details.

Much work has been performed in classification of EEG signals by various methods namely, linear and non-linear methods [2], support vector machine employing model based method [3-4], artificial neural network method [5-6], Hidden Markov model [7-8], non-iterative method based on Ambiguity function [9]. The cited works researched for better features and classifiers and reported the results. However, much work have not been anticipated towards the number of EEG signals to be used to train the classifier. In the present study, a statistical method is proposed to find the required minimum number of samples to be trained so as to get good classification accuracy with statistical stability.

Many works on minimum sample size determination have been reported in the field of bioinformatics and other clinical studies, to name a few, micro array data [10], cDNA arrays [11], transcription level [12] etc. Based on these works, data-driven hypotheses could be developed which in turn furthers EEG signal analysis research. An appropriate guideline was not proposed so far to choose minimum sample size of EEG signals for classification using machine learning approach. Hence, one has to resort to
some thumb rules which lack mathematical reasoning or blindly follow some previous work as basis for fixing the sample size. This is the fundamental motivation for taking up this study.

In this paper, F-test based statistical power analysis is used to find the minimum number of samples required to separate the classes with statistical stability. The minimum sample size is also determined using an entropy based algorithm called ‘J48’ algorithm. The results are compared and sample size guidelines are presented for EEG signals classification in the conclusion section.

2. Feature Extraction and Feature Selection

The feature extraction is performed through three techniques namely, Burg AR, Yule, Yule walker. Three features from each technique were obtained and thus totally nine features are available. Following the footsteps of [69], one can perform the feature selection using J48 decision tree algorithm to reduce the dimension of the dataset. In the present study, the most important five features were selected; they are f2, f4, f6, f7, f9. For the rest of the study, only these five AR features were considered.

3. Determination of Minimum Sample Size

This study uses a set of Burg AR, yule, and yule-walker features extracted from the EEG signals, instead of using EEG signals directly. With the selected set of five features along with their class labels form the data set to determine the minimum sample size using power analysis. The method of performing power analysis is discussed in section 4. The results of power analysis were verified with the help of a functional test namely J48 algorithm, as J48 algorithm could be used as a classifier with ten-fold cross validation method to validate the results. The results are presented and discussed in section 6.

4. Power analysis

Sample size has a great influence in any experimental study, because the result of an experiment is based on the sample size. Power analysis has been used in many applications [13-15]. It is based on two measures of statistical reliability in the hypothesis test, namely the confidence interval (1-α) and power (1-β). The test compares null hypothesis (H₀) against the alternative hypothesis (H₁). The null hypothesis is defined as the means of the classes are the same whereas alternative hypothesis is defined as the means of the classes are not the same. In hypothesis test, the probability of accepting null hypothesis is the confidence level and that of alternative hypothesis is the power of the test [10]. The estimation of sample size in power analysis is done such that the confidence and the power (statistical reliability measures) in hypothesis test could reach predefined values. The confidence level and the power are calculated from the distributions of the null hypothesis and alternative hypothesis. In case of multi-class problem (number of classes greater than two), instead of t-statistic, the F-statistic measure derived from Pillai’s V formula [16] is used for the estimation of sample size. Pillai’s V is the trace of the matrix defined by the ratio of between-group variance (B) to total variance (T). It is a statistical measure often used in multivariate analysis of variance
(MANOVA) [16]. The Pillai’s V trace is given by $V = \text{trace}(B^{-1}T) = \sum_{i=1}^{h} \frac{\lambda_i}{\lambda_i + 1}$ Where $\lambda_i$ is the $i^{th}$ eigen value of $W^{-1}B$ in which W is the within-group variance and $h$ is the number of factors being considered in MANOVA, defined by $h = c-1$. A high Pillai’s $V$ means a high amount of separation between the samples of classes, with the between-group variance being relatively large compared to the total variance. The hypothesis test can be designed as follows using F statistic transformed from Pillai’s $V$.

$H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_c$ ;

$H_1 : \text{There exists } i, j \text{ such that } \mu_i - \mu_j \neq 0 \quad H_0 : F = \frac{(V/s)/(ph)}{(1-(V/s))/[(s(N-c-p+s)]}$

$\sim F(ph, s(N-c-p+s))$ \hspace{1cm} $H_1 : F = \frac{(V/s)/(ph)}{(1-(V/s))/[(s(N-c-p+s)]}$

$\sim F[ph, s(N-c-p+s), \Delta = s\Delta cN]$

with $\Delta = \frac{V_{crit}}{(s-V_{crit})}$, where p and c are the number of variables and the number of classes, respectively. ‘$s$’ is defined by $\min(p, h)$. By using these defined distributions of $H_0$ and $H_1$, the confidence level and the power could be calculated for a given sample size and effect size. The method used for two-class problem is used here to estimate the minimum sample size for statistical stability whereby the sample size is increased until the calculated power reaches the predefined threshold of $1- \beta$.

5. J48 algorithm

The classification is done through a decision tree with its leaves representing the different brain conditions of the human beings. The sequential branching process ending up with the leaves here is based on conditional probabilities associated with individual features. It is reported that C4.5 model introduced by J.R. Quinlan has good predictive accuracy, good speed, minimum computational cost and better level of understanding. Decision tree algorithm (C4.5) has two phases: Building and Pruning. In building phase, the construction of decision tree depends very much on how a test attribute $X$ is selected. C4.5 uses information entropy evaluation function as the selection criteria. $Gain Ratio(X)$ compensates for the weak point of $Gain(X)$ which represents the quantity of information provided by $X$ in the training set. Therefore, an attribute with the highest $Gain Ratio(X)$ is taken as the root of the decision tree. In pruning phase, the C4.5 algorithm uses an error-based post-pruning strategy to deal with over-training problem. For each classification node C4.5 calculates a kind of predicted error rate based on the total aggregate of misclassifications at that particular node. The error-based pruning technique essentially reduces to the replacement of vast sub-trees in the classification structure by singleton nodes or simple branch collections if these actions contribute to a drop in the overall error rate of the root node.
6. Results and Discussion

In the present study, F-test based power analysis was used to determine the required sample size of the EEG signals to train the classifier. In this study, the test namely, priori sample size computation of multivariate analysis of variance (MANOVA) with repeated measures and within-between interactions has been performed on the data set. The expected power level was set to be 95% (it is equivalent to $\alpha = 0.05$ and $\beta = 0.05$). As a basis for this problem, a data set consisting of 100 samples from each class was considered. The data set was tested for normality, homogeneity and independence in standard error. The central and non-central distributions with critical F value of the result are shown in Fig. 1. While calculating sample size in power analysis, an important parameter namely, Pillai’s $V$ would be found as 0.3396. Power analysis was applied and sample size has been found for various set of confidence and power level. It was found that 81 samples are required to train the classifier so as to get good statistical stability with power level (1-$\beta$) of 95% . A series of experiments were carried out to calculate the sample sizes for 90%, 85%, 80%, 75% of power level and the results are tabulated in Table.1. The results of these experiments are presented as curves in a graph (Refer Fig.2). In another set of experiments, sample size was computed for various power levels with $\alpha$ error probability 5% and it was also found to be 16 per class. The experiment was repeated with the $\alpha$ error probabilities 10%, 20%, and so on up to 40%. The values in the Table 1 and the graph in Fig. 2 would enable the one working in this field to fix the sample size for training the classifier. The required total sample size could be viewed as a function of effect size for various $\alpha$ error probabilities (0.05 to 0.4 in steps of 0.05) and power level 1- $\beta$ (0.95 to 0.6 in steps of 0.05). To study the influence of the effect size on the required sample size, a set of experiments were carried out and it is found that effect size increase is inversely proportional to the sample size with increase in $\alpha$ error probability whereas power level increase is directly proportional to the total sample size with increase in effect size. Hence it is clear that effect size is also an important parameter while computing the sample size. Alternatively, one may be interested in knowing the results for $\alpha$ error probability as a function of effect size for various power levels. For a given power level, as the $\alpha$ error probability increases the effect size decreases. For a given $\alpha$ error probability, as power level increases the effect size also increases. It is to be noted that these effects are for a fixed sample size (81 in this case).

Table 1 Power analysis test results with Effect size f(V)= 0.3046038, Number of groups=5, Repetitions=5 with Pilli’s V formula value = 0.339623, Numerator df =16

<table>
<thead>
<tr>
<th>Output parameter</th>
<th>$\alpha=0.05$, $1-\beta=0.95$</th>
<th>$\alpha=0.10$, $1-\beta=0.90$</th>
<th>$\alpha=0.15$, $1-\beta=0.85$</th>
<th>$\alpha=0.20$, $1-\beta=0.80$</th>
<th>$\alpha=0.25$, $1-\beta=0.75$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noncentrality parameter $\lambda$</td>
<td>30.061846</td>
<td>21.8969</td>
<td>17.072159</td>
<td>13.360820</td>
<td>10.391749</td>
</tr>
</tbody>
</table>
The results of power analysis are validated through a decision tree algorithm namely J48 algorithm. The classifier’s job here is two-fold process: feature selection and feature classification. To study the effect of the sample size on classification accuracy, the sample size was constantly reduced with a decrement of 5 samples and the corresponding classification accuracies were noted. The variation in classification accuracy with respect to the sample size is shown in Fig. 3. One can observe that the classification accuracy falls abruptly when the sample size is reduced below 15 per class. This means if one has just 16 samples per class, it is possible to train a classifier however; the objective of the study is one step further – what is the minimum sample size that will have statistical stability. One can notice that the oscillation of the classification accuracy is more when the sample is low and it tends to stabilize as sample size increases. When sample size reaches 100, the oscillation is very small. In classification, the root mean square error and mean absolute error as a function of sample size is plotted and shown in Fig. 4 and Fig. 5 respectively.

One should note that the maximum classification accuracy of J48 algorithm for 100 samples per class data set was found to be 80%. It is implied that the reduction in classification accuracy is not due to variation in the input data. It may be due to lack of ability of the classifier for the given data set. For computing relative change in classification accuracy, 85% was used as reference value. For comparing the results obtained in power analysis and J48 algorithm, from Table 1, a representative value corresponding to 15% of $\alpha$ error probability was taken. As per power analysis result, from Table 1, if one can accommodate 15% of $\alpha$ error probability and willing to accept 85% of power level, then for given data set the minimum required sample size is 9. This means, if 9 samples were used for training the classifier, the maximum $\alpha$ error probability that likely to happen would be 15%. This has to be validated with J48 classifier’s classification results. From Fig.5, it is evident that the mean absolute error is below 15% for cases whose sample size is greater than or equal to 10. The corresponding root mean square error is shown in Fig. 4.

<table>
<thead>
<tr>
<th>Critical F</th>
<th>1.676863</th>
<th>1.503896</th>
<th>1.394382</th>
<th>1.3312269</th>
<th>1.245602</th>
</tr>
</thead>
<tbody>
<tr>
<td>Denominator df</td>
<td>304</td>
<td>216</td>
<td>164</td>
<td>124</td>
<td>92</td>
</tr>
<tr>
<td>Total sample size for 5 classes</td>
<td>81</td>
<td>59</td>
<td>46</td>
<td>36</td>
<td>28</td>
</tr>
<tr>
<td>Sample size per classes</td>
<td>$\approx$16</td>
<td>$\approx$12</td>
<td>$\approx$9</td>
<td>$\approx$7</td>
<td>$\approx$6</td>
</tr>
<tr>
<td>Actual power</td>
<td>0.952907</td>
<td>0.903578</td>
<td>0.856716</td>
<td>0.804849</td>
<td>0.751608</td>
</tr>
</tbody>
</table>
Fig. 1 F Test- MANOVA: Repeated measures, within-between interaction: Sample size calculation with Effect size $f(V) = 0.3046038$, $\alpha$ error probability=0.05, Power ($1-\beta$ error probability)=0.95, Number of groups=5, Repetitions=5 with Pilli’s $V$ formula.

Fig. 2 Total sample size as a function of $\alpha$ error probability for various power levels

Fig. 3 Total sample size as a function of classification accuracy
In classification problem, the mean absolute error of the classifier is a measure of type I error ($\alpha$ error probability). Type I error is an error due to misclassification of the classifier. In general, type I error is rejecting a null hypothesis when it is true. $\alpha$ error probability is a measure of type I error in hypothesis testing and hence, the equivalence is obvious. From the above discussion, the results of power analysis are true and the actual error did not exceed the upper bound (15%) found in power analysis. A similar exercise of validating the results at other points also assures the validity of the power analysis test. Thus one can confidently use the sample size suggested by power analysis for machine learning approach to classify EEG signals.

**7. Conclusion**

In the present study, machine learning approach was applied to classify the EEG signals of a data set consisting of 500 EEG signals with five different types. A statistical method for determination of minimum sample size using power analysis has been proposed and the results are validated by using a functional test J48 algorithm. From the results and discussions, it is evident that 16 samples per class are required in order to have a power level of 95% and an $\alpha$ error probability of 5%. For other paired values of power level and $\alpha$ error probability, the required minimum sample sizes were tabulated in Table 1. The effects of $\alpha$ error probability and power level on sample size were also discussed in detail. These graphs would definitely serve as a guideline for researchers working with EEG signals to fix the sample size (machine learning approach).
8. References