Engineering Applications of Artificial Intelligence 34 (2014) 154–167

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

A novel statistical algorithm for multiclass EEG signal classification

Siuly*, Yan Li

Faculty of Health, Engineering and Sciences, University of Southern Queensland, Toowoomba, QLD 4350, Australia

ARTICLE INFO

Article history:
Received 8 January 2014
Received in revised form
28 April 2014
Accepted 12 May 2014
Available online 20 June 2014

Keywords:
Electroencephalogram (EEG)
Optimum allocation
Multiclass least square support vector machine (MLS-SVM)
Multiclass classification

ABSTRACT

This paper presents a new algorithm for the classification of multiclass EEG signals. This algorithm involves applying the optimum allocation technique to select representative samples that reflect an entire database. This research investigates whether the optimum allocation is suitable to extract representative samples depending on their variability within the groups in the input EEG data. It also assesses whether these samples are efficient for the multiclass least square support vector machine (MLS-SVM) to classify EEG signals. The performances of the MLS-SVM with four different output coding approaches: minimum output codes (MOC), error correcting output codes (ECOC), One vs One (1vs1) and One vs All (1vsA), are evaluated with a benchmark epileptic EEG database. To test the consistency, all experiments are repeated ten times with the same classifying parameters in each classification process. The results show very high classification performances for each class, and also confirm the consistency of the proposed method in each repeated experiment. In addition, the performances by the optimum allocation based MLS-SVM method are compared with the four existing reference methods using the same database. The outcomes of this research demonstrate that the optimum allocation is very effective and efficient for extracting the representative patterns from the multiclass EEG data, and the MLS-SVM is also very well fitted with the optimum allocation technique for the EEG classification.

1. Introduction

The design of multiclass electroencephalogram (EEG) signal classification is a very challenging task because of the need to extract representative patterns from multidimensional time series generated from EEG measurements. EEG signals indicate the electrical activity of the brain, that is highly random in nature and contain useful information about the brain state to study brain function and neurological disorders (Niedermeyer and Lopes da Silva, 2005). Measuring brain activities through EEG leads to the acquisition of a huge amount of data. Visual inspection for discriminating the EEGs is a time consuming, error prone, costly process, and is not sufficient enough for reliable information. Hence, it is necessary to develop automatic EEG classification methods to ensure a proper evaluation and treatment of neurological diseases (Agarwal et al., 1998). In this paper, we focus on the classification of multiclass EEG signals as most of the EEG recordings are multi-categories.

As EEG recordings consist of a large amount of data, one key problem is how to represent the recorded EEG signals for further analysis, such as classification. Classification is one of the most frequently encountered decision making tasks in machine learning to assign a predefined class to each instance (segment of input data) (Duda et al., 2001; Brunelli, 2009). A classification problem occurs when an instance needs to be assigned into a predefined group or class, based on a number of observed attributes related to that instance. For classification, it is firstly important to extract useful representative instances from raw data, and then utilize those instances for classification. In the past a few years, many researchers have tried to apply different techniques, such as wavelet transform, eigenvector methods and Lyapunov exponents, to pull out a smaller number of values from the EEG data, which describe the key properties of the signals called 'features' (Lotte et al., 2007). The features significantly affect the accuracy of classifying EEG signals. Sometimes the feature values cannot represent all important information about the original signals and that's the reasons why the classification success rates can be limited. Again, none of the prior studies consider the variability of observations within a time window although it is an important issue to get consistent information about the signals for each time window. The majority of the existing methods cannot appropriately handle a large amount of EEG data and in addition, some existing methods of the feature extraction are not right choice for getting representative features from the original EEG data due to its non-stationary characteristic (e.g. Fourier transformation). Considering the mentioned issues, this paper proposes an optimum

* Corresponding author.
E-mail addresses: siuly@usq.edu.au, siuly_1976@yahoo.com (Siuly), liyan@usq.edu.au (Y. Li).

http://dx.doi.org/10.1016/j.engappai.2014.05.011
0952-1976/

allocation technique for finding representative samples, which can efficiently describe EEG signals based on the variability of their observations.

There are reasons of choosing the optimum allocation to extract representative sample points from the EEG data in this study. An optimum allocation method (Islam, 2007; Cochran, 1977) is used in sampling to allocate numbers of sample units into different groups with a minimum variation, providing the most precision with the least cost. This method is an appropriate choice when a population is heterogeneous and very large in size. Then it is required to divide the population into several groups according to their specific characteristics and used to select representative samples from the groups, such that those samples reflect the entire population. As EEG recordings normally include a huge amount of data and the data is heterogeneous with respect to a time period, this study uses the optimum allocation for getting representative samples from each group of the EEG data. 'Representative sample' means the sample is selected randomly from the population and each observation of the population has a known, non-zero chance of being selected in the sample. An effective sample of a population represents an appropriate extraction of the useful data which provides meaningful knowledge of the important aspects of the population. In this study, the EEG signals of a specific category are considered as a class (e.g. class 1 (Set A): health persons with eyes open; class 2 (Set B): health persons with eyes closed etc.). Here the whole EEG data of a class is considered as a population where a sample is considered as a representative part of the population that means a sample is a set of observations from a parent population. An observation in a sample is called sample unit, and the sample size is the number of observations that are included in a sample.

As it is not known how many sample units are required to describe the whole population, it is necessary to determine the sample size first. We determine the required size of samples to describe the characteristics of the whole EEG dataset for each class. The required sample size is determined by using a survey software called '*Sample size calculator*'. It is available publically from, http://www.surveysystem.com/sscalc.htm. Table 1 presents an example of the required sample size for the population with different sizes using the *Sample size calculator*. This table shows that the increment of the sample size is not with the same rate as the population size. It is seen from Table 1 that the sample size can reach to a maximum of 9604 at 95% confidence level and 16,641 at 99% confidence level when the population size is large as possible. This paper proposes a sampling idea based on the optimum allocation technique to obtain representative sample units from any size of a population data.

Based on the proposed methodology, after selecting the required samples from the EEG data of a class, the data is divided into non-overlapping groups with respect to a specific time period, where each group is called 'epoch' described in detail in Section 3.2. Using the optimum allocation process, the sample size of an epoch is determined depending on the variability of observations within that epoch, discussed in Section 3.3. In a class, the sum of all sample sizes for all epochs is equal to the obtained sample size of the class. Thus the samples are selected from each and every epoch of a class and make a vector set that is fed to a classification algorithm.

In recent years, many classification methods have been proposed, such as various types of neural networks (Subasi et al., 2005; Subasi and Ercelebi, 2005; Kousarrizi et al., 2009; Jahankhani et al., 2006) and support vector machines (SVMs), either for two-class or multiclass (Chandaka et al., 2009; Siuly et al., 2011a, 2011b, 2011c, 2013; Sengpur, 2009; Musselman and Djurdjanovic, 2012; Siuly and Li, 2012) EEG signal classification. The SVM is recognized as the most successful classifying model, owing to resulting in high performances for a wide range of classification and pattern recognition applications. A modified version of the SVM classifier is the least square support vector machine (LS-SVM) proposed by Suykens and Vandewalle (1999a). Since the LS-SVMs use equality constraints instead of inequality constraints and solve linear equations instead of the quadratic programming, the computational cost of the LS-SVM is reduced (Xing et al., 2008; Suykens et al., 2002). In addition, it has the advantage over other techniques that converge to a global optimum, not to a local optimum that depends on the initialization or parameters affecting the rate of convergence. The computation of the LS-SVM is faster compared with other machine learning techniques because there are fewer random parameters and only the support vectors are used in the generalization process. Hence this study is intended to apply the LS-SVM as a classification tool for classifying multiclass EEG data from the view point of efficiency, computational cost and classification performance.

A straightforward extension of LS-SVMs to multiclass problems has been proposed by Suykens and Vandewalle (1999a). A common way of solving the multiclass categorization problem is to reformulate the problem into a set of binary classification problems using different output coding schemes. There exist different output coding approaches; for example, minimum output codes (MOC), error correcting output codes (ECOC), One vs One (1vs1) and One vs All (1vsA) to construct a set of binary classifiers. In this research, the multiclass LS-SVM (MLS-SVM) is applied with the radial basis function (RBF) for the classification of EEG signals because the RBF kernel can nonlinearly map samples into a higher dimensional space and the polynomial kernel has more hyper parameters than the RBF kernel, and the sigmoid kernel is not valid under some parameters.

In this paper, we select the hyper parameters of the MLS-SVMs with the RBF after an extensive experimental evaluation before the classification process. We report the performances of the MLS-SVMs for each of the four output coding approaches, using the epileptic EEG benchmark dataset (EEG Time Series, 2005). To test the consistency, we repeat each experiment ten times with the same set of parameter values in each output coding approach. In order to further evaluate the performances, we compare our proposed algorithm with other four existing well-known algorithms. The experimental results show that the proposed algorithm can produce high classification performances for each class and outperforms the existing methods in terms of sensitivity, specificity and the total classification accuracy (TCA).

The rest of the paper is organized as follows. In Section 2, a review of the previous research is discussed. Section 3 presents the proposed approach. Section 4 describes the experimental data and implementation procedure. Section 5 discusses the results, and finally, the conclusions are drawn in Section 6.

Table 1
An example of the required sample size for various sizes of population.

Population	Sample size 99–100% confidence interval	
	95% confidence level	99% confidence level
500	475	485
1000	906	943
5000	3288	3845
10,000	4899	6247
50,000	8057	12,486
100,000	9604	16,641
50,00000	9604	16,641
100,00000	9604	16,641