Sharif University of Technology

Electrical Engineering department

Introduction to Machine Learning

Instructor: professor Hoda Mohammadzahed

Producer: Kasra Fallah

Jan 2022

Contents

Feature extraction

Time domain features:

Frist bunch of our features our time domain features that are useful in our problems, when we surf the related articles about EEG processing we face with lot of time domain features that have many strong source of back ground in "***EEG signal classification using PCA, ICA, LDA and support vector machines***" published in 2016 on IEEE and also from "***A comparative study on classification of sleep stage based on EEG signals using feature selection and classification algorithms***" published in 2018 on IEEE they give us better view for this problems but obviously all of them confirm that there is no grantee to have an accurate answer for this problem for any EEG signal.

1. Mean: we can easily calculate it with Numpy ready functions and it should be useful because many articles named it as the most important features

2. Variance: Also we can easily calculate it with Numpy ready functions, there is no need to say about the important of this parameter because our problem is based on statistics and the most important thing in statistics is VAR!

3. Original Signal: surprisingly it is one of the best parameters that we can use I will talk about its way of working in the conclusion and results more but there are lots of articles that supports the important of this parameter

4. Correlation: Calculated in our features for each person. We have

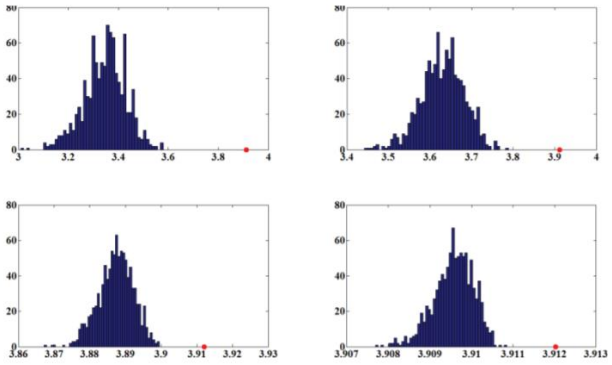
   $$8 * 7/2 = 28$$

   data for pairs of channels.

5. Entropy: Information entropy is the average rate at which information is produced by a stochastic source of data. The measure of information entropy associated with each possible data value is the negative logarithm of the probability mass function for the value:

   $$S = -X\ i\ Pi\ log\ Pi$$

   I apply entropy using the histogram function and then using the normalization method for making the results of our entropy better



I use the article named "Analysing Bin-width Effect on the Computed Entropy" for the reference of my mathematical works but obviously I check all of the ready function to work aligned with the used theories in our problem.

Entropy is a measure of the unpredictability of the state, or equivalently, of its average information content. To get an intuitive understanding of these terms, consider the example of a political poll. Usually, such polls happen because the outcome of the poll is not already known. In other words, the outcome of the poll is relatively unpredictable, and actually performing the poll and learning the results gives some new information; these are just different ways of saying that the a priori entropy of the poll results is large. Now, consider the case that the same poll is performed a second time shortly after the first poll. Since the result of the first poll is already known, the outcome of the second poll can be predicted well and the results should not contain much new information; in this case the a priori entropy of the second poll result is small relative to that of the first. Also many publications suggest this parameter as a critical point in EEG processing such as "***A comparative analysis of signal processing and classification methods for different applications based on EEG signals***" and another piece of art article by Jan Shaw published in 2019 named ***"A new feature extraction and classification mechanisms for EEG signal processing"*** this article are the main reason I use this parameter in my feature extraction.

6. phase-locking value (PLV): Phase-locking value (PLV) is a well-known feature in sensorimotor rhythm (SMR) based BCI. Zero-phase PLV has not been explored because it is generally regarded as the result of volume conduction. Because spatial filters are often used to enhance the amplitude (square root of band power (BP)) feature and attenuate volume conduction, they are frequently applied as pre-processing methods when computing PLV. However, the effects of spatial filtering on PLV are ambiguous. Therefore, this article aims to explore whether zero-phase PLV is meaningful and how this is influenced by spatial filtering. Based on archival EEG data of left and right hand movement tasks for 32 subjects, we compared BP and PLV feature using data with and without pre-processing by a large Laplacian. Results showed that using ear-referenced data, zero-phase PLV provided unique information independent of BP for task prediction which was not explained by volume conduction and was significantly decreased when a large Laplacian was applied. In other words, the large Laplacian eliminated the useful information in zero-phase PLV for task prediction suggesting that it contains effects of both amplitude and phase. Therefore, zero-phase PLV may have functional significance beyond volume conduction. The interpretation of spatial filtering may be complicated by effects of phase. I use two article as a reference for our work one of them is an amazing article published in IEEE named "***EEG based zero-phase phase-locking value (PLV) and effects of spatial filtering during actual movement***"

7. The mean of the absolute values of the first difference of the signal (MAVFDS): I find this feature from article "**Ranking of EEG Time-domain Features on the Negative Emotions Recognition Task**" published by IEEE in 2016 but this feature did not work well on our data set and the main reason I found was the selected channels that we worked on for our project.

Frequency domain features:

As the proposal mentioned the frequency domain features are really important especially for analysing the 8 passive channels that are responsible for distinguishing the differences. I read some article about this features that are piece of art articles such as "***A comparative analysis of signal processing and classification methods for different applications based on EEG signals***" that was focused on many types of this features and analyse their power to help us this process.

1. FFT: A fast Fourier transform (FFT) is an algorithm that computes the discrete Fourier transform (DFT) of a sequence, or its inverse (IDFT). Fourier analysis converts a signal from its original domain (often time or space) to a representation in the frequency domain and vice versa. The DFT is obtained by decomposing a sequence of values into components of different frequencies. This operation is useful in many fields, but computing it directly from the definition is often too slow to be practical. This features are not supported by many great article but one of the article that claimed they used this feature practically is **"Artificial neural network based epileptic detection using time-domain and frequency-domain features"**
2. Mean: **"EEG feature extraction based on wavelet packet decomposition for brain computer interface"** the only article I found that talking about this feature was this one that is not obviously a great article but in some samples the J score of this feature was not bad and we use that.

3. Median: This parameter is one of the most well-known parameters for EEG signal processing but in our problem it was not a game changing feature I tried to found the reason but I could not find anything except the factor that the 8 channel that we use are not responsible for the distinguish part of brain.

4. band power: One of the most important features of a signal is its energy that we calculated separately for each bond of signal.

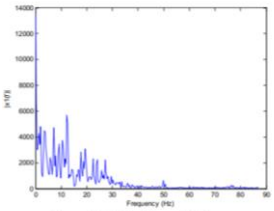
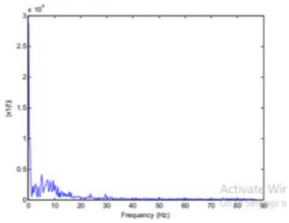


Figure 5(a) FFT of Normal EEG



As we can easily realize from the FFT, we see the different rang of energy that can help us to distinguish the signals but there are not main references that support this feature. In our samples they worked in some samples and they were helpful.

5. DCT : Like any Fourier-related transform, discrete cosine transforms (DCTs) express a function or a signal in terms of a sum of sinusoids with different frequencies and amplitudes. Like the discrete Fourier transform (DFT), a DCT operates on a function at a finite number of discrete data points. The obvious distinction between a DCT and a DFT is that the former uses only cosine functions, while the latter uses both cosines and sines (in the form of complex exponentials). However, this visible difference is merely a consequence of a deeper distinction: a DCT implies different boundary conditions from the DFT or other related transforms.
6. Modified discrete cosine transform (MDCT): he modified discrete cosine transform (MDCT) is a transform based on the type-IV discrete cosine transform (DCT-IV), with the additional property of being lapped: it is designed to be performed on consecutive blocks of a larger dataset, where subsequent blocks are overlapped so that the last half of one block coincides with the first half of the next block. This overlapping, in addition to the energy-compaction qualities of the DCT, makes the MDCT especially attractive for signal compression applications, since it helps to avoid artefacts stemming from the block boundaries. As a result of these advantages, the MDCT is the most widely used lossy compression technique in audio data compression. In the article named "**Using frequency-domain features for the generalization of EEG error-related potentials among different tasks**" published in EPFL they claimed that it is helpful for analysis but in our case it was not really game changing.

7. DST: Relative to DCT, this transforms express a function or signal in terms of a sum of sinusoids with different frequencies and amplitudes. their most important difference is that they 9 are imaginary part of DFTs but DCT is real part of DTF. A DCT is roughly equivalent to a DFT of a vector after it is doubled by mirroring by a symmetric reflection. This produces FFT input that does not have a discontinuity either in the middle or circularly. A DST is roughly equivalent to a DFT after an antisymmetric mirrored extension. This anti□symmetric addition can easily result in discontinuities both in the middle and around the circle. Discontinuities are rep□resented by energy in the high frequency bins in the FFT results. These high frequency artifacts are usually undesirable when using a transform for compression. Since the DCT does not have this potential high frequency content due to circular discontinuities (as does a DST or FFT), the same total energy is thus spread lower in frequency, which potentially allows for greater compression of the high frequency DCT bins, while remaining below some visible threshold.

Time-Frequency features:

In recent years, estimation of human emotions from Electroencephalogram (EEG) signals plays a vital role on developing intellectual Brain Computer Interface (BCI) devices. In this work, we have collected the EEG signals using 64 channels from 20 subjects in the age group of 21~39 years for determining discrete emotions (happy, surprise, fear, disgust, and neutral) under audio-visual induction (video/film clips) stimuli. Surface Laplacian filtering is used to pre-process the EEG signals and decomposed into five different EEG frequency bands (delta, theta, alpha, beta, and gamma) using Wavelet Transform (WT). The statistical features are derived from all these five frequency bands are considered for classifying the emotions using two linear classifiers (K Nearest Neighbour (KNN) & Linear Discriminant Analysis (LDA)). The main objective of this work is to consider a selected number of 24 channels for assessing emotions from the original EEG channels. There are three different wavelet functions ("db8", "sym8", and "coif5") are used to derive the linear and non-linear

features for emotion classification. There are many articles that support the important of time-frequency domain features and the most game changing one is "**Comparison of different wavelet features from EEG signals for classifying human emotions**" that published in 2015 and was a revolution in EEG signal processing.

1. Wavelet: in fact, the Fourier transform can be viewed as a special case of the continuous wavelet transform with the choice of the mother wavelet is

$$\psi(t) = e^{-2\pi it} .$$

The main difference in general is that wavelets are localized in both time and frequency whereas the standard Fourier transform is only localized in frequency. The Short-time Fourier transform (STFT) is similar to the wavelet transform, in that it is also time and frequency localized, but there are issues with the frequency/time resolution trade-off. However, consider a non-continuous signal with an abrupt discontinuity; this signal can still be represented as a sum of sinusoids, but requires an infinite number, which is an observation known as Gibbs phenomenon. This, then, requires an infinite number of Fourier coefficients, which is not practical for many applications, such as compression. Wavelets are more useful for describing these signals with discontinuities because of their time-localized behaviour (both Fourier and wavelet transforms are frequency-localized, but wavelets 10 have an additional time-localization property). Because of this, many types of signals in practice may be non-sparse in the Fourier domain, but very sparse in the wavelet domain. This is particularly useful in signal reconstruction, especially in the recently popular field of compressed sensing. (Note that the short-time Fourier transform (STFT) is also localized in time and frequency, but there are often problems with the frequency-time resolution trade-off. Wavelets are better signal representations because of multi resolution analysis.)

2. STFT: STFTs as well as standard Fourier transforms and other tools are frequently used to analyse music. The spectrogram can, for example, show frequency on the horizontal axis, with the lowest frequencies at left, and the highest at the right. The height of each bar (augmented by colour) represents the amplitude of the frequencies within that band. The depth dimension represents time, where each new bar was a separate distinct transform. Audio engineers use this kind of visual to gain information about an audio sample, for example, to locate the frequencies of specific noises (especially when used with greater frequency resolution) or to find frequencies which may be more or less resonant in the space where the signal was recorded. This information can be used for equalization or tuning other audio effects.

Feature selection

Over fitting

In this case, when we have a lot of features, (for example more than our size of test), then our net will over fit and for example for each output, our net learns to specify a feature so that our accuracy will be about one hundred percent for our train data but we will not get appropriate answer for our test data. All this makes us to think about using methods such as J fisher score for choosing the features.

Implementing J score for different methods of classification

For deciding threshold, we move the threshold on J for deciding features, then learn the net and evaluate its accuracy using part of train that we selected randomly for evaluating accuracy of network. I also use plotting the accuracy using the K-fold cross validation method for all of the methods of classification and then I chose how many of the features are important for our problem and method. I used the J score for finding the best number of features, after that I use a kind of cross validation procedure to find the best number of the parameter that gives use the best accuracy. Remember that it is important to check the validity of both types of data separately. The main reason is that there are different number of 1 and 0 in the data set and of the best approaches to solve this problem is using class_weight for your learning approach. In the following table I write the number of the features that is the best choose for each method. It was a time-taking action to find all of these feature vales but I found them and they are available in the following table.

#person	Logistic regression	SVM–linear-2	SVM-linear-3	SVM-poly -2	forest
1	29	51	48	73	110
2	38	43	35	80	123
3	45	73	54	120	42
4	32	24	35	34	43
5	36	58	65	17	87
6	40	28	34	45	45
7	61	65	45	65	85
8	40	24	48	46	45
9	32	56	39	37	79

Number of features for each approach

I chose the best possible accuracy for each reason between the different methods and different persons. As we know the standard approach for this problems are the SVM with linear kernel. Many famous articles included "**EEG signal classification using wavelet feature extraction and a mixture of expert model**" named this approach as the best method. Also the others articles mainly use logistic regression.

Accuracy of the results

First of all, I want to talk about the meaning of accuracy in our question. It is so important to mention that our data is totaly unbiased and this fact make us to have many problems in this classification. The meaning of the accuracy in these types of problems can be defined in two ways. *1. Using confusion matrix 2. Using the balanced* accuracy of data. I used the second method for the total accuracy. The accurcy that is shown below is the form the best possible method with the best the best number of features for that method.

#Person	The average accuracy of Best approach
1	(SVM-3) 69.4%
2	(LR) 75.6%
3	(SVM-3) 71.6%
4	(SVM-3) 73.2%
5	(SVM-2) 68.4%
6	(SVM-3) 73.9%
7	(RL) 69.4%
8	(RL) 73.7%
9	(SVM-3) 76.5%

Predictions

In this part I want to show the result of the test dataset and the predicted words. Unfortunately, I could not find all of the word in the best way but these are the best results that I get from this code.

#Person	Predicted Word	My Prediction based on result
1	L05Ay	LUKAS
2	LUKAS	LUKAS
3	LUKA5	LUKAS
4	LUK4S	LUKAS
5	WAT83	WATER
6	HV4NT	HUANT
7	WSDDF	?????
8	WATER	WATER
9	WATER	WATER

conclusion

The thing that was really impressive for me was the changed results in the predicted word for example I found out that many time A was replaced by 4 that are really similar and also S was replaced by 5.

$$A -> 4$$

$$S -> 5$$

$$R -> 3$$

$$U -> V$$

Another valuable result that we found was the importance of wavelet features in our detection and this was all happened in my project, and these features helped me to improve the results and somehow make better predictions.

Thanks for your consideration.