Exploratory Data Analysis Of The Gesture Phase Segmentation Dataset

BAKHTI Yassine   EL BADAOUI Khalil   BENFATTOUM Ilyas

I. Introduction

In this report we present our exploratory data analysis of our dataset. The dataset is composed by features extracted from 7 videos with people gesticulating, aiming at studying Gesture Phase Segmentation.

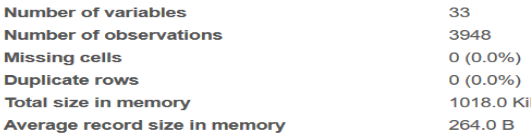
The different phases to predict in our case are S (stroke) and H (hold).

The rest of this report is organized as follows. Section2 summarizes the main characteristics of our dataset. Features description and multivariate analysis are provided in section3. The problem of imbalanced dataset is solved in section4. Normalization and PCA analysis is detailed in section5 and section6. Finally, section7 and section8 provide some machine learning test techniques and a description of our chosen algorithms.

II. Dataset Description

Figure 1 illustrates the analysis of our dataset especially : the number of variables, observations and missing values which we can conclude that it is complete.

Moreover, we analysed the number of duplicated data rows and the total size in memory.



Number of variables	33
Number of observations	3948
Missing cells	0 (0.0%)
Duplicate rows	0 (0.0%)
Total size in memory	1018.0 KiB
Average record size in memory	264.0 B

Fig. 1. Useful information about our dataset.

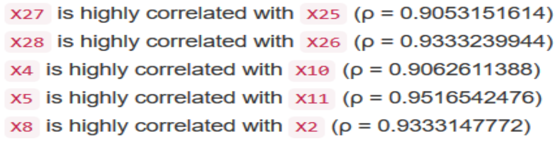
III. Multivariate Analysis and Features Selection

Correlation analysis is a statistical assessment method used to study the strength of the relationship between two, numerically measured, continuous variables.

If correlation is found between two variables it means that when there is a systematic change in one variable, there is also a systematic change in the other; the variables alter together over a certain period of time.

Pearson's product-moment coefficient ρ is the measurement of correlation and ranges in our case between +1 and -1. +1 indicates the strongest positive correlation possible, and -1 indicates the strongest negative correlation possible and 0 indicates no correlation.

In our case we defined a threshold of 0.9. Therefore, any variables with a $|\rho| \geq 0.9$ are considered highly correlated as shown in figure 2.



X27 is highly correlated with X25 (ρ = 0.9053151614)
X28 is highly correlated with X26 (ρ = 0.9333239944)
X4 is highly correlated with X10 (ρ = 0.9062611388)
X5 is highly correlated with X11 (ρ = 0.9516542476)
X8 is highly correlated with X2 (ρ = 0.9333147772)

Fig. 2. Highly correlated features.

IV. Handling Imbalanced Dataset with Clustering

In order to handle the problem of the imbalanced classes "H" for Hold and "S" for Stroke in our case as shown in figure 3 we Under sampled the majority class by replacing a cluster of majority samples by the cluster centroid of a KMeans algorithm.

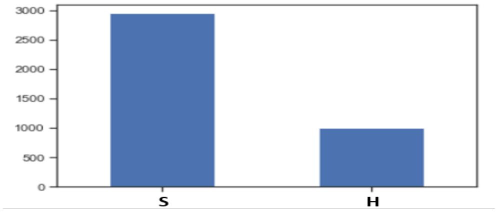


Fig. 3. Histogram of the 2 classes "H" and "S"

This algorithm keeps N majority samples by fitting the KMeans algorithm with N cluster to the majority class and using the coordinates of the N cluster centroids as the new majority samples.

Here we will resample only the majority class to be equal to the minority class which contains 998 samples.

Therefore we got 998 cluster which means 998 sample in the majority class.

V. Normalization

From a numerical point of view having normalized similar data improves convergence of algorithms and accuracy of calculations. This normalization is done in order to obtain good results while doing the pricipal component analysis. After normalizing our dataset, we applied PCA (detailed in the next section) and took the most significant variables, this has given better results in contrast to the case where we consider that the dataset is initially normalized.

VI. Principal Component Analysis

After reducing our dataset from 32 variables to only 27 using correlation property we will try to reduce more this number of variables in order to speed up the training of our models using dimensionlity reduction technique called Principal Component Analysis (PCA).

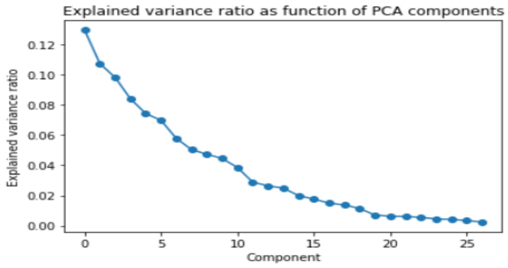


Fig. 4. Explained variance ratio as function of PCA components

Feature elimination using PCA is what we are looking for: we want to reduce the feature space by eliminating some variables. In our dataset, instead of considering every single variable, we might drop some of them based on the explained variance ratio of each variable using the elbow method as shown in figure 4.

In our case we took 17 variables which explain 93% of our dataset information.

Advantages of feature elimination methods include simplicity and maintaining interpretability of our variables.

As a disadvantage, though, we gain no information from those variables we have dropped. Thus, we have entirely eliminated any benefits those dropped variables would bring.

VII. Machine Learning Algorithms Evaluation

The adopted method for evaluating our model's performance is k-fold cross-validation, where the original dataset is partitioned into k equal size subsamples, called folds. This is repeated k times, such that each time, one of the k subsets is used as the test set/validation set and the other k-1 subsets are put together to form a training set. The following figure models the k-fold cross-validation.

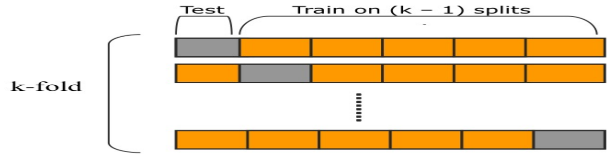


Fig. 5. k-fold cross-validation

The model evaluation metrics are required to quantify model performance, we've used tree metrics:

- Accuracy: refers to closeness of the measurements to a specific value.
- Precision: refers to the closeness of the measurements to each other.
- Loss: function used to optimize a machine learning algorithm.

VIII. Our chosen Algorithms

A. Support Vector Machine (SVM)

SVM algorithms use a set of mathematical functions that are defined as the kernel. The function of kernel is to take data as input and transform it into the required form. Different SVM algorithms use different types of kernel functions. These functions can be different types. For example linear, nonlinear, polynomial, radial basis function (RBF), and sigmoid. The most used type of kernel function is RBF. Because it has localized and finite response along the entire x-axis.

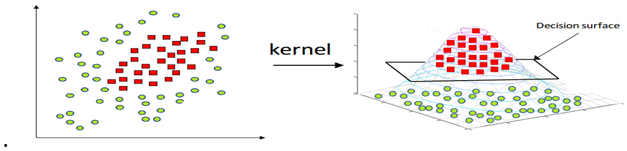


Fig. 6. kernel trick to separate a non linearilly separable data

B. Random Forest (RF)

Random forest is a classification algorithm consisting of a set of decisions trees. It uses bagging and randomness when constructing each individual tree to try to create an uncorrelated forest of trees, whose prediction by the committee is more accurate than that of any individual tree.

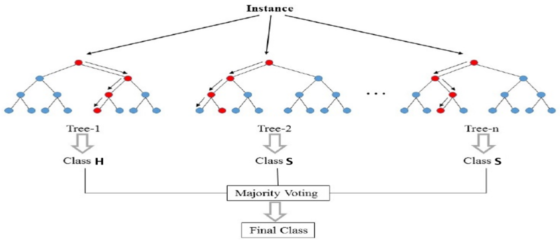


Fig. 7. Random Forest Making a Prediction on our classification problem

C. Artificial Neural Network (ANN)

Neural networks are a set of algorithms, modeled loosely after the human brain, that are designed to recognize patterns. A neural network is described as a network of layers (input, hidden and output layers), each layer consists of several nodes. We start with an input layer in which each node corresponds to an explanatory variable. Then one or more hidden node layers and finally an output node layer that corresponds to the predicted variable. The synapses that connect each of these nodes to the nodes of the next layer are weighted. The learning phase of the network will make it possible to optimize these weights so that the predicted value is as close as possible to the value to be predicted.

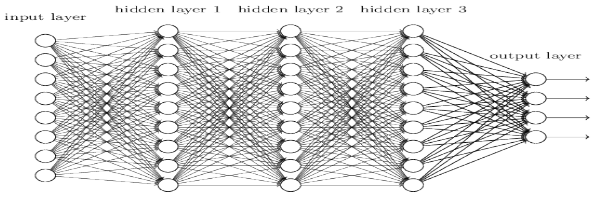


Fig. 8. Artificial Neural Network