Alexandre Benayoun

Case Study

1. Data analysis

This case study is a classification problem. I defined the classes as:
0: ‘dribble’
1: ‘tackle’
2: ‘no action’
3: ‘shot’
4: ‘pass’
5: ‘cross’
6: ‘run’
7: ‘walk’
8: ‘rest’

First, I investigated the data I had. I calculated a few mathematical properties of the norms. The following table gathers the **number of gaits, their length, average value, maximum, minimum and standard deviation** for each class, coming from match_1.

Class	0	1	2	3	4	5	6	7	8
number	42	20	0	5	23	3	207	243	21
length	41.6	44.2	-	32.6	41.7	60.0	38.7	54.8	67.4
average	52.9	70.7	-	123	63.8	46.7	54.6	28.1	22.0
maximum	203	390	-	501	257	302	203	76	41
minimum	13.1	10.8	-	27.9	16.4	8.6	13.9	13.0	15.0
std	36.1	73.2	-	102	53.1	48.1	38.2	11.9	4.00

The classes are not balanced as some have a lot more samples than others. Indeed, classes 6 and 7 have more than 200 samples, while the others have less than 50. This could be an issue when building the model.

We see that the action of rest is usually longer than the others, the maximum value and the standard deviation are very small. On the contrary, a shot is the shortest action and its average, maximum and standard deviation are high. These observations make sense with what we could expect from these respective actions.

I also used the peaks detection method to have more knowledge about the data. With this method, I was able to calculate the **number of peaks in a gait, their heights and widths**. Moreover, I calculated the **skewness and kurtosis** that measure the symmetry of the data distribution and whether the data is heavy-tailed or light-tailed.

However, these metrics only give information about the whole gait, but nothing on the evolution of the values during this gait. To capture that, I divided the gait into 3 parts in chronological order, and calculated the metrics individually on these 3 parts. For example, I calculated the average values of the 3 parts, respectively m1, m2 and m3, and then calculated the gap between the different parts to evaluate the transitions: M21=m2-m1, M32=m3-m2.
From this analysis, I noticed that during a shot or a pass, there usually is an increase of the intensity followed by a decrease. During a tackle, the increase continues throughout the gait. Finally, during dribbles, runs, walks and rests, the intensity is constant.

Then, I had to choose between all these indicators. I used the **Pearson correlation coefficient** to calculate the correlation between these metrics. This coefficient gives a number between -1 and 1, and the more the value is close to 1 (or -1), the more the two parameters are correlated. The following table gathers the Pearson coefficients between 4 metrics (length of the norm, maximum, average and the number of peaks).

	length	maximum	average	peaks
length	-	-0.27	-0.41	0.93
maximum	-0.27	-	0.84	-0.18
average	-0.41	0.84	-	-0.30
peaks	0.93	-0.18	-0.30	-

The number of peaks and the length have a very high coefficient of 0.93. It means that these two parameters are correlated. This is not surprising, as the number of peaks has more chance to be high if the length of the gait is also high. Therefore, they are redundant and would bring the same information to the model.

Another way to evaluate the importance of the features is to use the feature_importance tool from the Random Forest classifier.

2. Methodology

I chose to use a **Random Forest classifier** because this is a good algorithm for classification problems. It is composed of several decision trees that are able to capture nonlinear relationships without requiring a heavy preprocessing work. Besides, as mentioned before, it can give the importance of each feature for the prediction, and therefore be useful for the features selection.

During cross validation, I noticed that the models were predicting almost exclusively the labels ‘run’ and ‘walk’. This is because we have unbalanced classes and these two are the most dominant ones. The solution I propose is to divide the problem into two phases:

- Model A: a classification problem composed of 3 classes: **[‘run’, ‘walk’, ‘other’]**, ‘other’ being the 7 other classes.

- Model B: another classification problem composed of 7 classes: **[‘dribble’, ‘tackle’, ‘no action’, ‘shot’, ‘pass’, ‘cross’, ‘rest’]**

First, model A will predict if the current input is from the class ‘run’, ‘walk’ or another class. And then, if the prediction is ‘other’, model B will predict between the 7 other classes.

The advantage of this approach is that the 3 classes of model A are more balanced. Indeed, the class distribution is: ‘run’: 207, ‘walk’: 243, ‘other’: 114.
Similarly, model B will predict the classes more easily without the overwhelming majority of the ‘run’ and ‘walk’ classes.

However, this method can lead to cumulative errors. Moreover, model B will only be trained on a few samples, as these classes are more rare.

Concerning the features, I chose to use:
(length, max, average, standard deviation, M21, M32).

M21 being the difference between the means of the 2nd and 1st parts of the gait. Similarly for M32 between the 3rd and 2nd parts.

I trained and tested the two models separately. I trained model A with all the data from match_1 and model B with the data that are not ‘run’ nor ‘walk’. I did not use the gaits that are lower than 0.1s and higher than 3s for the training as it does not translate a realistic action. Then, I tested it on the data from match_2.

After that, I tested the two models together,as explained before, on the data from match_2.

3. Other approaches

Other algorithms such as SVM or Neural Networks could be used.

Another interesting approach would be to consider it as a time series problem and use RNN or LSTM models.

This could be a time series inside the gait. Meaning that the evolution of the norm’s values will lead to the prediction of the action. However, we should do some preprocessing as the gaits do not have the same length.

We could also consider a time series that takes multiple gaits, or multiple features such as the mean. This approach would mean that the evolution of the gaits (or the means) from the previous actions will lead to the prediction of the current action. This implies that the current action is influenced by the previous ones.