Support Vector Machines applied to Enterococcus

Clover Biosoft

September 25, 2018

1 Introduction

The aim of this report is to show a different method for clustering Enterococcus Faecium samples that have been previously preprocessed.

In a first report, we used the Principal Component Analysis (PCA) to perform the clustering. PCA is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.

Now, we use the Support Vector Machines (SVM) which is a supervised machine learning algorithm that, based on the samples, calculates a decision function employing a Kernel function. In this report we show the difference between four Kernel functions: Radial Basis Function (RBF), Polynomial, Linear and another linear separation Kernel called LinearSVC, which separates classes linearly but being more precise for a large amount of data and having slightly different parameters to variate than Linear Kernel.

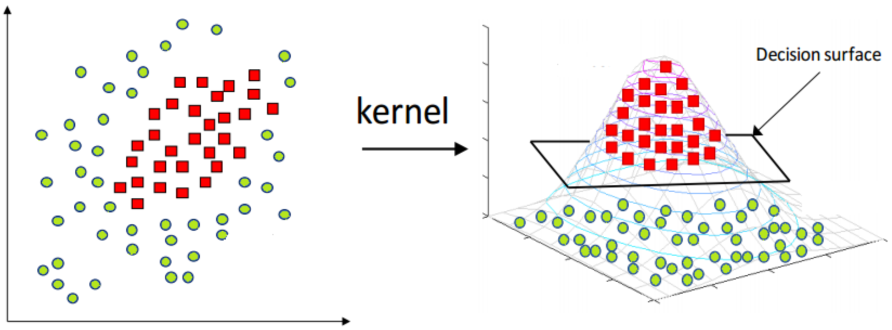

Figure 1: Applying a Kernel in SVM.

2 Clustering: E. F. Sensibles Vs E. F. Van B

The preprocesing methods used are the same than the ones used in the first report.

At each picture, the blue region is the one where all the E. F. Sensibles samples are supposed to be and the red region is the one for the E. F. Van B. Also, blue points are the E. F. Sensibles samples and red points are the E. F. Van B ones.

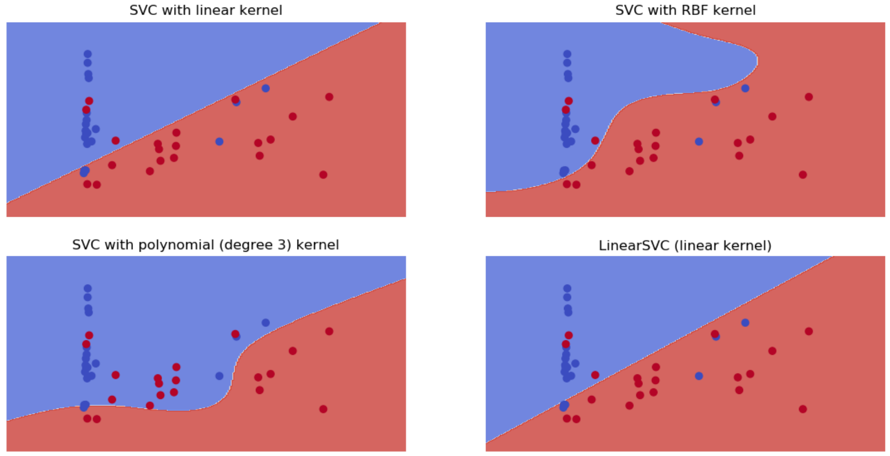

Figure 2: SVM with four different kernels for 40 samples of Enterococcus.

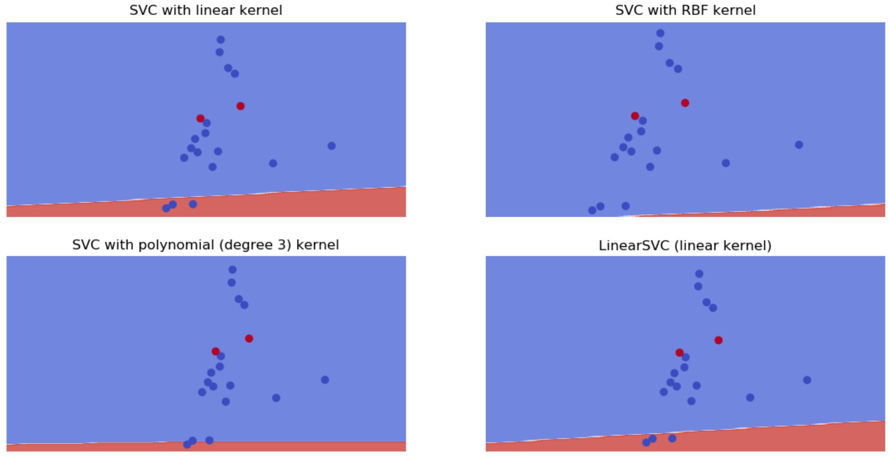

Figure 3: Zoom in the top left corner.

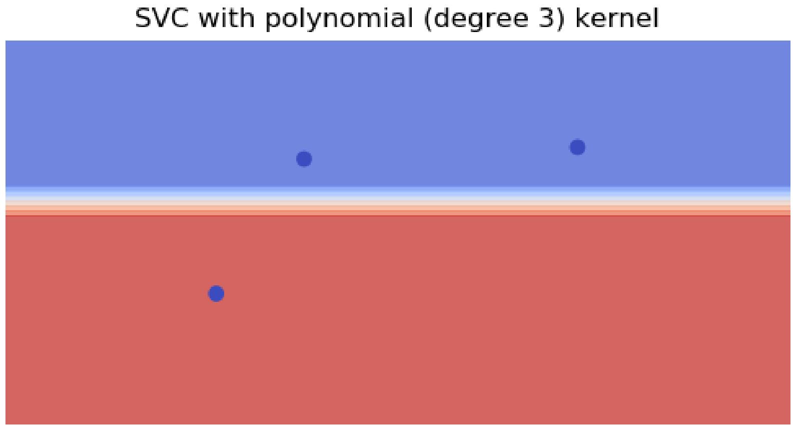

Figure 4: Zoom in for the polynomial border points.

As we can see, in Figure 2 it is shown clearly that SVM-RBF can group better than the other different Kernels. It is also shown that the polynomial case locates the E. F. Sensibles (blue points) properly but loses a lot of precision when it comes to group the E. F. Van B (red points). By zooming in the top-left corner (Figure 3), we have a clear view of all the points. We also see that linear cases do not group E. F. Sensibles (blue points) so good in such area.

2.1 Clustering results

For this case, we have:

- Red dots: E. F. Van B (20 samples)
- Blue dots: E. F. Sensibles (20 samples)

Now we will show how points are located in each region by comparing with the number of total points of each type of sample. Like this we show the percentage of points of each type of sample located in each region for each kernel:

- Red Background Region:
  - Linear
    - 17/20 E. F. Van B (85%).
    - 6/20 E. F. Sensibles (30%).
  - RBF
    - 17/20 E. F. Van B (85%).
    - 3/20 E. F. Sensibles (15%).
  - Polynomial
    - 8/20 E. F. Van B (40%).
    - 1/20 E. F. Sensibles (5%).
  - LinearSVC
    - 17/20 E. F. Van B (85%).
    - 6/20 E. F. Sensibles (30%).
- Blue Background Region:
  - Linear
    - 3/20 E. F. Van B (15%).
    - 14/20 E. F. Sensibles (70%).

- RBF
    - 3/20 E. F. Van B (15%).
    - 17/20 E. F. Sensibles (85%).
- Polynomial
    - 12/20 E. F. Van B (60%).
    - 19/20 E. F. Sensibles (95%).
- LinearSVC
    - 3/20 E. F. Van B (15%).
    - 14/20 E. F. Sensibles (70%).

3 Clustering: E. F. Sensibles Vs E. F. Van B Vs E. F. Van A

In this part 21 E.F. Van A samples have been added to see how the clustering works with this method.

The preprocessing of these samples is different than the one performed at the first example:

- Baseline Subtraction. The baseline component is an intensity offset in the raw spectrum caused by the presence of the matrix. It happens when the spectrum information component overlaps.

- Noise removal. We use the Savitzky Golay filter which gives as a result a smoothed version of the original spectra.

- Peak Alignment. We apply a small tolerance to align peaks among different spectra.

- Peak detection. To obtain the peak matrix with a reduced amount of masses we try different methods such as keeping just the peaks above a certain threshold or obtaining a fixed number of peaks. Between these two methods, the best results are obtained for a fixed value of the most representative peaks. Several values have been tested (10, 40, 80, 100 peaks for each spectrum) with different tolerances, and it is with 40 peaks that the result of the SVM is better, since with a higher value the value of 0 in the matrix (no peaks found) increases, and a lower value of peaks give a too restricted matrix. So we will consider only the 40 most representative peaks of each one of the 61 spectra.

- Comparison of E.F. Sensibles, E.F. Van B and E.F. Van A. Once each of the most representative peaks of each spectrum is obtained, the peak alignment is performed, (according to a tolerance of 5Da) to get a unique peak matrix to perform the comparison.

We apply the algorithm with the same Kernels of the first case. Region and points colored with blue belong to E. F. Sensibles; red to E. F. Van B and white to E. F. Van A.

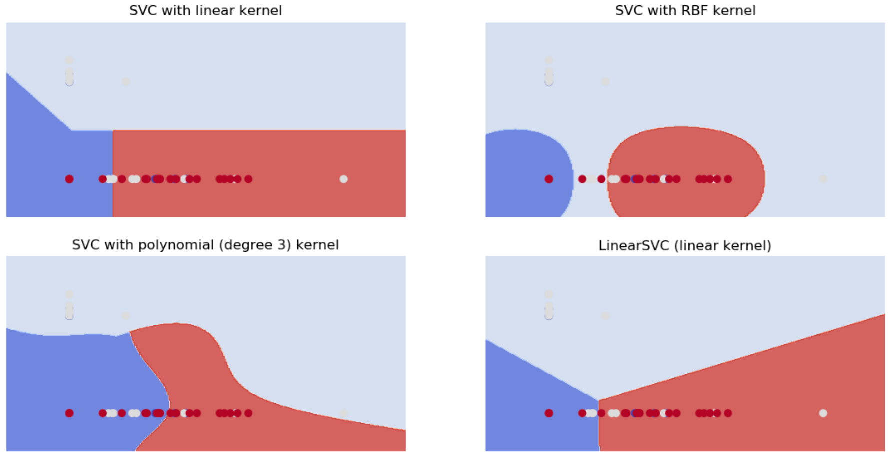

Figure 5: SVM with four different kernels for 61 samples of Enterococcus.

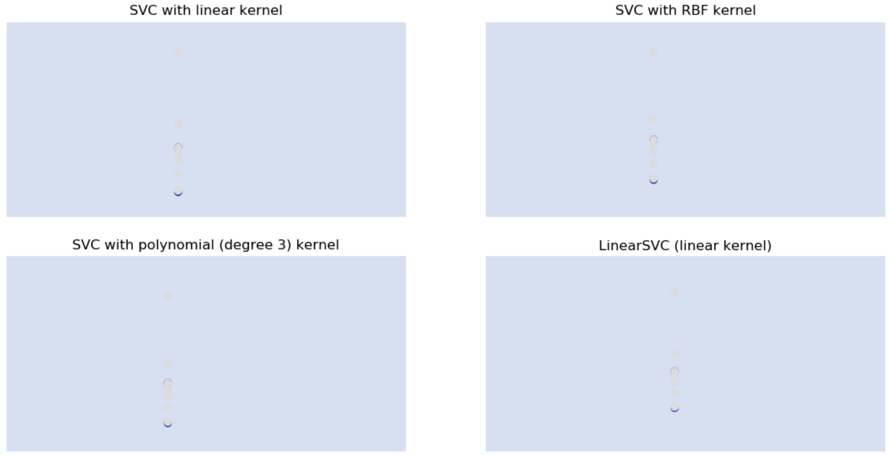

Figure 6: Zoom in the top-left corner of the image.

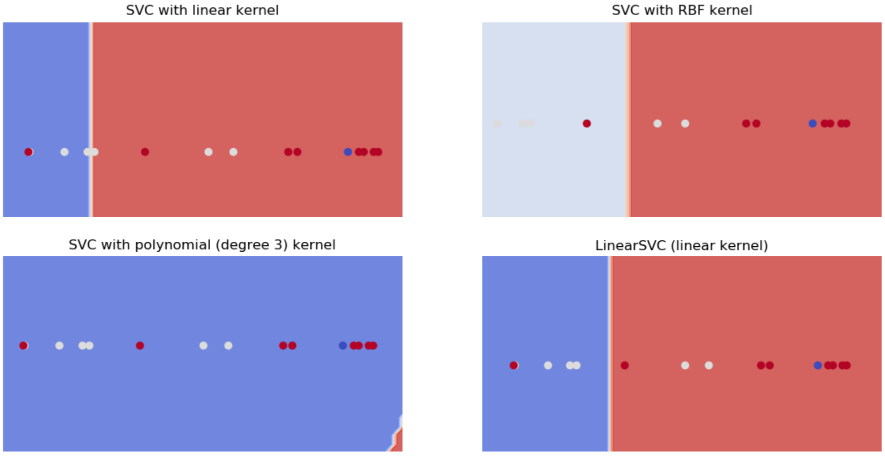

Figure 7: Zoom in the middle down of the image.

The remaining points are overlapping the red point in the botton-left corner.

As we can see in Figure 5, 6 and 7, most of the points looks aligned. It can be explained if we remember how the preprocesing was performed and how the algorithm works: by aligning the spectra considering the 40 most representative peaks of each sample, for certain masses the number of samples where a peak is found is low. This type of preprocessing causes the peak matrix to have a higher number of values equal to 0 since no peaks are found in all the samples for all the masses. When applying the SVM to this type of matrix, the representation of the samples in the different regions tends to be linear, since in our SVM, two masses are selected (by means of a selection algorithm), and we represent the intensities of one of the selected masses against the intensities of the other one, so most of the represented points are located on the axes $x = 0$ and $y = 0$, and at the origin.

We also see that the linear cases are going to group better the E. F. Van B and, knowing that most points of the E.F. Sensible are in the botton left corner of each chart, they will be well grouped for all cases. However, it is clear how RBF is able to provide a better grouping of the entire data set. When zooming in the middle down of the image (Figure 7), apart from seeing the hidden points, we see that there is a lot of variation in the transition zone between classes if we switch between one kernel and another.

3.1 Clustering results

For this case, we have:

- Red dots: E. F. Van B (20 samples)
- White dots: E. F. Van A (21 samples)
- Blue dots: E.F. Sensibles (20 samples)

Let´s see how well located the points are in each group for each kernel:

- Blue Background Region
  - Linear
    - 15/20 E. F. Sensibles (75%)
    - 7/21 E. F. Van A (33%)
    - 4/20 E. F. Van B (20%)

  - RBF
    - 15/20 E. F. Sensibles (75%)
    - 4/21 E. F. Van A (20%)
    - 3/20 E. F. Van B (15%)
  - Polynomial
    - 16/20 E. F. Sensibles (80%)
    - 10/21 E. F. Van A (50%)
    - 11/20 E. F. Van B (55%)
  - LinearSVC
    - 15/20 E. F. Sensibles (75%)
    - 8/21 E. F. Van A (38%)
    - 4/20 E. F. Van B (20%)
- Red Background Region
  - Linear
    - 3/20 E. F. Sensibles (15%)
    - 6/21 E. F. Van A (28%)
    - 16/20 E. F. Van B (80%)
  - RBF
    - 3/20 E. F. Sensibles (15%)
    - 4/21 E. F. Van A (20%)
    - 15/20 E. F. Van B (75%)
  - Polynomial
    - 2/20 E. F. Sensibles (10%)
    - 2/21 E. F. Van A (10%)
    - 9/20 E. F. Van B (45%)
  - LinearSVC
    - 2/20 E. F. Sensibles (10%)
    - 5/21 E. F. Van A (36%)
    - 16/20 E. F. Van B (80%)
- White Background Region
  - Linear
    - 2/20 E. F. Sensibles (10%)
    - 8/21 E. F. Van A (38%)
    - 0/20 E. F. Van B (0%)
  - RBF
    - 2/20 E. F. Sensibles (10%)
    - 13/21 E. F. Van A (62%)
    - 2/20 E. F. Van B (10%)
  - Polynomial
    - 2/20 E. F. Sensibles (10%)
    - 9/21 E. F. Van A (43%)
    - 0/20 E. F. Van B (0%)
  - LinearSVC
    - 3/20 E. F. Sensibles (15%)
    - 8/21 E. F. Van A (38%)
    - 0/20 E. F. Van B (0%)

4 Conclusion

As a conclusion, we can remark that the best Kernel for both cases is the RBF:

- For the two groups case, 85% of E. F. Sensibles and 85% of E. F. Van B are properly located. Comparing with PCA, we find:

	E. F. Sensibles Region		**E. F. Van B Region**	
	E. F. Sensibles	E. F. Van B	E. F. Sensibles	E. F. Van B
PCA	75%	30%	25%	70%
SVM-RBF	85%	15%	15%	85%

- For the three groups case, 75% of E. F. Sensibles, 62% of E.F. Van A and 75% of E.F. Van B are located in their proper region. Unfortunatelly, PCA cannot classify it.

So, for the three group case, we do not get a very precise classification with SVM-RBF but for the two group case, with SVM-RBF there is more percentage of points which are located in their corresponding region, and less percentage of points which are not located in their proper region. So, SVM-RBF performs a more accurate analysis.