Project: DNA Sequence Classification

KERNEL METHODS IN MACHINE LEARNING-AMMI 2020

Francis ANOKYE
Aissatou NDOYE

Monday 1st June, 2020

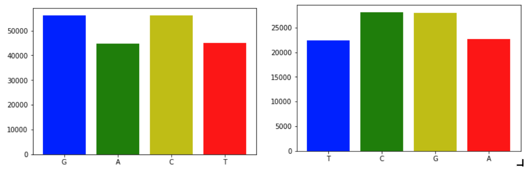
1 Summary

The data challenge is a project taken to climax the kernel methods in machine learning course at AMMI, aimed at the implementation of machine learning algorithms to gain understanding and further adapt them to structural data (DNA sequence data). In this report, we present our approach to the challenge which was hosted on Kaggle with the goal of predicting whether a DNA sequence region is binding site to a specific transcription factor or not. Our best result ranked 3rd on the public leader board with a score of 69.80% at the time of writing this report.

2 Introduction

According to the National Human Genome Research Institute (genome.gov), sequencing of DNA simply means to determine the order of the four chemical building blocks called ”bases” that make up the DNA molecule. These four chemical bases always bond with the same partner to form ”base pairs”. Adenine (A) always pairs with thymine (T); cytosine (C) always pairs with guanine (G).

The challenge provided a balanced data set of 2000 training sequences with their labels indicating bound (1) or not (0) and 1000 test sequences for prediction and submission to the Kaggle competition. Additional numeric form of the data were provided which consisted of the feature matrices calculated respectively from the train sequences and test sequences based on bag of words (BoW) representation. Figure 1 illustrates the representation of the sequence in our data.



(a) Training Data (b) Testing Data

Figure 1: Distribution of each character.

Our primary objective was to predict whether a DNA sequence region is binding site to a specific transcription factor or not. Inasmuch as several simple strategies were tried at the beginning which made us gain better understanding of the problem at hand, we only report on the best approach and algorithms due to our constraint on space.

3 Methods

In our quest for finding the right preprocessing and algorithms that met the needs of our task, a couple of these methods were attempted starting from the use of a simple logistic regression model which proved to be not too useful for the nature of the data as visualized with 3D-t-SNE in Figure 2 and the given size of data, hence, we resorted to kernel methods based on the inspiration from our lecture notes, tutorial sessions and reports from past data challenges.

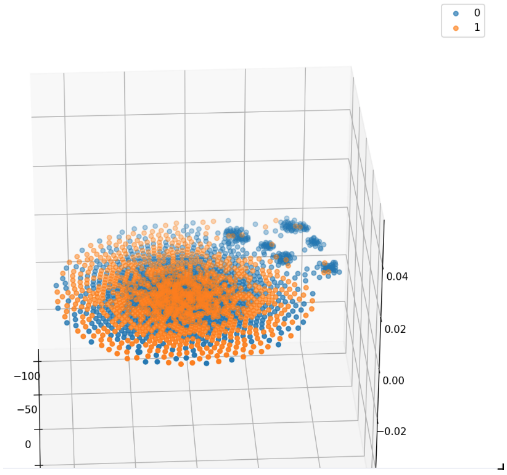


Figure 2: TSNE 3D visualization of the preprocessed data

More so, literature supports the use of support vector machine (SVM) with kernels as one of the most viable options for binary classification on sequential data as observed in our case. The k-spectrum kernel which relates to all possible substrings of length k that are contained in a sequence mostly referred to as k-mers with its variants were tested with kernel SVM on the numerical data representation which achieved accuracy score not beyond 66% and also directly on the given sequences. In our use of the kernel SVM with the sequence data, input sequences are mapped into a high-dimension vector space where the feature values generate the coordinates, then the SVM finds a linear decision boundary in the new space and tests weather the sequences are in the positive or negative side of the boundary. The features that we used in the spectrum kernel are the set of all possible sub-sequences of a fixed length k. If two sequences contain many of the k-length subsequences, their ”inner product” is then computed by the k-spectrum kernel. We optimized the weights in our prediction task by linearly combining three k-spectrum kernels to create multiple kernels with k values of 12,13 and 15 creating mismatch kernels respectively ie MM(k,m). Multiple spectrum allows us to have information about sub-strings of different lengths for a given sequence which is more useful than having information about sub-strings of a single length by combining several k-spectrum kernels.

4 Conclusion

The performance of the kernel SVM with multiple spectrum on our dataset in this challenge presents the classifier as a promising algorithm that is effective in predicting whether a DNA sequence region is binding site to a specific transcription factor or not. Our final result was an accuracy of 69.80% in the public leaderboard. It is possible that we could have achieved a better score if we had performed thorough hyper-parameter search on a wide range of k. We look forward to this in the future even after the challenge.