Kernel Methods : DNA sequence classification

Hadia Mohmmed Osman Ahmed Samil - Domguia Joseph Marie
AIMS AMMI

Abstract

The goal of this work is to use kernel algorithm to classify is a DNA is bound or not with a transcription factor. We apply standard kernel (quadratic, linear, ...) and implement some specific for sequential data (spectrum, mismatch). The results show that the good choice of the kernel is very important to get a good performance.

Keywords: kernel, DNA, classification

Introduction

DNA is a sequence of 4 molecules represented by letter {A, C, T, G}. The is to build a classifier able to predict whether the sequence is bind to a Transcription factor. Our dataset consists of sequences of 101 characters, 2000 sequences and labels for the training and 1000 sequences without a label for the test. We have 4^{101} possibles sequences we have to think about the complexity of the task.

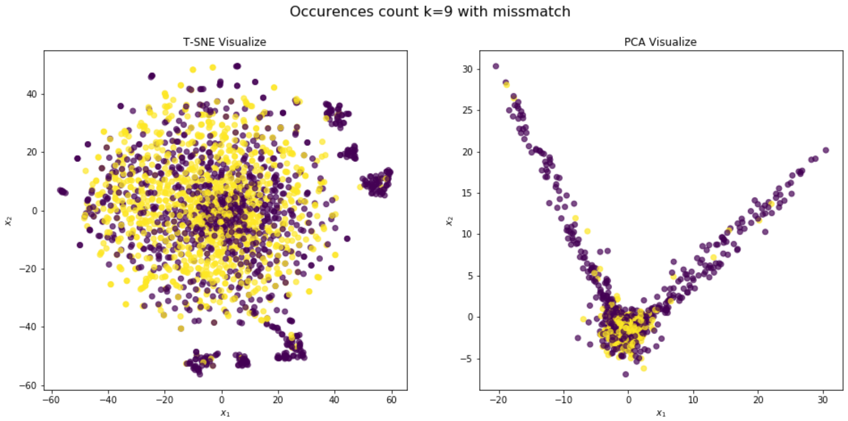
Explored Approach

Ordinal Encoding

Because we cannot train directly on the characters we just covert them to number according to their position {A=0.25, C=0.5, G=0.75, T=1.0}. After we train with Ridge Regression the result was not impressive.

Occurrence Counts

For this, we count the numbers of occurrence of a subsequence of length k in the DNA sequence and we use a vector of occurrences count. The problem of this approach the dimension of the matrix increase exponentially for k=8, we have an input dimension of 262144. To deal with this problem we implement sparse matrix and did dimensional reduction. We tried with different dimension 400, 1000 and applied ridge regressor with the linear and quadratic kernel.



Occurrence count of a sequence of 9 characters with 1 mismatch visualisation with TSNE and PCA

We discover than we increase the dimension give a better spacial structure to the data. We tried to use ridge on 2 dimensions to perform classification according to the visualization of PCA, to discretize the yellow region. Not impressive results.

One hot encoding

We encode each subsequence of length k with one-hot encoding and concatenated all the one hot sequence the sequence. The output dimensions increase very fast we use the sparse matrix to deal with it. Thanks for all the zeros in one hot. The result was similar to the occurrences count.

Spectrum kernel and mismatch kernel

This one is similar to occurrence count just in this case we build compute the kernel with the occurrence count. For the implementation, we just reuse our implementation of occurrence count and combine with the SVM kernel we did in the practical session to build the kernel with the sparse matrix representation and convert it to a dense matrix. We use gaussian and min/max normalisation. Gaussian give us a better result.

One hot Kernel.

Same as the previous one we use the implementation. We just reuse our implementation of one-hot encoding.

Results

We split the dataset with 80% train and 20% for the validation. And we did cross validated random hyper parameter with 5 fold

	Parameters	Valid Accuracy
Occurrence count + Ridge RBF kernel	lambda=4.67e-11 / sigma=8.93	66.5
Spectrum Kernel + SVM	k=10, C=1	69.5
Mismatch Kernel + SVM	k=12, mismatch=1, C=0.1	70.75

Discussion

We think the good results given by the spectrum and mismatch kernel is due to the capacity of the kernel to represent information.

References

1. A. Cohen, C. S. Leslie, E. Eskin, J. Weston, and W. S. Noble.
Mismatch string kernels for discriminative protein classification. Bioinformatics, 20(4):467–476, 01 2004.

2. C. Leslie, E. Eskin, and W. Stafford Noble.
The spectrum kernel: A string kernel for svm protein classification. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, 7:564–75, 02 2002.