KERNEL METHOD FOR MACHINE LEARNING

GENE CLASSIFICATION REPORT

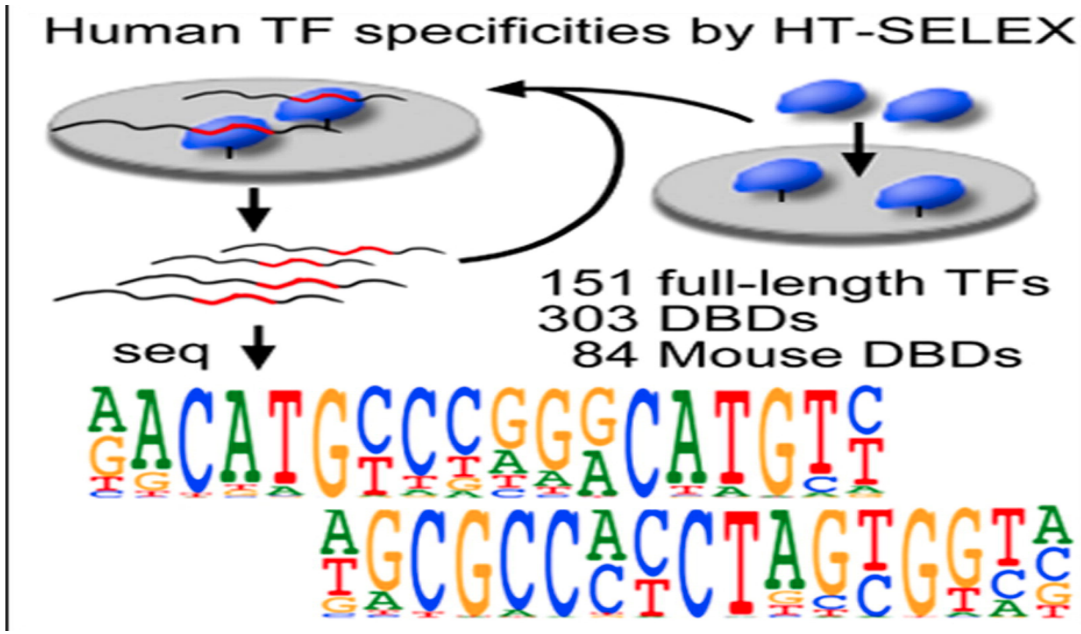
Emmanuel Owusu Ahenkan
Safiia Mohammed

AMMI GHANA

June 1, 2020

INTRODUCTION

- The main task of this project is to classify gene sequence: thus predicting whether a DNA sequence region is binding site to a specific transcription factor.



[1]

DATA SET

- The data is of two forms, namely the principal files and the optional files.
- The principal files contain data that has 2000 training points and 1000 test sequence.
- The optional files contains numeric data.
- We initially worked with the optional files but later used the principal files after performing data preprocessing.

DATA PREPROCESSING

- We performed two different kinds of data preprocessing.
- We first splitted the gene sequence related to each ID, into their various alphabetic letters and used Label Encoder on them (training data).
- We later used the spectrum kernel for preprocessing and we achieved significant improvement on our results compared to the previous.

MODELS

The models we implemented are :

- Kernel Ridge Regression
- Logistic Regression
- Kernel Logistic Regression
- Weighted Kernel Logistic Regression
- Kernel Support Vector Machine

MODELS

The Kernel methods we implemented are :

- Linear Kernel
- Quadratic Kernel
- Polynomial Kernel
- Exponential Kernel
- Radial Basis Kernel (RBF)
- Laplacian Kernel

RESULTS AND FINDINGS

- The summary of our results and findings are shown in the table below:

Model	Kernel	Parameters	Accuracy
Logistic Regression	None	-	0.614
Kernel Ridge Regression	polynomial p=5	λ=0.00001	0.680
Kernel Logistic Regression	RBF	λ=0.0001, sigma=8.0	0.660
Weighted Kernel Logistic Regression	Polynomial P=5	λ=0.0001	0.644
Support Vector Machine(SVM)	RBF	C=20, sigma=8	0.676

Table: Summary of used Models and Kernels (Public Scores).

RESULTS AND FINDINGS

- From the above table (Table), we can observe that, the Polynomial (with degree 5) and RBF kernels gives the best scores for different models. Models implemented with these kernels had better accuracies.
- Regarding the parameters, the smaller the value of Lambda (λ), the higher(better) the accuracy obtained.
- The values of sigma(σ) between [5-12] gave us better accuracies.
- The parameter C, with high values gave high accuracy during training time, however we preferred less values of C to avoid overfitting during testing time.

RESULTS AND FINDINGS

- On the Private score, the three best accuracies are: 0.684, 0.662 and 0.648 which were obtained by kernel logistic regression (polynomial kernel), Kernel ridge Regression(Polynomial kernel) and SVM with RBF kernel respectively.
- This indicates that, these kernels work well on the data set.
- In addition, simple models performed better than SVM in general.

CONCLUSION

- We observed that data preprocessing was very important. It improved the performance of the model significantly.
- We again observed that applying kernel methods to models was necessary and aided in the models obtaining better accuracies.
- Regularization and cross-validation are very important to avoid overfitting of models.
- Simple models can perform better than complex models.

REFERENCE

 Jolma, Arttu, et al. "DNA-binding specificities of human transcription factors." Cell 152.1-2 (2013): 327-339.