DNA Splice Site Detection Using Naive Bayes and K-NN

Zhongyue Zhang
June 10, 2014
Statistics 391 Final Project
zhangz6@cs.washington.edu

Abstract

Naive Bayes and K-Nearest Neighbors (K-NN) are two types of classifier. Here we trained both classifier with different feature vector to identify whether a 60 bases long DNA sequence is an intron/exon or exon/intron splicing site. Our experiments show that K-NN performs better than Naive Bayes. In particular, K-NN achieve better result than the baseline, which is to classify everything not being a splicing site, while Naive Bayes does worse than that. However, neither classifier is suitable for this task, since both of them have a low accuracy.

1. Introduction

Splice site are points on a DNA sequence where part of the DNA, called exon, is removed from the sequence during the process of protein creation. The part that is kept is called intron. The training dataset contains 3000 DNA sequences that are 60 bases long. Each base can be either be A, G, T, C, D, N, S or R. D, N, S and R represent ambiguous bases (see fig. 1). The class labels have three category, intron/exon splicing site represents by 1, exon/intron splicing site represents by 2, and 0 means it is not a splice site. The test set contains 150 examples with no ambiguous bases. We trained and validated our classifier using the training set and then test on the test set. We also use the validation to tune our hyper parameters, if there are for the specific classifier. The experiments results indicated that neither K-NN nor Naive Bayes do well on this task.

base	meaning
D	A or G or T
N	A or G or C or T
S	C or G
R	A or G

Figure 1

2. Naive Bayes

For Naive Bayes classifier, we used the frequency of AGTC at each 60 position as feature vector. We experimented both with and without removing ambiguous DNA sequences, as well as Laplace smoothing. When ambiguous DNA are removed, the Naive Bayes simply take the frequency vector as features. When ambiguous DNA are not removed, each base that can be represented by the ambiguous base are added by 1/#of possible base represented by the ambiguous base. For example, if we detect a D, 1/3 are added to A, G and T.

3. K-NN

For K-NN classifier, we tried base frequency at each position (BFP), codon (groups of three bases, since one codon can be translated to an amino acid) counts (CC) and amino acid counts (AAC). There are also two ways of counting codon and amino acid frequency. One being splitting the 60 base long DNA in to 20 groups, another being counting each 58 position with consecutive three bases. The rationale of implementing the second method is that the DNA may not align to start from exactly the first base. We also experimented with several different distance measurement. They are gram edit distance for BFP, Manhattan and Euclidean distance and cosine similarity for CC and AAC. For cosine similarity, we choose the largest k instances instead of the smallest in contrast to three other distance measurement.

4. Experiments

For training and validation, we used holdout method, where we randomly generated a holdout set from the training set with a ratio of 4 to 1, 4 being the training set and 1 being the validation set. This result in a training set of size 2400 and validation set of size 600. During training and validation we also tune our K-NN to maximize accuracy with optimal k. After validation, we train on both training and validation set, and then test on the test set.

Table 1 Validation result: Naive Bayes without Ambiguous DNA without smoothing

			Predicted Class			
			Not Splice	*IE Site*	*EI Site*	*Recall*
Actual	*Class*	*Not Splice*	*255*	*22*	*23*	*85.000%*
		IE Site	*122*	*14*	*16*	*9.211%*
		EI Site	*122*	*5*	*9*	*6.618%*
	Precision		*51.102%*	*34.146%*	*18.750%*	*47.279%*

Table 2 Validation result: Naive Bayes without Ambiguous DNA with smoothing

			Predicted Class			
			Not Splice	*IE Site*	*EI Site*	*Recall*
Actual	*Class*	*Not Splice*	*253*	*22*	*25*	*85.333%*
		IE Site	*121*	*12*	*19*	*9.211%*
		EI Site	*120*	*6*	*10*	*6.618%*
	Precision		*51.200%*	*34.146%*	*19.149%*	

Laplace smoothing made the classifier return 1 less EI Site and one more Not Splice, which shows no significant improvement. The main reason here is that the dataset is quite large, and there is no zero probability.

Table 2 Validation result: Naive Bayes with Ambiguous DNA without smoothing

			Predicted Class			
			Not Splice	*IE Site*	*EI Site*	*Recall*
Actual	*Class*	*Not Splice*	*300*	*0*	*0*	*100.000%*
		IE Site	*152*	*0*	*0*	*0.000%*
		EI Site	*136*	*0*	*0*	*0.000%*
	Precision		*51.020%*	*0.000%*	*0.000%*	

5. Conclusion