

Extension of Likelihood Ratio with Naive Bayes with Memory

Raja Loganantharaj¹

Keywords: predictive mining, computational effectiveness, improving Bayes model.

1 Introduction.

Naïve Bayes, which assumes positional independence, has been successfully used in many applications in Bioinformatics and in other fields such as artificial intelligence and data mining. Some techniques, such as positional weighted matrix and hidden Markov models, use Naïve Bayes in the underlying model. The positional independence assumption is the strength as well as its weakness of Naïve Bayes. The positional independence assumption of the Naïve Bayes makes the computation of the joint probability value easier in the expense of the accuracy or the underlying reality. In this paper we will develop a compromise between the accuracy of the underlying model and the computational efficiency. We will also demonstrate the effectiveness of our approach in applying it to a plant TATA and TATA-less promoters [1] to discriminate putative TATA box from a known TATA box

2 Formulation.

Let us formulate with respect to some prediction of motifs or some model, say M , in a DNA sequence. Let a string, say $e_{-k}, e_{-k+1}, \dots, e_{-1}, e_1, \dots, e_k$, represents k nucleotides upstream and downstream from a given pattern designated by M . Then $P(M|e_{-k}, e_{-k+1}, \dots, e_{-1}, e_1, \dots, e_k)$ and $P(\neg M|e_{-k}, e_{-k+1}, \dots, e_{-1}, e_1, \dots, e_k)$ respectively represent the conditional probability of the sequence $e_{-k}, e_{-k+1}, \dots, e_{-1}, e_1, \dots, e_k$ being a model M or not a model M . Each of e_r takes one of $\{a, c, g, t\}$. The likelihood ratio

$$\begin{aligned} & P(M|e_{-k}, e_{-k+1}, \dots, e_{-1}, e_1, \dots, e_k) / P(\neg M|e_{-k}, e_{-k+1}, \dots, e_{-1}, e_1, \dots, e_k) \\ & \text{is rewritten using Bayes theorem as} \\ & = P(M, e_{-k}, e_{-k+1}, \dots, e_{-1}, e_1, \dots, e_k) / P(\neg M, e_{-k}, e_{-k+1}, \dots, e_{-1}, e_1, \dots, e_k) \\ & = P(e_k|M, e_{-k}, e_{-k+1}, \dots, e_{-1}, e_1, \dots, e_k) \cdot P(M, e_{-k}, \dots, e_{-1}, e_1, \dots, e_{k-1}) / \\ & \quad (P(e_k|\neg M, e_{-k}, e_{-k+1}, \dots, e_{-1}, e_1, \dots, e_k) \cdot P(\neg M, e_{-k}, \dots, e_{-1}, e_1, \dots, e_{k-1})) \\ & \text{Using chain rule and applying positional independence} \\ & = P(e_k|M) \dots P(e_{k-1}|M) \dots P(e_{-k}|M) P(M) / \\ & \quad (P(e_k|\neg M) \dots P(e_{k-1}|\neg M) \dots P(e_{-k}|\neg M) P(\neg M)) \end{aligned}$$

The log likelihood ratio becomes

$$\begin{aligned} & \text{Log}(P(M/\text{sub_sequence})/P(\neg M/\text{sequence})) \\ & = \log(P(e_k|M)/P(e_k|\neg M)) + \log(P(e_{k-1}|M)/P(e_{k-1}|\neg M)) + \dots + \log(P(e_{-k}|M)/P(e_{-k}|\neg M)) + \\ & \quad \log(P(M)/P(\neg M)) \\ & = C + \sum(\log(P(e_r|M)/P(e_r|\neg M))) \text{ for all } r \text{ from } -k \text{ to } k, \text{ where } C \text{ is a constant representing} \\ & \quad \log(P(M)/P(\neg M)). \end{aligned}$$

Given a sequence of length $2k$, one can determine whether a given model M , say a transcription start site, is at location k using the likelihood ratio formulated using a Naïve Bayes. The values

¹ Center for Advanced Computer Studies, University of Louisiana, PO Box 44330, Lafayette, LA 70504 America, E-mail: logana@cacs.louisiana.edu

$P(e_r|M)/P(e_r|\neg M)$ are computed from the training set. $P(e_r|\neg M)$ is called the background probability and in many cases taken to be 0.25 for convenience. The likelihood ratio as has been defined may be viewed as 0 memory indicating the probability of a given nucleotide at a position does not depends on the surrounding nucleotides. To improve the quality of the solutions, we may want to impose some constraint that the probability is dependent on a single neighboring nucleotide, which we call a memory 1, if the probability depends on two previous neighboring nucleotides, we call it memory 2. In general we can extend the approach to memory m to denote the probability of a nucleotide for an outcome depends on the previous m nucleotides. The likelihood ratio model with memory m is given by

$$\begin{aligned} & \text{Log}(P(M/\text{sub_sequence})/P(\neg M/\text{sequence})) \\ &= C + \sum (\log(P(e_r|e_{r-1}, e_{r-2}, e_{r-3}, \dots, e_{r-m}, M)/P_{\text{back}}(e_r)) + \log(P(e_k|M)/P_{\text{back}}(e_k)) + \\ & \log(P(e_{-k+1}|e_{-k}, M)/P_{\text{back}}(e_{-k+1})) + \dots + \log(P(e_{-k+m-1}|e_{-k+m-2}, e_{-k+m-3}, \dots, e_{-k}, M)/P_{\text{back}}(e_{-k+m-1})) \quad \text{for all } r \\ & \text{from } k \text{ to } -k+m, \text{ where } C \text{ is a constant representing } \log(P(M)/P(\neg M)). \end{aligned}$$

To test the effectiveness of the extension with memory, we have tested the model with the data set that discriminate TATA box with putative TATA box [1]. We have trained the model with 80% of the data set and test it against the remaining 20%. We have run the experiment for 60 times and have tabulated the results of true positive, true negative and the overall prediction accuracy.

	Memory 0		Memory 1		Memory 2		Memory 3	
	Mean	STD	Mean	STD	Mean	STD	Mean	STD
Prediction Accuracy	50.94	4.64	53.71	4.37	57.54	5.01	54.23	5.21
True Positive	63.04	7.26	64.57	6.65	78.44	6.91	71.51	8.16
False Negative	66.23	8.48	61.68	7.83	72.08	9.45	70.26	7.92

Table 1: Results of testing the extension of likelihood ratio with memory with data set that discriminate TATA and putative TATA box.

3 Discussion.

In this brief paper we have proposed an extension to Naïve Bayes and likelihood ratio method to accommodate neighboring information to make decision of the supporting model. We have tested the proposed method on a data set on plant TATA and TATA-less promoters available at [1]. The results show that the prediction accuracy increases with memory or associated neighboring nucleotides. Having the model with memory does not increase the time complexity, but the space complexity increases. We have used a very simple threshold to determine whether it is a model M or not the model. We can improve the prediction accuracy by the optimization method that we have introduced in [2].

4 References and bibliography.

[1] PlantProm DB (<http://mendel.cs.rhul.ac.uk/mendel.php?topic=plantprom>).

[2] Loganantharaj, R. Prediction Transcription of different Genome, (Being reviewed for a Journal Publication).