Anchors: Pre-Classification and its Effects on Hidden Markov Models

Jeremy Fisher[1] and Alan Sprague[2]

Keywords: Hidden Markov Models, Data Mining

1 Introduction.

Hidden Markov Models have been prominent in the categorization of biological data in the past decade. Hidden Markov Models (HMM) make predictions about an observable sequence from a finite alphabet.

Genemark and Genscan, two popular tools used to identify regions in DNA sequence, incorporate Hidden Markov Models into their design. A large interest in these programs is to classify sequences of DNA that are composed of the small alphabet of four nucleotides. Part of the classification of these sequences is to locate introns and exons. While performing admirably, there seems to be a cap of the performance of these programs [1].

Verbumculus is a program that analyzes substrings within a sequence. It can be used on DNA sequences to try to find under-represented and over-represented substrings within that sequence. It calculates the expected value and variances of all substrings of length n in O (n^2) worst case and O(n log n) expected time. From this you can give a score to the substrings [2].

A previous experiment constructed a HMM to classify the languages of English and Spanish based on consonant and vowel rhythms, the rationale being that insights might be applied to the classification of introns and exons, and eventually other parts of the gene such as promoter regions. The experiment yielded promising results at low language transition probabilities [3].

We apply the concept of under-represented and over-represented substrings to find unusual substrings to pre-classify a small portion of an observable sequence at a high accuracy. Once pre-classified, the sequence is run through the HMM and the pre-classified parts are used as 'anchors' that alter the probabilities of the HMM to suggest a specific (pre-determined) classification. We revisited the previous experiment classifying English and Spanish with this pre-classification to investigate the results.

The addition of the pre-classified 'anchors' in a sequence increased the accuracy of the HMM. Figure 1 shows the comparison of the unaided classification with that of the pre-classified sequence. Languages were run through the HMM at both frequent (10%) and infrequent (as low as 0.1%) language transition probabilities. The sequence was pre-classified at 0.3%, 0.6%, 1.2% and 3.0 % of total sequence.

These promising results indicate that pre-classification might increase accuracy of existing classification models which incorporate HMM.

[1] University of Alabama at Birmingham. E-mail: fisherje@cis.uab.edu
[2] University of Alabama at Birmingham. E-mail: sprague@cis.uab.edu

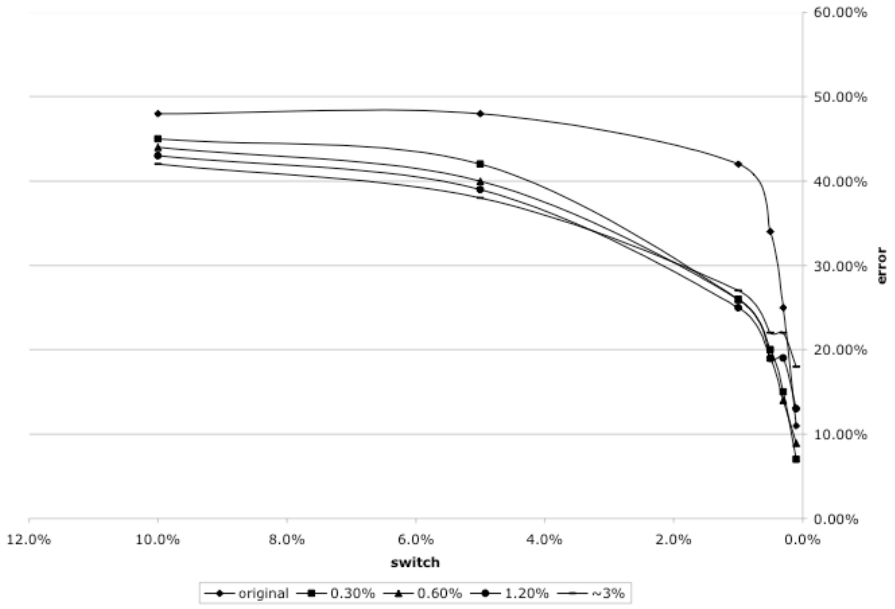

Figure 1: Alignment error of HMM given at 0%, 0.3%, 0.6%, 1.2% and 3.0% pre-classfication. Horizontal axis depicts the language transition probabilities. Vertical axis denotes the alignment error.

References

[2] Apostolico, et al. "Efficient Detection of Unusual Words" *Journal of Computational Biology*, Volume 7, Number 1/2, 2000. P 71-94.

[3] J. Fisher, F. Hernandez, A. Sprague. "Language Patterns: Comparison and Prediction Using Hidden Markov Models". *Proceedings of the ACMSE'03- ACM Southeastern Conference*

[1] Rogic et al. "Evaluation of Gene-Finding Programs on Mammalian Sequences" *Genome Research*, 2001. May 11 (5): 817-832.