Lecture 20: Machine Learning for Protein Classification

Paper reference

Abstract
This paper uses Support Vector Machines (SVMs) for the tasks of protein classification, remote homology detection and motif extraction. SVMs are the state-of-the-art among classification algorithms, and carefully designed SVM kernels can lead to substantial improvement in performance over more naïve approaches. This paper proposes a profile-based string kernel for classification of proteins, which uses probabilistic profiles such as those produced by the PSI-BLAST algorithm. This performs better on the tasks of protein classification and remote homology detection when compared to simpler SVM kernels. The SVM is trained using the kernel trick and due to which the number of parameters to be learned equals the number of training examples which is much less than the number of distinct k-length profiles. The trained SVM is then used in order to find discriminative motifs, i.e. the k-length profiles whose presence strongly indicates the presence of the positive class. This is done in a manner similar to feature selection techniques often used in Machine Learning, except that this also takes into account the number of occurrences of the particular k-mer and similar k-mers. The overall running time of the whole algorithm i.e. the kernel computation and the SVM training is much less than the running time of the PSI-BLAST algorithm.

Discussion
The profile-based kernel method performs better than methods such as Mismatch-SVM, which is a similar method but with a simpler kernel, SVM pairwise which uses a feature vector of pairwise alignment scores between the input sequence and a set of training sequences, the Fisher-SVM method which uses fisher scores extracted from a profile Hidden Markov Model as features, and also the baseline PSI-BLAST algorithm, whose results are used by the profile-based method and refined further by making features of the PSI-BLAST scores. The improvements are obtained on homology detection by way of classification and also discriminative motif extraction, which are evaluated on measures such as accuracy and ROC curves.

The paper improves on existing approaches primarily due to the design of better real valued features. The mismatch-SVM method had a feature corresponding to each k-mer and the feature value was the number of k-mers in its mutations neighborhood i.e. those that differed from the feature k-mer in less than t positions. This paper defines the mutation neighborhood to be those which have a certain minimum threshold probability of getting generated, given the emission probabilities from the profile. The probability estimates from the profile are smoothed with the background frequencies of amino acids in the training data, in order to prevent overfitting. Overfitting is making extreme generalizations from small amounts of training data and smoothing is a common way to reduce it in probabilistic models.

This paper also adapts the mismatch-SVM method of efficient computation of the kernel matrix to this problem. Here k-length profiles are passed down the tree branches instead of k-mers and a certain path in the tree is discarded if the probability so far already falls below the threshold.

A different form of feature selection is done in order to find the most discriminating positive k-length profiles (motifs). The score assigned to a k-length profile is the sum of the product of the feature weight and feature score for all positive examples. The feature weight is computed from the weights learned for the training examples.

This method can be considered semi-supervised since the PSI-BLAST algorithm also uses unlabeled data to better performance.

Experimental results are reported on the SCOP dataset and the homology detection task is to detect members that belong to the same SCOP family as the positive instances.