Poster H-58

An Automated Combination of Sequence Motif Kernels for Protein Subcellular Localization

Authors:
Alexander Zien (*MPI for Biological Cybernetics*)
Cheng Soon Ong (*Friedrich Miescher Laboratory and MPI for Biological Cybernetics*)

Short Abstract: We propose an elegant multiclass prediction approach for protein subcellular localization. First we define a family of protein sequence kernels which consider variable length motifs with gaps. Second, we generalize the multiclass SVM to automatically optimize over multiple kernels. We compare to other subcellular localization predictors on different protein datasets.

Long Abstract:

INTRODUCTION

Protein subcellular localization is a crucial ingredient to many important inferences about cellular processes, including prediction of protein function and protein interactions. While many predictive computational tools have been proposed, they tend to have complicated architectures and require many design decisions from the developer.

We focus on kernel methods, such as the popular support vector machine and proceed to define individual kernels on each set of features and combining them in a principled manner. One difficulty with adding kernels is that doing so with uniform weights is not always optimal. Multiple kernel learning (MKL) is a way of optimizing kernel weights. In this work, we propose two novel ideas to the problem of protein subcellular localization. First we define a family of sequence kernels which consider variable length motifs with gaps on amino acid sequences. This kernel compares the histograms of various motifs. Second we extend the multiple kernel learning paradigm to the multiclass case. This results in a support vector machine like algorithm which is convex and can be solved by off the shelf solvers. Hence we can obtain a globally optimal solution without resorting to complex hand crafted architectures.

MOTIF COMPOSITION KERNELS

We develop a family of kernels on sequences that is based on the occurence of motifs, that is for each motif we define a kernel. Our motif kernel is a special case of kernels on probability measures.

We compute a kernel between pairs of amino acids by utilizing the BLOSUM62 amino acid substitution scores, and extend this to k-mers by adding kernel values over each component. The final kernel between sequences is obtained from the histogram of k-mers. Instead of just considering subsequences of k consecutive amino acids, we allow for patterns consisting of k amino acids in any (fixed) positional arrangement, including gaps. For each such pattern, a separate kernel matrix can be computed from the data. Note that the combinatorial explosion of possible motifs for increasing order k is not a real problem, because the

number of motifs with positive probability is bounded by the protein length, and we employ sparse representations. The problem remaining is to combine the set of kernels in a meaningful way.

MULTICLASS MULTIPLE KERNEL LEARNING

Standard binary SVMs learn a decision function which is linear in a feature space implicitly defined by a kernel on the data points. The sign of this function corresponds to the predicted class. Its magnitude can be seen as a measure of confidence in the prediction.

Recently there have been developments in machine learning on optimizing over a set of kernels while choosing the best predictor. The algorithms consider a linear combination of kernels and optimize to find the best combination for the prediction task at hand. In this paper, we propose the first multiclass version of this approach. In other words we generalise the SVM to multiclass and multiple kernels. For the mSVM with a single kernel, training amounts to solving a quadradic program. Training the mSVM with multiple kernels includes the optimization of the kernel weights. We use a 1-norm regularizer on the convex combination weights to promote sparsity, such that only a small number of informative kernels is selected in the solution.

We show that the result of the mathematical derivation is a semi-infinite linear program (SILP). This is a linear program which can be solved by off the shelf solvers, except that it has infinitely many linear constraints. The SILP can be solved by a column generation strategy, in which we alternate training an mSVM with a single kernel and reoptimizing the combination weights by solving a linear program.

COMPARISON WITH OTHER LOCALIZATION PREDICTORS

We compare our approach with four other approaches on the TargetP dataset, PSORTdb dataset and MIPS dataset. The regularization parameter of our predictor was optimized to maximise the F1 measure, which is the geometric mean of the precision and recall. Note that this is the only parameter which we have to set in our method. To perform the comparison, we also computed the relevant measures of performance used by the other predictors. All the results were computed as an average of 10 random 80%/20% splits of the data. We performed equally well on a MIPS dataset detecting membrane proteins which is a binary classification dataset of another paper, and demonstrated that we can also further predict whether the non membrane proteins were in the nucleus, mitochondria or cytoplasm.

On both plant and non-plant data of the TargetP dataset, we performed equally well in terms of the Matthew's correlation coefficient. Our predictor is more specific but less sensitive compared to TargetP. Compared with the newer predictor TargetLoc on the same dataset, we performed significantly worse in detecting proteins in the cytoplasm and nucleus. We observed by analysing the kernel weights that 6-mers with 1 gap in the N-terminus region of the protein sequence were the most useful motifs for this task.

Compared to PSORTb on Gram positive bacteria, we were significantly better at extracellular proteins but significantly worse at locating proteins in the cell wall. For cytoplasmic proteins, although the F1 scores were similar for both predictors, our method was much better at recall

while PSORTb was much better at precision. From the kernel weights, we observed that 2-mers at the C-terminus end and 4-mers at the N-terminus end with 1 to 3 gaps were the most useful on Gram positive bacteria.

CONCLUSION

We demonstrate that our method performs comparably to the current state of the art in protein subcellular localization. Further, this is already achieved with only using information from the amico acid sequence, while our method offers a principled way of integrating other data types.

Finally, the proposed approach is very general, and could be beneficial for other multiclass bioinformatics prediction problems. Since the derivation works with arbitrary monotonically decreasing convex loss functions, class-dependent (and even data point-dependent) loss can easily be implemented. This provides another possibility for piping prior knowledge on the classes or the data into the learning process.