
Thomas Abeel, Yvan Saeys and Yves Van de Peer
VIB / Ghent University, Departement of Molecular Genetics
Bioinformatics & Evolutionary Genomics
Technologiepark 927, B-9052 Gent, BELGIUM
+ 32 (0) 9 33 13 695 (phone)
thomas.abeel@psb.ugent.be

Computer-aided gene prediction is one of the hot topics in genome analysis because it allows for computational annotation of genomes. The region before a gene is called the promoter. This promoter and in particular the core promoter is responsible for the initiation of the transcription of a gene. The identification of gene promoters and their regulatory elements is one of the biggest challenges in bioinformatics. The core promoter is the region close around the transcription start site (TSS) (we took a region of -200, +50 around the TSS). Core promoter prediction techniques try to locate the core promoter region. Machine learning techniques are often used to detect putative TSSs, although several problems exist. First, the datasets involved are rather large, ranging from several thousand training instances for the positive data to several hundred thousand negative samples. The second problem is the class imbalance between positive and negative samples. In the human genome only 3% of the sequence codes for genes and an even smaller portion is located within the core promoter. This large imbalance is very difficult to model in e.g. support vector machines (SVM) as there is so little positive information.

Here, we explore a technique to reduce the training time for support vector machines, while increasing their prediction performance. The used technique is an ensemble of support vector machines, each trained on a different part of the training set. Every one of these support vector machines is then validated on a separate validation dataset which is different from the training data. This step is needed to determine the optimal number of support vector machines that must have a positive output to classify an instance as positive. While the training time of one SVM using a RBF kernel with increasingly large datasets does not increase linearly, the time needed for training of separate SVM's on random subsamples of the large dataset increases linearly when we go for single coverage of the original dataset. We have selected more subsets than strictly necessary to have higher chances to cover the original dataset completely.

We have compared the performance using different measures for the ensemble of support vector machines and a single support vector machine. This technique was applied to a core promoter classification task. Here we have tried to distinguish core promoters from gene and intergenic sequences. We have performed our analysis on four different species: rice, arabidopsis, mouse and human. The datasets can be considered to be large, containing 2500-7000 positive training examples and ten times more negative ones, each sample consisting of 250 features. Training on a dataset of 4500 positive and 13500 negative samples on a single SVM took nearly 4 days and gave a recall of 77%, precision 88% and a F-measure value of 0.82. The validation was done on the trainingdata using a 10-fold cross validation.

Our approach with 26 SVM's each trained on 3200 samples (800 positive and 2400 negative) with a validation of 8000 sequences (80 positive, 7920 negative) performed better. The validation used is much stricter as it is much more difficult for the SVM to predict true positives. The training and validation only took a couple of hours instead of 4 days. We have obtained a recall of 94%, precision 92% and a F-measure value of 0.93.

This analysis clearly shows that the use of an ensemble of support vector machines is superior to the use of one single SVM, both in time as in classification performance. This approach can also be very valuable for similar problems with very skewed classes, like splice-site prediction.