Stochastic Models alongside Deep Learning Methods for Gene Prediction

Waldir Edison Farfan Caro, Alan Durham

Universidade de São Paulo

Abstract

Sequence labeling is the task of, given an observed sequence, determine the best label for each element according to a set of predefined categories. It has applications in many research areas where the detailed understanding of the sequence is mandatory, such as bioinformatics, natural language processing and computer vision. To address this problem, stochastic methods were developed such as Hidden Markov Models (HHMs) and Conditional Random Fields (CRFs) due to their ability to model the relationship of the members of the sequences. Lastly, Deep Neural Networks (DNNs) have been used satisfactorily for sequence labeling problem since they are suitable to automatically learn complex feature representation from data, which is an advantage to models designed by hand. Since CRFs have good achievements in modeling and DNNs have a high capacity in learning the representation. In this work we propose the use of mixed methods of CRFs and DNNs capable of improve the task of sequence labeling, in special the problem of gene prediction, by giving suitable models for the sequences with a rich representation of their features. For this porpouse, we use the ToPS framework, as a source of efficient implementation of Markov Models and the recently integration of CRFs; and the PyTorch framework which offers optimized tensor implementation for deep learning methods. Both frameworks follow an object oriented approach giving the chance to a better blending of the techniques. Thus, for gene prediction we gain better feature representation of genome sequences beside rich modeling of their structure, even for complex organisms.

Funding: