Chapter 5
A Prediction Approach Based on Self-Training and Deep Learning for Biological Data

Mohamed Nadjib Boufenara
LIRE Laboratory, Abdelhamid MEHRI, Constantine 2 University, Algeria

Mahmoud Boufaida
LIRE Laboratory, Abdelhamid MEHRI, Constantine 2 University, Algeria

Mohamed Lamine Berkane
LIRE Laboratory, Abdelhamid MEHRI, Constantine 2 University, Algeria

ABSTRACT

With the exponential growth of biological data, labeling this kind of data becomes difficult and costly. Although unlabeled data are comparatively more plentiful than labeled ones, most supervised learning methods are not designed to use unlabeled data. Semi-supervised learning methods are motivated by the availability of large unlabeled datasets rather than a small amount of labeled examples. However, incorporating unlabeled data into learning does not guarantee an improvement in classification performance. This paper introduces an approach based on a model of semi-supervised learning, which is the self-training with a deep learning algorithm to predict missing classes from labeled and unlabeled data. In order to assess the performance of the proposed approach, two datasets are used with four performance measures: precision, recall, F-measure, and area under the ROC curve (AUC).

INTRODUCTION

The increasing volume and complexity of biological data has created the need for powerful data analysis tools and methods (Stephens et al., 2015). This increase in data volume presents new opportunities, but also poses new challenges.

DOI: 10.4018/979-8-3693-3026-5.ch005

New technologies have made possible to examine biological data such as genome, transcriptome or phenotype (Hasin et al., 2017). In order to build a prediction system, it is necessary to dispose a set of training data that includes labeled examples. However, with the advent of next-generation sequencing (Chang, 2018; Davey et al., 2011) and biobanks (Allen et al., 2014) a lack of labeled biological data was found during the collection of the training set. The imbalance between the amount of tagged and unlabelled data causes a problem when building a prediction system. This imbalance is due to the difficulty in obtaining the labels because it is an expensive operation and / or takes a long time and may not be possible in the case of rare diseases for example. This limitation prohibits learning a specific classifier in many scenarios (Rajpurkar et al., 2017). Although unlabeled data are comparatively plentiful than labeled examples, most supervised learning methods are not designed to use unlabeled examples (Patnaik & Popentiu-Vladicescu, 2018).

A class of machine learning strategy called Semi-Supervised Learning (SSL) is able to deal with this problem (Zhu & Goldberg, 2009). The SLL is a class of algorithms that uses unlabeled data as well as labeled examples when forming a model. Co-training (BLUM & MITCHELL, 1998) and self-training (Zhu & Goldberg, 2009) are instances of SSL, which are often used in scenarios where the number of labeled examples is small and the number of unlabeled instances is large. This imbalance is due to the high cost of labeling biological data. The SSL can be articulated on a wide variety of prediction methods such as Random Forest (Criminisi et al., 2012), Support Vector Machine (SVM) (Li et al., 2008) or Deep Learning (BELLOT et al., 2018). The latter permits the computational models composed of several processing layers to learn data representations with several levels of abstraction. An open question is whether unlabeled data can help improve the accuracy of the prediction when the amount of labeled data is small.

This paper is interested in self-training, which is a class of SSL strategies. One of the major difficulties of self-training remains the amplification of errors. If a predictor predicts incorrectly certain classes and the predicted ones are added to the labeled training dataset, the next predictor can learn incorrect examples and generate more misclassified examples. This can create more erroneous classifiers in each iteration and the classification error can be magnified (Lee et al., 2017).

However, the quick development of artificial intelligence, and more specifically deep learning, offers new options for the improvement of prediction accuracy (LeCun et al., 2015).

One of the important applications of deep learning in the biological field is to predict the effect of mutations from a DNA sequence. Such model-based evaluations of the effect of sequence changes complement methods based on Quantitative Trait Locus (QTL) mapping, and can in particular help to accurately map probable causal genes (Angermueller et al., 2016).

Encouraged by these recent successes, the authors propose in this paper a prediction approach based on deep learning and self-training to impute missing classes and improve the accuracy of biological data prediction. The objective of this study is to:

- Improve the accuracy of prediction of biological data by integrating unlabeled data;
- Try to decrease the error amplification.

This approach is mainly based on three phases. The first phase is a training phase with labeled data only. The second phase involves self-training with unlabeled data. The third one is a test phase to assess the efficiency of deep learning used to measure the accuracy of the model built with measured data only, and model built on measured and self-trained data. A full evaluation of the model formation process