Active Semi-Supervised Learning for Analysis of Biological Data

Guilherme Camargo[1], Pedro Henrique Bugatti[2], Priscila T M Saito[2],

1 Programa de Pós-Graduação em Bioinformática – PPGBIOINFO, Universidade Tecnológica Federal do Paraná, Cornélio Procópio
2 Federal University of Technology – Parana

Abstract

In the last few years, new data capture devices have made it possible a major technological breakthrough. Thus, large complex databases (e.g. images, sounds or texts) are obtained daily. In order to allow the storage and the retrieval of information from these databases, it is necessary specialists to annotate the samples. However, the annotation by specialists can bring inconsistencies to the samples, since different individuals can interpret the samples in divergent ways. Another reason to consider is the cost to perform the task of annotating the samples, which is high and exhaustive to specialists. Therefore, a solution to the problem would be to automate the process of identifying the most informative samples using computational methods. In this way, a label is assigned to each sample, classifying it according to the scope of the problem. One way to develop a suitable solution is applying machine learning techniques in order to build a pattern classifier. To take benefit from the large number of unsupervised samples available in disproportion to the scarcity of supervised ones, semi-supervised learning techniques have been explored using partially supervised and unsupervised training samples, where supervised samples propagate their labels to the unsupervised ones. However, such techniques neglect the existence of redundant samples, as well as the existence of more relevant samples that could boost the classifier learning. In this context, active learning techniques associated with semi-supervised techniques are interesting, since a smaller number of more informative samples automatically selected through the active learning strategy, and then annotated by a specialist can propagate the labels to a set of unsupervised samples (through the semi-supervised learning strategy). Therefore, we developed a new active semi-supervised learning approach for biological data, exploring new strategies for selecting more informative samples for the classifier learning. Preliminary results show that the union of active and semi-supervised learning improves accuracy for some biological datasets, reaching higher values faster, showing that the obtained active semi-supervised classifiers are more efficient than the supervised and the semi-supervised ones.

Funding: CNPq (#431668/2016-7, #422811/2016-5), CAPES, Fundação Araucária, SETI, and UTFPR.