One class support vector machine for textual gene profiles prediction

Shi Yu

Ph.D Student
Bioinformatics group, SISTA-SCD-ESAT, Department of Electrical Engineering, K.U.Leuven

Abstract

The integration of machine learning and bioinformatics enables biologists to infer the property of unknown instance with pattern of known dataset. However, the underlying assumptions in machine learning are not exactly applicable for biological problems. For example, in biological studies the known data and unknown data might be heterogeneous and their number might not be equally balanced. In some cases, only positive samples are well studied while negative samples are not so easy to find, which is not perfectly applicable for binary classification case. In biological field, “inferring” is more or less conducted in the nature way of human learning and sometimes the counter samples are not necessary to learn.

In the present work, we used a version of Scholkopf et al. (1995)’s one class SVM to detect the similarity of gene profiles based only with a set of positive instances. One class SVM is an extension of binary SVM classifier which tries to describe one class of samples and distinguish it from other possible samples. Based on one class SVM, we create a web application *TextPredictor* to predict the similarity between gene profiles. The gene expressions we used in our experiments are based on TxtGate, a former biomedical text mining system created in year 2004. The learning process in TextPredictor is implemented in full genomic size scanning of certain species. Different configurations of text mining and kernel settings are also benchmarked by the disease relevant datasets. According to the comparison with other methodologies such as Person correlation measure and non-kernel one class learning methods, one class SVM shows nice properties in the efficiency of learning and the specificity of prediction.

Keyword
Bioinformatics, Machine Learning, Text Mining, One Class SVM, Gene Prediction