Semi-Supervised Embedding for Predicting Protein-Protein Interactions from Multiple Sources

Yanjun Qi, NEC Laboratories America
Oznur Tastan, Carnegie Mellon University
Jaime Carbonell, Carnegie Mellon University
Judith Klein-Seetharaman, Carnegie Mellon University
Jason Weston, NEC Laboratories America

Protein-protein interactions (PPIs) are critical for virtually every biological function. Recently through the integration of direct and indirect biological sources, researchers suggested using supervised learning methods for the task of classifying pairs of proteins as interacting or not. One key difficulty with the supervised approaches is that they require a large number of labeled training examples to learn accurately. In contrast to vastly explored yeast or human interactomes, for many biologically interesting organisms or intra-species PPIs, e.g. between pathogen and host organisms, availability of reliable sets of known interaction partners is often rare. This data restriction strongly reduces the predictive power of computational PPI algorithms.

Meanwhile there exist a considerable amount of pairs where an association is proposed between the partners, but not enough experimental evidence is presented to support it as a direct interaction (*partially labeled*). In this paper we propose a semi-supervised learning technique for predicting PPIs from not only *labeled*, but also *partially labeled* data sets. The basic idea is to combine an embedding-based regularizer with a supervised classifier to perform semi-supervised classification. We train a multi-layer perceptron network to uncover the data clustering structure by embedding *partially labeled* examples. Simultaneously we train this network on *labeled* data for PPI classification. We applied our method to predict the set of interacting proteins between HIV-1 and human proteins. Our method improved upon the best previous method for this task indicating the benefits of semi-supervised embedding with auxiliary information.

Topic: learning application

Preference: oral/poster