Resampling Methods to Handle the Class-Imbalance Problems in Predicting Protein-Protein Interaction Site and Beta-Turn

様式　4.

学　位　論　文　概　要

学位論文題名

Resampling Methods to Handle the Class-Imbalance Problems in Predicting Protein-Protein Interaction Site and Beta-Turn

（和訳）

タンパク質間相互作用予測およびβターン予測におけるクラス不均衡問題を扱うためのリサンプリング手法

電子情報科学　専攻　知能情報・数理　講座

氏　　名　Nguyen Thi Lan Anh

主任指導教員氏名　Kenji Satou

学位論文概要

Protein is a functional biomolecule with complicated three-dimensional structure. Only proteins have the suitable 3D structures can contact together to perform the vital functions in cells. Learning about protein structure and protein-protein interaction are the important tasks in bioinformatics.

This thesis aimed at solving the class-imbalance problems in predicting (i) protein-protein interaction sites; and (ii) β-turns and their types. Firstly, we proposed a novel over-sampling method for handling the severe imbalanced datasets to improve the performance of predicting protein-protein interaction sites. The combinations of our new algorithm with KSVM-THR and random under-sampling methods were also proposed. Experimental results showed that our new methods achieved higher sensitivity, precision, G-mean, F-measure, and AUC-PR in comparison with the state-of-the-art methods. In addition, we found that predicted shape strings enhance the performance of the prediction.

Secondly, we investigated the information of predicted protein blocks and applied for predicting β-turns and their types. The use of this feature can improve the prediction results in comparison with the most recent publications. We utilized random under-sampling method to deal with the imbalanced datasets and feature selection to remove the redundant features. Results of experiments on three standard benchmark datasets showed that our methods are comparable with the state-of-the-art methods.