Genome-wide Protein Function Prediction via Multi-instance Multi-label Active Learning

Jiansheng Wu1, \*, Weijian Wu2, Jun Wang 1

吴建盛1, \*， 吴伟坚2， 王俊1

1School of Geography and Biological Information, Nanjing University of Posts and Telecommunications, Nanjing 210023

1地理与生物信息学院，南京邮电大学，南京 210023

2College of Computer and Information, Hohai University, Nanjing 210024

2计算机与信息学院，河海大学，南京 210024

*Email: jansen@njupt.edu.cn* (Jiansheng Wu), *[1773468916@qq.com](mailto:1773468916@qq.com)* (Weijian Wu), [*wangj@njupt.edu.cn*](mailto:wangj@njupt.edu.cn) (Jun Wang)

Abstract

As the number of sequenced genomes rapidly grows, there are a large number of proteins with unknown function in new genomes. Active learning methods can assist biologists for selecting the most valuable ones as candidates for biological experiments. Previously, it is proved that the protein function prediction task is naturally Multi-Instance Multi-Label (MIML) learning problem. In this paper, we formulate the problem of selecting the most valuable proteins in annotating genome-wide protein functions as a MIML active learning task, Then, we propose a MIML active learning framework named MIMLAL and design two algorithms MIMLAL-A and MIMLAL-R for genome-wide protein function prediction. MIMLAL minimizes an approximated surrogate loss by stochastic gradient descent using the most valuable bag-label pairs, which are chose by the selection criterion combining label cardinality inconsistency (LCI) and diversity. We have performed experiments on multiple real-world organisms including *Geobacter sulfurreducens, Azotobacter vinelandii, Shewanella loihica PV-4, escherichia coli strain K12, Saccharomyces cerevisiae, Caenorhabditis elegans, Drosophila melanogaster，Mus musculus，Homo sapiens*. The proposed approaches MIMLAL-A and MIMLAL-R achieve excellent performance in all cases.

Keywords: Protein Function Prediction, Multi-instance Multi-label Learning, Active Learning

Reference

[1] Wu J, Huang SJ, Zhou ZH. Genome-wide protein function prediction through multi-instance multi-label learning[J]. IEEE/ACM Transactions on Computational Biology & Bioinformatics, 2014, 11(5), 891-902.

[2] Zhou ZH, Zhang ML, Huang SJ, Li YF, et al. (2012) Multi-instance multi-label learning [J]. Artificial Intelligence, 2012,176(1), 2291-2320.

[3] Huang SJ, Gao W, Zhou ZH. Fast multi-instance multi-label learning [C]. In: Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI'14), Quebec City, Canada, 2014, pp.1868-1874. (CORR abs/1310.2049)

[4] Huang SJ , Zhou ZH. Active query driven by uncertainty and diversity for incremental multi-label learning [C]. In: Proceedings of th 13th IEEE International Conference on Data Mining (ICDM'13), Dallas, TX, 2013, pp.1079-1084.