IGI PUBLISHING

701 E. Chocolate Avenue, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.igi-global.com

ITJ 5528

This paper appears in the publication, International Journal of Computational Models and Algorithms in Medicine, Volume 1, Issue 1
edited by **Aryya Gangopadhyay © 2010, IGI Global**

Classification Systems for Bacterial Protein-Protein Interaction Document Retrieval

Hongfang Liu, Georgetown University Medical Center, USA

Manabu Torii, Georgetown University Medical Center, USA

Guixian Xu, Minzu University of China, China

Johannes Goll, The J. Craig Venter Institute, USA

ABSTRACT

Protein-protein interaction (PPI) networks are essential to understand the fundamental processes governing cell biology. Recently, studying PPI networks becomes possible due to advances in experimental high-throughput genomics and proteomics technologies. Many interactions from such high-throughput studies and most interactions from small-scale studies are reported only in the scientific literature and thus are not accessible in a readily analyzable format. This has led to the birth of manual curation initiatives such as the International Molecular Exchange Consortium (IMEx). The manual curation of PPI knowledge can be accelerated by text mining systems to retrieve PPI-relevant articles (article retrieval) and extract PPI-relevant knowledge (information extraction). In this article, the authors focus on article retrieval and define the task as binary classification where PPI-relevant articles are positives and the others are negatives. In order to build such classifier, an annotated corpus is needed. It is very expensive to obtain an annotated corpus manually but a noisy and imbalanced annotated corpus can be obtained automatically, where a collection of positive documents can be retrieved from existing PPI knowledge bases and a large number of unlabeled documents (most of them are negatives) can be retrieved from PubMed. They compared the performance of several machine learning algorithms by varying the ratio of the number of positives to the number of unlabeled documents and the number of features used.

Keywords: *Document Retrieval, Manual Curation, Protein-Protein Interaction (PPI) Networks*

INTRODUCTION

Protein-protein interaction (PPI) network is essential to understand the fundamental processes governing cell biology. PPI network data for several organisms has already been generated by high-throughput studies and submitted to/ or collected by various PPI interaction databases (Morrison et al., 2005). However, the majority of PPIs are reported in the scientific literature. To collect such data in a standard-

DOI: 10.4018/jcmam.2010072003

Copyright © 2010, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

ized way and to avoid duplication of efforts, IMEx[1] databases such as IntAct (http://www.ebi.ac.uk/intact), DIP (Database of Interacting Proteins; http://dip.doe-mbi.ucla.edu), MINT (Molecular Interactions Database; http://mint.bio.uniroma2.it/mint) and MPIDB (Microbial Protein Interaction Database; http://www.jcvi.org/mpidb) conduct coordinated manual literature curation. Text mining system to prioritize articles for curators according to their PPI relevance can accelerate such curation processes significantly. For example, MPIDB curators scan a whole issue (20 to 50 articles) of the Journal of Bacteriology or Molecular Microbiology and find approximately 10% of these articles report interaction experiments. Thus, the curators spend roughly 90% of their time reading irrelevant articles. A text mining system to prioritize articles for curators can be developed using supervised classification algorithms that provide certain kinds of confidence scores during classification. In order to build such systems, a class-labeled corpus is needed where PPI-relevant documents are labeled as positive and those irrelevant as negative. In many real-world applications, it is common that positive instances are explicitly included in a designated database, but it is uncommon to also include negatives in the database (Elkan & Noto, 2008). In developing a PPI mining application, PPI-relevant documents can be retrieved from existing PPI knowledge bases and unlabeled documents are available in large literature repositories such as PubMed. Learning with only positively labeled documents has great importance in this application.

We consider learning with only positive labeled documents as learning from a noisy and imbalanced training set where unlabeled documents are considered as negatives with some mislabeled documents. We build a document retrieval system to assist the curation of MPIDB (Goll et al., 2008) and report our investigation of the stability of two document classification algorithms with respect to the ratio of positives and unlabeled documents in the training set and also of the impact of feature selection on the classification performance. We also propose to use different subsets of unlabeled documents and form an ensemble of classifiers.

In the following, we first describe the background of classification algorithms. The experimental methods are introduced next. We then present the results and discussion, and conclude our work.

BACKGROUND

Learning from Positives and Unlabeled

Existing methods for learning from positives and unlabeled (LPU) commonly adopt a two-step strategy where the first step is to obtain a reliable negative data set (RN) based on keywords or available classifiers and the second step is to refine or augment RN using various learning approaches such as clustering or boosting. One popular approach to obtaining RN is to define a classification task which considers unlabelled documents as negatives, build a binary classifier, and treat those predicted negatives as RN. For example, Li and Liu proposed a method where they first build a Rocchio classifier based on positives and regard negatives according to that classifier as RN. Similar technology can be used to augment training data for PPI document classification. For example, Tsai et al. trained a SVM classifier with 3,536 positives (TPs) and 1,959 negatives (TNs) from BioCreAtIvE II workshop. Then the training set was augmented by applying the SVM classifier to the likely positive data sets and 50,000 unlabeled MEDLINE. Finally, an improved SVM classifier was derived from the original corpus using additional features based on the augmented corpus.

Recently, Elkan and Noto showed that if positive documents were randomly sampled, the conditional probability of a given document being labeled (and thus positive) differ from the conditional probabilities of a given document being positive only by a constant factor (Elkan & Noto, 2008). We have also found that when defining document retrieval

Copyright © 2010, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

9 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the product's webpage:
www.igi-global.com/article/classification-systems-bacterial-protein-protein/38943?camid=4v1

This title is available in InfoSci-Journals, InfoSci-Journal Disciplines Medicine, Healthcare, and Life Science. Recommend this product to your librarian:
www.igi-global.com/e-resources/library-recommendation/?id=2

Related Content

A Novel Particle Swarm Optimization Algorithm for Multi-Objective Combinatorial Optimization Problem
Rahul Roy, Satchidananda Dehuri and Sung Bae Cho (2013). *Trends in Developing Metaheuristics, Algorithms, and Optimization Approaches (pp. 1-16).*
www.igi-global.com/chapter/novel-particle-swarm-optimization-algorithm/69714?camid=4v1a

Digital (or Virtual) Hoarding: Emerging Implications of Digital Hoarding for Computing, Psychology, and Organization Science
Jo Ann Oravec (2018). *International Journal of Computers in Clinical Practice (pp. 27-39).*
www.igi-global.com/article/digital-or-virtual-hoarding/210558?camid=4v1a

Scalability of Piecewise Synonym Identification in Integration of SNOMED into the UMLS
Kuo-Chuan Huang, James Geller, Michael Halper, Gai Elhanan and Yehoshua Perl (2011). *International Journal of Computational Models and Algorithms in Medicine (pp. 26-45).*
www.igi-global.com/article/scalability-piecewise-synonym-identification-integration/60649?camid=4v1a

A Survey of Ant Colony Optimization Algorithms for Telecommunication Networks

Ilham Benyahia (2012). *International Journal of Applied Metaheuristic Computing (pp. 18-32).*

www.igi-global.com/article/survey-ant-colony-optimization-algorithms/67331?camid=4v1a