Applying Lazy Local Learning in BC II.5 Article Categorization Task

Cheng-Ju Kuo

Cheng-Ju Kuo[1a], Maurice HT Ling[3,4c], and Chun-Nan Hsu[1,2b]

[1] Institute of Information Science, Academia Sinica, Taiwan
[2] USC/Information Science Institute, Marina del Rey, CA, USA
[3] School of Chemical and Life Sciences, Singapore Polytechnic, Republic of Singapore
[4] Department of Zoology, The University of Melbourne, Parkville, Victoria, Australia
a) cju.kuo@gmail.com,
b) chunnan@iis.sinica.edu.tw,
c) mauriceling@acm.org

Article categorization task (ACT) in BC II.5 is to classify whether the input article contains protein interaction descriptors.

In our approach, given an article for classification, the system will invoke the selection of top 100 abstracts from the 5,495 abstracts in the BC II - IAS data corpus based on the document cosine similarity between the query article and each abstract in the IAS dataset. These 100 abstracts were used to train a classifier by AdaBoost (with Decision Tree as the weak learner), implemented in MALLET. The trained classifier will then be used to class the query article into either containing protein interaction or not. Therefore, for each query, there will be a unique classifier trained by a different set of selected abstracts. This is a lazy local approach in the sense that a classifier will not be trained until a query article is given and the training examples are selected locally in the neighborhood of the query article in the feature space.

In order to reduce the feature size generated from full-text article, only title, abstract and captions of figures of FEBS full-text article were used to generate feature vectors for model training and testing after the removal of common words and stemming of tokens.

Independent evaluation demonstrated 17.4% precision for identifying articles containing protein interaction descriptors (58 true positives out of 333 identified positives) but 98.9% precision in identifying articles not containing protein interaction descriptors (259 true negatives out of 262 identified negatives). This result suggests that our system can be applied as an effective filter to eliminate articles without protein interaction descriptors and thus greatly reduce the number of articles for subsequent processing of extracting protein interactions.