

Tackling Class Imbalance and Data Scarcity in Literature-Based Gene Function Annotation

Mathieu Blondel
Kobe University
1-1 Rokkodai, Nada, Kobe
657-8501, Japan
mblondel@ai.cs.kobe-u.ac.jp

Kazuhiro Seki
Kobe University
1-1 Rokkodai, Nada, Kobe
657-8501, Japan
seki@cs.kobe-u.ac.jp

Kuniaki Uehara
Kobe University
1-1 Rokkodai, Nada, Kobe
657-8501, Japan
uehara@kobe-u.ac.jp

ABSTRACT

In recent years, a number of machine learning approaches to literature-based gene function annotation have been proposed. However, due to issues such as lack of labeled data, class imbalance and computational cost, they have usually been unable to surpass simpler approaches based on string-matching. In this paper, we propose a principled machine learning approach based on kernel classifiers. We show that kernels can address the task's inherent data scarcity by embedding additional knowledge and we propose a simple yet effective solution to deal with class imbalance. From experiments on the TREC Genomics Track data, our approach achieves better F_1 -score than two state-of-the-art approaches based on string-matching and cross-species information.

Categories and Subject Descriptors

H.4 [Information Systems Applications]: Miscellaneous;
I.2.7 [Artificial Intelligence]: Natural Language Processing—*Text analysis*

General Terms

Algorithm, Experimentation

Keywords

Text Classification, Kernel Methods

1. INTRODUCTION

Since the completion of the Human Genome Project, a large number of studies have been conducted to identify the role of individual genes. However, the active research in the domain has been producing numerous publications. As a result, it is extremely labor-intensive for biomedical researchers alone to collect the information relevant to their need, since obtaining such information requires intensive reading. To remedy this problem, a number of organizations have been working on annotating each gene of model organisms with a controlled vocabulary organized as a Directed Acyclic Graph, called Gene Ontology (GO) terms, based on the contents of the published scientific articles. However, the annotation requires trained human experts with extensive domain knowledge. With limited human resources and

the ever-growing literature, it was reported that it would never be completed at the current rate of production.

To alleviate the burden, TREC 2004 Genomics Track and BioCreative targeted automatic GO domain/term annotation. Overall, participants from the Genomics Track reported the effectiveness of supervised classification techniques while participants from BioCreative mainly adopted string-matching techniques. These different strategies can be attributed to the fact that Genomics Track only considered the three top-level GO terms (called GO domains) while BioCreative targeted all GO terms, which can amount to up to 30,000.

In this paper, we propose a principled machine learning approach based on kernel classifiers. Kernel are computationally efficient and allow us to address the data scarcity issue by embedding additional knowledge. Moreover, existing methods for automatic GO annotation usually neglect the so-called class imbalance problem. We propose a simple but effective method to deal with the problem. From experiments on the Genomics Track data, our approach outperforms two existing state-of-the-art methods based on string-matching and cross-species information.

2. PROPOSED METHOD

2.1 System overview

The goal of this research is to design a system that can leverage annotated articles from the scientific literature in order to *learn* how to automatically map the function of genes mentioned in new articles to appropriate GO terms. More precisely, for a new non-annotated article, we would like to be able to answer the question: does the article contain supporting evidence that gene X has GO term Y? Our system is divided into a learning phase and a prediction phase, as depicted in Figure 1. In the information extraction step, articles are represented by concatenating three kinds of textual evidence: article title, abstract and relevant text fragments extracted by gene mention matching. In the learning step, since GO annotation is a multi-label task (i.e., one article can be associated with zero, one or more GO terms), we use a one-vs-all scheme, which consists in learning one binary classifier per GO term. In the post-processing step, we remove illogical GO term predictions based on the DAG structure.

2.2 Class imbalance

To deal with class imbalance (the number of annotated articles varies quite a lot from a GO term to another), we

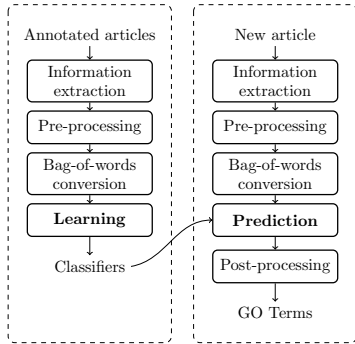


Figure 1: System overview: the learning (left) and prediction (right) phases.

employ a simple technique inspired by Osuna et al. [2] which consists in applying a stronger weight to positive examples than to negative examples. When learning the classifier for the c^{th} GO term, we set the soft-margin SVM parameter C to $\lambda(n_{-c}/n_c)$ when an article is associated with this GO term and to λ otherwise. Here, we defined n_{-c} as the number of articles which are not labeled with the c^{th} GO term and n_c as the number of articles which are. λ is optimized by stratified cross-validation. Since $n_{-c} > n_c$, this consists in assigning a stronger weight to positive examples than to negative examples. This simple heuristic produced better results than other techniques such as negative example downsampling.

2.3 Data scarcity

In this research, vectors (bag of words) are very sparse and contain on average less than 300 unique words. We can therefore reasonably expect potentially discriminative words to be either absent or underrepresented in the vectors. For this reason, we propose to use the probabilistic Latent Semantic Analysis (pLSA) kernels [1]. Since they are estimated from all the articles, we can expect the latent topics used in these kernels to smooth the probability of words based on their semantic context. This is especially useful for GO terms which are annotated with only few articles.

3. EXPERIMENTS

For evaluation, we used the test dataset provided for the TREC 2004 Genomics Track and supplemented it by GO term information. The resulting dataset consists of 863 test instances. Each instance is a pair, represented by a PubMed ID and a gene mentioned in the article. For training data, we used the Genomics Track training dataset (1418 instances with full text available) supplemented with GO terms in the same way and MGD database (6750 instances with abstract only). These datasets are dealing with mouse genes only. We performed the usual document preprocessing: stop-word, punctuation and long-word (> 50 characters) removal, stemming and lowercase conversion. As an evaluation metric, we used the F_1 score for direct comparison with the previous work.

We used SVM as our classifier and compared three different kernels: \mathcal{K}_{linear} , $\mathcal{K}_{polynomial}$ and \mathcal{K}_{plsa} . A summary of the results is reported in Table 1. $\mathcal{K}_{plsa} (+U)$ corresponds to the results obtained when an additional 10,000 unlabeled

Table 1: Performance comparison.

KERNEL	PRECISION	RECALL	F_1 -SCORE
\mathcal{K}_{linear}	0.36	0.20	0.26
$\mathcal{K}_{polynomial} (d = 2)$	0.35	0.19	0.25
\mathcal{K}_{plsa}	0.38	0.20	0.26
$\mathcal{K}_{plsa} (+U)$	0.39	0.22	0.28
$\mathcal{K}_{plsa} (+U + T)$	0.38	0.24	0.29
$\mathcal{K}_{plsa} (+U + T + O)$	0.42	0.23	0.30
STOICA & HEARST	0.19	0.46	0.27
SEKI ET AL.	0.26	0.27	0.26

abstracts from the MGD database were used to learn the pLSA model (semi-supervised learning). $\mathcal{K}_{plsa} (+U + T)$ corresponds to the results obtained when the test set was also used to learn the pLSA model, thereby tailoring the classifiers to the task of interest (transductive learning).

For direct comparison, Table 1 provides the results of the methods of Stoica and Hearst [4] (re-implementation by the authors) and Seki et al. [3] (results taken from their paper). Our method achieves the highest precision and F_1 score. On the other hand, Stoica and Hearst accomplish the highest recall. This is because they perform string matching using the entirety of the GO term descriptors but eliminate inconsistent candidates using cross-species constraints. $\mathcal{K}_{plsa} (+U + T + O)$ are our results when cross-species constraints were also used in post-processing. Based on the t-test, the difference between our best method and the method of Stoica and Hearst was found statistically significant at the 0.05 significance level ($p = 0.03$).

4. CONCLUSION

This study proposed an approach to GO term annotation based on kernel classifiers. From the experiments on the Genomics Track data, we observed that 1) our system exhibits high precision and performs better in terms of F_1 -score than two state-of-the-art methods based on string-matching and cross-species information 2) latent topics can be advantageously embedded into kernels and are a promising way to address the inherent data scarcity in GO term annotation 3) per-class regularization is a simple yet effective way to deal with the class imbalance problem.

5. REFERENCES

- [1] T. Hofmann. Learning the similarity of documents: An information-geometric approach to document retrieval and categorization. In S. A. Solla, T. K. Leen, and K.-R. Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 914–920, 1999.
- [2] E. E. Osuna, R. Freund, and F. Girosi. Support vector machines: Training and applications. Technical report, Massachusetts Institute of Technology, 1997.
- [3] K. Seki, Y. Kino, and K. Uehara. Gene functional annotation with dynamic hierarchical classification guided by orthologs. In *Proc. of Discovery Science*, volume 5808, pages 425–432, 2009.
- [4] E. Stoica and M. Hearst. Predicting gene functions from text using a cross-species approach. In *Proc. of Pacific Biocomputing Symposium*, volume 11, pages 88–99, 2006.