Learning sets of positive rules of amino acid properties to classify protein functional classes

Aik Choon Tan [(1)], Ali Al-Shahib [(1)], David Gilbert [(1)] and Yves Deville [(2)]

[(1)] Bioinformatics Research Centre, Department of Computing Science, University of Glasgow,
17 Lilybank Gardens, G12 8QQ Glasgow, United Kingdom.
{actan, alshahib, drg}@brc.dcs.gla.ac.uk

[(2)] Department of Computing Science and Engineering, Université catholique de Louvain, Place Sainte Barbe, 2,
B-1348 Louvain-la-Neuve, Belgium.
deville@info.ucl.ac.be

Keywords. Machine learning, pattern sets, imbalanced data, protein functional class, annotation.

Introduction

With the availability of full-scale genome of various organisms, one of the recent bioinformatics challenges is to accurately assign gene products into their functional classes. Standard bioinformatics tools such as detecting sequence homology by using PSI-BLAST [1] and FASTA [3] provide initial hints to the experimental determination of function.

At the abstract level, protein functional class prediction can be regarded as mapping a sequence to its biological function(s). In the field of machine learning, the annotation of gene products can be viewed as a standard classification problem. For some multi-class classification problems, the set of positive examples is very small compared to the set of negative examples; this is the common scenario in the functional annotation problem where there exist a lot of functional classes but the number of the examples (protein sequences) in each class is relatively low. This imbalanced proportion of examples in each class contributes to the poor performance of standard machine learning techniques (e.g. decision trees). These approaches tend to produce a strong discrimination classifier (high overall predictive accuracy) with very low sensitivity (positive coverage) when learning on these types of problems.

In this paper we propose a novel machine learning approach POS-SET, that generates a classifier with a better sensitivity, while not loosing too much positive prediction accuracy. We present the general framework of POS-SET and describe its application to learn sets of positive rules in classifying protein functional classes of *P.gingivalis*.

Methods

For a supervised classification problem, a set of training data (positive and negative examples) in the form of $\{x, y \mid x \in \text{attributes}, y \in \text{classes}\}$ is provided to the learner. The learner's task is to induce a set of rules that can discriminate positive examples (E+) from negative ones (E-), and thus propose a classification for new instances. For a multi-class classification problem, the *n* classes are transformed into *n* two-class problems (e.g., C_1 = E+, C_2,…,C_n = E-) and a PART rule-based machine learning technique [2] can then be applied to induce the *n* classifiers. To simplify the presentation, we assume that each classifier contains *k* positive rules. The positive rules of the initial classifier of class *i* will be denoted R_{i1}, R_{i2}, …, R_{ik}.

The new classifier is then constructed from the set of rules. The basic idea of our approach is to consider any rule R_{ij} as a potential rule for each of the new classifiers. A rule R_{ij} will be included in the new classifier of a given class if it correctly classifies the positive examples of that class. As a decision measure, we use the normalised confidence measurement as the cut-off point for rule selection. The rules of the new classifier for class C_i are all the rules that satisfy the cut-off point. Obviously, some (but not all) of the rules of the initial classifiers will be in the new classifier, as well as rules from other initial classifiers.

The proposed method has been designed to increase the sensitivity (positive coverage) of the classifiers. One should then expect the method to have a reduced specificity (also called soundness). As will be shown in the example, this approach is useful when the ratio E+/E- is very low, and when the initial classifiers yield little

sensitivity. In that case, the loss of specificity is small compared to the increase of sensitivity, yielding more useful classifiers. Obviously, for some classes the initial classifiers can be preferred to the new one.

Data set

We have applied our learning method in classifying the protein functional classes of fully sequenced *Porphyromonas gingivalis*, which is a gram-negative oral anaerobe that causes human gum disease. The initial data with protein functional classes was taken from the Oral Pathogen Sequence Databases, Los Alamos National Laboratory Bioscience Division (http://www.stdgen.lanl.gov/oragen/). The protein functional classification used in this study was based on Monica Riley's functional categories [4] which has been predominately used by TIGR. We used the 752 open reading frames (ORFs) with known functional classes as our training data. The attributes that were used in this experiment are the Grand Average of Hydropathicity (GRAVY), the percentage of every amino acid, the pI value, the net charge, the aliphatic index, the length and the number of the amino acid, and the molecular weight of the protein. These attributes were obtained from Los Alamos National Laboratory and Protparam tool (http://ca.expasy.org/tools/protparam.html).

Results and Discussion

We have performed 10-fold cross-validation on the training data and compared the performance of PART and POS-SET. The result shows that POS-SET increases the sensitivity and also the normalised positive predictive accuracy compared to PART. Although our method increases the True Positive-rate (TP-rate), as a trade-off it also increases the False Positive-rate (FP-rate). Since the objective of this study is to improve the rule coverage when classifying protein functional classes, we permit the rule-set to cover some false positives as a consequence to improving the positive coverage of classical machine learning. However, the results show that the increase of TP-rate is higher than the corresponding increase of the FP-rate. In general, POS-SET performs well in learning from a small set of positive examples compared to the negative examples. This is due to the fact that our method is capable to generating a softer boundary for the classifier and thus avoiding problems connected with the strong discriminative boundary generated by standard machine learning systems.

Another interesting finding from this experiment is that the rule sets generated from POS-SET are much smaller than those of the original PART system. We would have expected POS-SET rule-sets to contain more rules compared to PART due to "collecting" additional rules from other classifiers, but it turns out they have with increased sensitivity. We believe that these rule-sets are useful for classifying protein functional class and thus can assist wet experimental biologists in understanding the co-relationship between amino acid properties and functions.

Conclusions

We have devised a novel machine learning approach by combining sets of positive rules, which can be generated by classical machine learning techniques. We have applied this method in classifying protein functional classes of *P.gingivalis*. Our approach increases sensitivity and normalized positive predictive accuracy compared to the classical rule-based system. We believe our method performs well and is capable of generating sensitive classifiers in the functional classification problem where the E+/E- ratio is very low.

Acknowledgements
AC Tan and A Al-Shahib were supported by the studentships from University of Glasgow.

References

[1] Altschul, S.F.,et al. *Nucleic Acids Research*, 25: 3389-3402. 1997
[2] Frank, E. and Witten, I.H. In *Proceedings of the Fifteenth ICML*, pp. 144-151, 1998.
[3] Pearson, W.R. and Lipman, D.J. *Proc. Natl. Acad. Sci. USA*, 85: 2444-2448, 1988
[4] Riley, M. *Microbiol. Rev.*, 57:862-952, 1993