Learning Sets of Positive Rules of Amino Acid Properties to Classify Protein Functional Classes

UNIVERSITY of GLASGOW

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, Glasgow G12 8QQ, U.K. {actan, alshahib, drg}@brc.dcs.gla.ac.uk http://www.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

1. 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.

We propose a novel ensemble machine learning approach eKISS, that generates a classifier with a better sensitivity, while not loosing too much positive prediction accuracy.

2. Machine Learning Background

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 *L*. 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. The common approach of treating multi-class learning is to transform the *K* classes into a set of two-class problems, which is also known as one-against-others method. This approach faces one serious pitfall when learning in multi class problems: when we transform the *K* classes into *K* two-class problems, the positive examples of a class C_1 will be under-represented compared to the large number of negative examples for class $C_2,\ldots,C_K$. The presence of large amount of negative examples in the training data poses several pitfalls for classical machine learning systems.

The major problem of applying discriminative classical machine learning techniques (e.g. decision trees, artificial neural networks) in this situation is they either generate a trivial rejector classifier, which classifies everything as a negative class (due to the negative examples being the majority class); or overfits the positive examples by generating large decision trees or highly complex neural networks.

3. Research objective

The specific problem that we would like to address in this research is learning from multi-class protein functional classes of imbalanced data sets where the protein examples from one class heavily outnumber those from the other class (e.g. 1 to 5%). The goal of this work is to develop a learning system to classify multi-class problems in an imbalanced data situation. We have devised eKISS (**e**nsemble **K**nowledge for **I**mbalance **S**ample **S**ets), an ensemble learning method to tackle these types of problems. The objective of eKISS is to generate one-against-others classifiers which are capable of learning over multi-class examples under the skewed normal distribution of the training examples, as well as providing explanation to the user. We present the general framework of eKISS and describe its application to learn sets of positive rules in classifying protein functional classes of *P.gingivalis*.

4. eKISS (ensemble Knowledge for Imbalance Sample Sets) Method

In our approach, we have applied the PART rule-based machine learning technique to generate the base classifiers for our ensemble learning system. PART (Frank and Witten, 1998) is a rule-induction algorithm that avoids global optimisation, and generates accurate and compact rule sets by combining the paradigms of "divide-and-conquer" (C4.5, Quinlan, 1993) and "separate-and-conquer" (RIPPER, Cohen, 1995).

The basic idea of eKISS is to consider any rule R_{ij} as a potential candidate rule for each of the new ensemble classifiers. The main assumption made in eKISS is that all the rules generated by the PART learning algorithm represent possible classification rules, hence enlarging the search space.

The eKISS search strategy is to find all the rules that correctly classify the examples in the positive class, hence improving the coverage of the positive examples under the multi-class imbalanced data situation. We also believe these positive rules are useful for providing insights to the human expert in understanding the relationships between protein structure and sequence information compared to a trivial rejector classifier.

Technically, a rule R_{ij} will be included in the new ensemble 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, cf_norm = (TP-0.5)/(TP+FP(E+/E-)) 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.

5. Data Sets *Porphyromonas gingivalis*

P. gingivalis is a gram-negative oral anaerobe that causes human gum disease.

Initial data with protein functional classes from the Oral Pathogen Sequence Databases, Los Alamos National Laboratory Bioscience Division (http://www.stdgen.lanl.gov/oragen/).

The protein functional classification based on Monica Riley's functional categories which has been predominately used by TIGR.

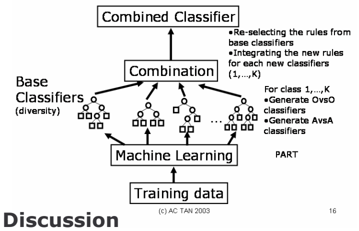
752 open reading frames (ORFs) with known functional classes

Attributes:

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.

Pbtained from Los Alamos National Laboratory and Protparam tool (http://ca.expasy.org/tools/protparam.html).

eKISS Learning System


Combined Classifier
•Re-selecting the rules from base classifiers
•Integrating the new rules for each new classifiers (1,…,K)
Combination
Base Classifiers
(diversity)
…
For class 1,…,K
•Generate OvsO classifiers
•Generate AvsA classifiers
Machine Learning
PART
Training data
(c) AC TAN 2003
16

6. Results and Discussion

We have performed ten-fold cross-validation on the training data and evaluated the test set by comparing the performance of PART and eKISS. The results show that eKISS 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 of 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. We also tested eKISS on a set of randomly generated data set, where eKISS is not performing well as expected. In general, eKISS performs well in learning from a small set of positive examples compared to the negative examples. This is due to the fact that eKISS is capable of generating a softer boundary for the classifier and thus avoiding problems connected with the strong discriminative boundary generated by classical learning systems.

An example of a set of positive rules for cellular processes functional class:

IF gravy = medium and his = low and thr = medium and pro = low and tyr = low and ser = low and leu = low and net_charge = medium OR
asn= low and numbaa= low and gln= low and gravy= high and pi= high and ala= medium OR
cys = low and his = low and leu = medium and ala = medium THEN
functional class = cellular processes

Acknowledgements: We would like to thank Gilleain Torrance for his useful comments. AC Tan & Ali Al-Shahib were funded by a University of Glasgow studentship.