Multi-class protein fold classification using a new ensemble machine learning approach

Aik Choon Tan[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, 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

Several new initiatives (e.g. structural genomics) and improvement of the methods for structure determination will result in a rapid increase in the number of structures. These high-throughput structure determination projects will produce structural data on proteins for which very little is known about their biology. Thus sophisticated computational methods are needed to detect, search for and compare remote protein homology in the hope that knowledge can be transferred to the new unknown protein (e.g. inference about function). Machine learning is one such approach that has been widely used in the development of automatic protein structure classification and prediction.

One of the aim of structural genomics is to enhance the understanding of the relationships between amino acid sequence and its corresponding protein fold. Hence, one of the advantages of using symbolic machine learning approaches for this purpose is to generate human understandable classifiers (rules) from some biological background knowledge that can explain the current proteins in the Protein Data Bank (PDB).

2. Machine Learning Background

For a supervised classification problem, a set of training data (positive and negative examples) in the form of {x, y | x $\in$ attributes, y $\in$ 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 SCOP fold imbalanced data sets where the protein examples from one class heavily outnumber thoses 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.

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.

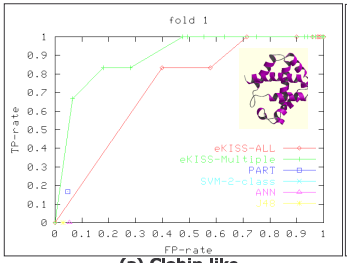
5. Data Sets (adapted from Ding and Dubchak, 2001)

SCOP 1.61 (Sept 2002) and Astral 1.61 (Sept 2002)

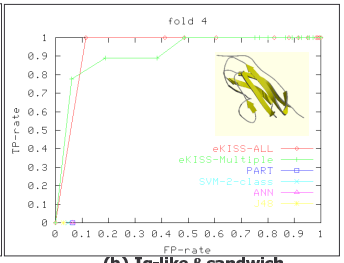
- 25 SCOP folds
- 125 amino acid physico-chemical properties
- 408 Training set
- 174 Test set

6. Results and Discussion

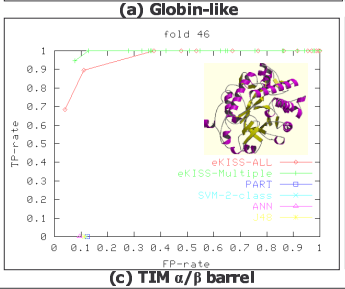
We have performed ten-fold cross-validation on the training data and evaluated the test set by comparing the performance of PART and eKISS. Table 1 summarises the performance on the training and test set. From the results, eKISS outperforms PART on 20 classes based on the F_1-measure. The results show 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.



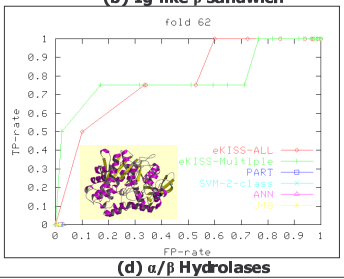
(a) Globin-like



(b) Ig-like β sandwich



(c) TIM α/β barrel



(d) α/β Hydrolases

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