Assignment 1: Database Search Results

Pengyi Yang*

1 A List of Articles Relevant to My Current Research Topic

Papers are grouped by “Review Papers”, “Evaluation Measures”, “Methodology Papers”, and “Application Papers”. Those of particular interest have been highlighted as:

- of special interest
- of outstanding interest

References

[*] Review Papers:


[*] Evaluation Measures:


*Name: Pengyi Yang; Name of Supervisor: Bingbing Zhou; Unit of Study: INFO5993; Title of Research Topic: Imbalanced Data Sampling (temporary subproject); Research Interests: Bioinformatics, Data Mining, Hybrid System; Date: 18 March 2009
[*] Methodology Papers:


[*] Application Papers:


2 Paper Review

2.1 Paper Details

Title: “Protein Classification with Imbalanced Data”
Authors: X Zhao; X Li; L Chen; K Aihara
Source: Proteins: Structure Function and Bioinformatics
Times Cited: 5 (google scholar), 3 (web of science)
ISSN: 0887-3585
DOI: 10.1002/prot.21870

2.2 Summary

To annotate a new protein, classifying it into some known protein class is often performed. This is known as the protein classification problem and it is generally a multi-class classification problem. When decomposing the problem into a binary classification problem where samples from the interested class are compared against all samples from other classes, the dataset become extremely imbalanced because the number of proteins in this class is usually much smaller than that of the proteins outside the class.
In this paper, the authors proposed to combine over-sampling with an ensemble of classifiers to overcome the imbalanced problem introduced in protein classification. The novelty of the paper is that it introduces the idea of using ensemble of classifiers with each classifier working on a sub-feature space. This is the strength of this work because the ensemble classifier helps to overcome the bias introduced by relying on a single classifier. The weaknesses of this study, however, is that the evaluation and comparison are conducted on a single measuring matrix–AUC (Area Under the ROC Curve) while many alternative measuring matrixes such as F-Measure and Geometric Mean are available and may provide radically different indications. To make the paper more solid, I would suggest to evaluate the proposed method on these alternative measuring matrixes as well and compare if the improvement is consistent.

This work is related to my research topic because it is an important application of using sampling method to alleviate imbalance problem of protein classification. Future directions could include further study of the proposed method with more evaluation matrixes and classification algorithms, and apply the method to other types of biological and medical data analysis such as microarray and clinical datasets which are also commonly highly imbalanced.

3 Top Conferences and Journals in My Research Area

3.1 Top Conferences

Tier of A+:

- ACM SIGKDD Conference On Knowledge Discovery and Data Mining (KDD) (Top 1 Data Mining conference)
- International Conference on Intelligent Systems for Molecular Biology (ISMB) (Top 1 Bioinformatic conference)

Tier of A – B:

- Annual International Conference on Research in Computational Molecular Biology (RECOMB)
- Asia Pacific Bioinformatics Conference (APBC)
- International Conference on Bioinformatics (InCoB)

3.2 Top Journals

- Bioinformatics. Impact Factor: 5.039 (Top 1 Bioinformatics Journal)
- BMC Bioinformatics. Impact Factor: 3.493
• IEEE/ACM Transactions on Computational Biology and Bioinformatics. Impact Factor: 1.803

• Machine Learning. Impact Factor: 1.742 (Top 1 Machine Learning Journal)

• IEEE Intelligent Systems. Impact Factor: 1.438

### 4 Main Research Groups in My Area

- **Sandrine Dudoit**, Associate Professor, Biostatistics and Statistics Chair and Head Graduate Advisor, Graduate Group in Biostatistics Division of Biostatistics School of Public Health University of California, Berkeley 101 Haviland Hall, 7358 Berkeley, CA 94720-7358. (Development and application of statistical and computational methods to address problems in biomedical and genomic research)

- **Xingming Zhao**, Ph.D., Institute of Systems Biology, Shanghai University, No 99, Shanga Road, Shanghai, 200444, China. (Pattern recognition, Bioinformatics, Systems Biology)

- **Nathalie Japkowicz**, Associate Professor, School of Information Technology and Engineering (SITE), University of Ottawa 800 King Edward Ave. Ottawa, Ontario, Canada, K1N 6N5. (Learning from Imbalanced Data Sets and Evaluation Methods for Machine Learning)

- **Nitesh V. Chawla**, Assistant Professor, Dept. of Computer Science & Engg. 384 Fitzpatrick Hall University of Notre Dame, IN 46556. (Data, Inference Analysis, and Learning)