

Interactive Learning for Microarray Analysis

Qi Tian¹, Yijuan Lu², Maribel Sanchez³, Yufeng Wang⁴

Keywords: self-supervised learning, DEM, relevance feedback, plasmodium falciparum

1 Introduction.

Two problems plague efforts to analyze high throughput genomic/transcriptomic/proteomic data: (i) the high dimensionality of the data, and (ii) the relatively small sample size. The above problems can be alleviated by self-supervised learning techniques, which take a hybrid of labeled and unlabeled data to train classifiers. Discriminant-EM (DEM) [2] proposed a framework for such tasks by applying self-supervised learning in an optimal discriminating subspace. Moreover, the linear DEM is extended to a nonlinear kernel algorithm, Kernel DEM (KDEM) to capture the non-linearity in the data distribution in this paper. In order to incorporate the specialists' knowledge to further improve the classification, we also propose and implement an interactive learning framework Relevance Feedback for gene classification and retrieval. Extensive experiments on the *Plasmodium falciparum* dataset show the effectiveness and promising performance of the approach.

2 Discriminant-EM and Relevance Feedback.

Discriminant-EM (DEM) [2] is a self-supervised learning algorithm by taking a small set of labeled data with a large set of unlabeled data. The basic idea is to learn discriminating features and the classifier simultaneously by inserting a multi-class linear discriminating step in the standard expectation-maximization (EM) iteration loop. DEM differs from other discriminate analysis methods such as multiple discriminant analysis (MDA) and biased discriminant analysis (BDA) in the use of unlabeled data and the symmetric/asymmetric way they treat the positive and negative examples in the discrimination step. However, the discriminating step is linear in both DEM and BDA, they have difficulty in handling nonlinearly separable data. In this paper, we generalize the DEM algorithm from linear setting to a nonlinear one, and extensive experiments are performed on the *Plasmodium falciparum* dataset for gene classification.

Relevance feedback [3], which has been successfully used in informational retrieval, is rarely used in the field of bioinformatics. In this paper, we introduce relevance feedback framework to microarray analysis. The aim is to incorporate specialists' feedback to retrain our classifier which can bridge the gap between the temporal expressions and the associated semantics. Through this procedure, specialists provides judgment on the classification result as to whether, and to what degree, genes belong to that class based on their knowledge such as Gene Ontology classification and functional annotation and our classifier is retrained continuously to achieve more and more correct classification.

¹ Department of Computer Science, University of Texas at San Antonio, TX, USA, E-mail: qitian@cs.utsa.edu

² Department of Computer Science, University of Texas at San Antonio, TX, USA, E-mail: lyijuan@cs.utsa.edu

³ Department of Biology, University of Texas at San Antonio, TX, USA, E-mail: msanchez@lonestar.utsa.edu

⁴ Department of Biology, University of Texas at San Antonio, TX, USA E-mail: YWang@utsa.edu

3 Experiments and Analysis.

In the experiment, we test various discriminant algorithms such as MDA, BDA, DEM and their kernel algorithms on *Plasmodium falciparum* microarray dataset [1]. After standard quality control filtering and normalization, a complete dataset consists of signals for 7091 oligonucleotides and training set contains 523 genes in 12 classes, including components involved in genetic information flow, metabolic pathways, cellular regulatory networks, organellar activities, and parasite-specific activities. In each class, we perform two-class classification and training set size varies from 100 to 400. Table 1 shows the classification results for class 10. In average, we can see DEM outperforms MDA, BDA and most kernel algorithms perform better than their linear algorithms. From our extensive tests on all the gene groups in the dataset, we find that the average classification error rate of all the algorithms is below 15% and our classification results are fairly consistent with the Gene Ontology classification and the classification according to temporal expression phases [1].

Through relevance feedback, we find self-supervised learning has successfully predicted a number of putative genes that may belong to a specific functional class. For example, in addition to essential enzymes (DNA-directed RNA polymerase complex), transcriptional factors such as Gas41 and Sir2 homolog and transcriptional activators may play a role in the regulation of transcription (Table 2). Besides it also offers a powerful means for an annotation feedback. For instance, two oligonucleotide probes, f23846_3 and opfh0036, both correspond to gene PF08_0034; however, the former is positively classified into Group 1, whereas the latter is negative. This discrepancy is probably due to the error in gene model.

It is worth emphasizing that the relevance feedback appears to improve learning significantly. A simple trial of correcting four ambiguous training examples (PF14_0601, PF14_0104, PF13_0178, and PF11020c) based on Gene Ontology predictions, the classification accuracy increase from 84.5% to 87.2%.

Error rate (%)	Training Dataset Size			
	100	200	300	400
MDA	3.9113	2.2243	1.9767	1.0417
BDA	7.4462	10.9559	9.3023	14.8611
DEM	3.9247	1.6544	1.3953	1.1806
KMDA	2.3522	1.8934	1.4244	1.25
KBDA	4.2876	1.8566	1.8314	1.1806
KDEM	2.0833	1.3603	1.8605	1.1111

Table 1: Comparison of linear and kernel algorithms (gene group 10)

Oligo_ID	Gene_ID	Annotation	Prob.
f22770_1	PFC0805w	DNA-directed RNA pol II	96.4%
opfc0750	PFC0805w	DNA-directed RNA pol II	77.9%
m44300_14	PF13_0152	sir2 homologue	99.6%
f21506_2	MAL8P1.131	Transcription factor Gas41 homologue	51.3%
M33088_1	MAL13P1.213	Putative transcription activator	98.2%

Table 2: Putative genes that may be involved in transcriptional machinery.

4 References.

- [1] Bozdech, Z. Llinas M., Pulliam, B. L., Wong, E. D., Zhu, J. DeRisi, J. L. 2003. The transcriptome of the intraerythrocytic development cycle of plasmodium falciparum. *Plos Biology*, 1(1), pp. 1-16.
- [2] Wu, Y., Tian, Q., and Huang, T. S. 2000. Discriminant EM algorithm with application to image retrieval, *Proc. of IEEE Conf. Computer Vision and Pattern Recognition*.
- [3] Zhou, X. and Huang, T. S. 2003. Relevance feedback in image retrieval: a comprehensive review, *ACM Multimedia Systems Journal, special issue on CBR*, 8(6), pp. 536-544.