Multiclass Classification of SRBCT Tumors

Gene Yeo

Artificial Intelligence Laboratory and
The Center for Biological and Computational Learning
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

http://www.ai.mit.edu

The Problem: There currently exists no single biological or chemical test that can precisely distinguish small, round blue cell tumors of childhood (SRBCTs) into their subclasses, which include neuroblastoma (NB), rhabdomyosarcoma (RMS), non-Hodgkin lymphoma (NHL) and the Ewing family of tumors (EWS) [2]. A recent paper by Khan et al. demonstrated that using gene expression profiles obtained from cDNA microarrays of samples which included both tumor tissue as well as cell lines, artificial neural networks (ANNs) can accurately distinguish the tumor subtypes using 96 top genes obtained from Principal Component Analysis. In order to identify candidate targets for therapy, it is important to find a smaller subset of genes and yet retain high classification accuracy.

Motivation: Khan et al. [2] used ANNs with no hidden layers as a classifier, which is a simple linear disciminant method. As such, perhaps other simple linear classification methods may give similar results, and with appropriate feature selection, fewer and more biologically salient genes can be retrieved. In an analysis similar to that of Yeang et al [4], multiclass classification is performed using 3 binary classifiers (k-nearest neighbors (kNN), weighted-voting (WV), and linear support vector machines (SVM)) in a one-versus-rest fashion with 3 feature selection techniques (Golub's Signal to Noise (SN) ratios [1], Fisher scores (FSc) and Mukherjee's SVM feature selection (SVMFS))[3], to obtain equal accuracy with fewer genes (features).

Previous Work: Murkherjee et al [3] introduced a feature selection technique for SVMs, which has been successfully applied to Golub et al's [1] AML-ALL problem. This technique is applied to this multiclass dataset, and these features are used as inputs to other classifiers (Weighted Voting, kNN) as well.

Approach: In solving multiclass problems using binary classifiers combined in a one-versus-rest fashion, each classifier trained on one class versus the rest of the classes makes a prediction about a given test sample. A sample's predicted label is the class from which that classifier returned a positive label, and that the rest of the classifiers returned negative labels. Sample predictions can be rejected in these two cases: (1) If the 4 binary classifiers each do not predict that the sample is in their respective classes. (2) If conflicting predictions arise i.e. 2 or more binary classifiers predict that the sample belongs in their respective classes. Results with rejection are indicated in the tables below by "w.r.". In the tables below, (signed) refers to using the signed SVM labels to determine the class prediction (hard errors [4]), and (max) refers to assigning the class label for which the distance from the margin in the positive direction (i.e. the direction in which that class resides rather than the "rest") is maximal. We can also determine the confidence of the prediction using the magnitudes of the SVM outputs.

Leave one out cross validation on the training dataset, using top genes ranked by the various feature selection methods were performed. In the tables below, the top X genes means X genes discriminated one class from the rest of the classes, and because there are four classes, the total number of genes used are 4 times X, or fewer, as there are overlaps in some of the gene subsets. The bracketed numbers are the Leave-One-Out Cross Validation (LOOCV) error for the 64 training samples.

kNN	kNN w.r.	WV	WV w.r.	SVM (signed)	SVM (signed) w.r.	SVM (max)
2 (4)	2 (4)	9 (19)	4 (4)	0 (2)	0 (1)	0 (1)

Table 5: Test errors (20 samples) using all genes. LOOCV errors (64 samples) in brackets. (signed) refers to signed output as predicted class, (max) refers to maximal output as predicted class. w.r. stands for "with rejections".

Impact: Small round, blue cell tumors can be easily classified into their classes by simple methods, such as Weighted voting, kNN, and linear SVM classifiers. It may be of interest to use the features selected by SVMFS

FSc	kNN	kNN w.r.	WV	WV w.r.	SVM (signed)	SVM (signed) w.r.	SVM (max)
100	0 (2)	0 (1)	3 (2)	0 (0)	0 (2)	0 (0)	**0 (0)**
60	2 (3)	0 (0)	4 (2)	0 (0)	1 (2)	0 (0)	**0 (0)**
20	2 (3)	1 (0)	2 (2)	1 (0)	2 (3)	1 (0)	**1 (0)**
10	3 (2)	1 (0)	1 (2)	1 (0)	4 (3)	1 (0)	**1 (0)**
5	4 (2)	0 (0)	6 (4)	0 (0)	4 (4)	0 (0)	**1 (0)**
S2N	kNN	kNN w.r.	WV	WV w.r.	SVM (signed)	SVM (signed) w.r.	SVM (max)
100	0 (2)	0 (1)	2 (0)	0(0)	0 (2)	0 (0)	**0 (0)**
60	0 (2)	0 (0)	1 (0)	0(0)	1 (3)	0 (0)	**0 (0)**
20	2 (2)	0 (0)	1 (1)	0(0)	3 (2)	0 (0)	**1 (0)**
10	1 (2)	1 (0)	1 (2)	1(0)	2 (3)	1 (0)	**1 (0)**
5	2 (2)	1 (0)	2 (4)	0(0)	2 (2)	0 (0)	**0 (0)**
SVMFS	kNN	kNN w.r.	WV	WV w.r.	SVM (signed)	SVM (signed) w.r.	SVM (max)
100	1 (1)	0 (1)	**0 (0)**	0 (0)	0 (1)	0 (0)	**0 (0)**
60	**0 (0)**	0 (0)	1 (0)	0 (0)	0 (1)	0 (0)	**0 (0)**
20	2 (2)	1 (0)	2 (1)	1 (0)	**1 (0)**	1 (0)	**1 (0)**
10	3 (3)	1 (1)	4 (2)	1 (0)	1 (2)	1 (0)	**1 (0)**
5	6 (5)	1 (0)	7 (6)	1 (0)	7 (4)	1 (0)	**1 (1)**

Table 6: Test errors (20 samples). LOOCV errors (64 samples) in brackets. (signed) refers to signed output as predicted class, (max) refers to maximal output as predicted class. w.r. stands for "with rejections". Left column is top number of genes ranked by FSc, S2N or SVMFS used to train one classifier (multiply by 4 to get max possible number of distinct genes).

to discriminate samples when large feature spaces are available, and go to features selected by say S2N when it is of pharmaceutical interest to pick few features. Using the maximal SVM outputs gives the best accuracy (1 low confidence error for Test 13 (non-SRBCT)) with just 20 genes (in total) ranked by their Signal to Noise ratios. Rejection of samples in this multiclass scheme is useful for reducing errors, and may indicate biological perculiarity in the rejected samples. Lastly, the top genes are candidates for further histochemical and biological validation.

Future Work: The important features by the different feature selection algorithms (including Khan et al's method) overlap to some extent. It may be of interest to explore the extent of the overlap, and the biological significance of the features selected. Combining features selected by different methods may give better results, as SVMFS improves classification accuracy for all methods using larger sets of features, but S2N and FSc helps accuracy when fewer features are used.

Research Support: Research at CBCL is sponsored by grants from: Office of Naval Research (DARPA) under contract No. N00014-00-1-0907, National Science Foundation (ITR) under contract No. IIS-0085836, National Science Foundation (KDI) under contract No. DMS-9872936, and National Science Foundation under contract No. IIS-9800032 Additional support was provided by: Central Research Institute of Electric Power Industry, Center for e-Business (MIT), Eastman Kodak Company, DaimlerChrysler AG, Compaq, Honda R&D Co., Ltd., Komatsu Ltd., Merrill-Lynch, NEC Fund, Nippon Telegraph & Telephone, Siemens Corporate Research, Inc., Toyota Motor Corporation and The Whitaker Foundation.

References:

[1] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caliguiri, C. Bloomfield, and E. Lander. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, 286:531–537, 1999.

[2] J. Khan, J. Wei, M. Ringner, L. Saal M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. Antonescu, C. Peterson, and P. Meltzer. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, 7(6):673–679, 2001.

[3] S. Mukherjee, P. Tamayo, D. Slonim, A. Verri, T. Golub, J. Mesirov, and T. Poggio. Support Vector Machine Classification of Microarray Data. AI MEMO, CBCL Paper 1677, 182, MIT Artificial Intelligence Laboratory, 1998.

[4] C. Yeang, S. Ramaswamy, P. Tamayo, S. Mukherjee, R. Rifkin, M. Angelo, M. Reich, E. Lander, J. Mesirov, and T. Golub. Molecular Classification of multiple tumor types. *Bioinformatics Suppl. 1 ISMB 2001*, 17:S316–S323, 2001.