Applying Feature Selection to Gene Expression Data: Cell-type Classification & Gene Signature Identification.

Gustavo Empinotti Susan Tu Raf Mertens

CS229

gustavoe | susanctu | rafm @stanford.edu

December 14, 2012

Abstract

DNA microarrays allow biologists to capture the expression of tens of thousands of genes in a single tissue sample. In this paper, we explore the possibility of using this gene expression data to identify gene signatures that allow for accurate classification of cells. We describe our work on two datasets, one derived from cancerous and normal breast tissue, and one derived from blood cells. For the first dataset, we attempted to use feature selection to find the genes that indicate cancerous breast tissue. This problem turned out to be easy in the sense that many sets of genes result in high classifier performance, so we concluded that we needed a more difficult classification problem in order to find a gene signature of biological significance. In the second dataset, cells were labeled by flow cytometry according to their stage in hematopoietic differentiation. For this multiclass classification problem, we identified gene signatures that give reasonably good classification results. We also report our classifier's results when run on test datasets with early progenitors and AML (Acute Myeloid Leukemia) cells, where the classification of cancerous cells is of particular interest because accurate classification would allow us to determine which types of normal cells cancerous cells developed from.

1 Introduction

The wealth of gene expression data obtained from microarray technology has triggered research on the application of computational methods –in particular those from statistics and machine learning- to biology problems. On the topic of cancer, for example, work has been done on predicting survival [1, 2], determining which genes' expression correlate with certain diseases, and classifying cancers by subtype [1, 3, 4]. Popular approaches include entropy theory, χ^2 and t-statistics for feature selection, k-nearest neighbors, naive Bayes, and support vector machines for classification [5].

We attempt to identify gene signatures for two problems: in the first dataset we try to predict whether or not the sample comes from a cancer patient. In the second, we attempt to classify a cell according to its stage of hematopoietic differentiation, a highly regulated process by which the body generates blood cells. The process is also a widely studied model for multilineage differentiation in humans [8]. This second problem involved multiclass classification, a field in which no dominant method has emerged.

For this project we used Scikit-Learn, a collection of Python implementations of common machine learning algorithms available at http://scikit-learn.org.

2 Binary Classification: Breast Cancer

2.0 Dataset & Preprocessing

In this section, we describe our attempts to use feature selection to identify a gene signature that distinguishes between cancerous and normal breast tissue. The dataset was originally obtained from the National Cancer Institute's Cancer Genome Atlas and includes expression data of 17814 genes for 599 people, of whom 533 have breast cancer and 66 do not. Data was obtained from Agilent G4502 microarrays.

We normalized the data so that the expression of each gene had mean 0 and standard deviation 1. At least one gene expression value was missing for each of 319 people. These values collectively pertained to 537 distinct genes and added up to a total of 1740 missing values. We set these values to be equal to the mean (0).

2.1 Results & Discussion

Method	Precision	Recall	# Features
Naive Bayes	99%	99%	all
Logistic Regression with l_1	99%	100%	71
SVM with χ^2	99%	100%	50
Backwards Selection	99%	100%	64
SVM with Random	91%	100%	50

Table 1: Average Precision and Recall on Breast Cancer Dataset from 4-fold Cross Validation. Number of features for logistic regression is average over 4 runs. For backwards selection we removed 25 features on each iteration due to limitations in computational resources.

As demonstrated by the high percentages in Table 1, this problem is very easy to solve. The high precision and recall resulting from picking random features suggests that any set of genes gives good predictions, which is consistent with recent literature that claims that random subsets of genes are good at predicting breast cancer survival, and sometimes even better than some published gene signatures [6] (although we did not find the latter to be the case in our classification problem). The feature selection methods in Table 1, although they all resulted in excellent classification accuracy, resulted in gene signatures that had little to no overlap. These results suggest that there are many distinct reasonably small gene signature that indicate whether breast tissue is cancerous or not.

3 Multiclass Classification: Blood Cells

3.0 Dataset & Preprocessing

This section regards our attempt to develop a classifier that categorizes cells according to their stage of hematopoietic differentiation. Labels were obtained through flow cytometry and cDNA amplification, as further described by Novershtern et al. [8]. The dataset contained data from 211 arrays. Each array had gene expression levels for 11927 genes, and each cell was classified in one of 38 categories (shown in Figure 1). Novershtern et al. grouped the 38 classes into 5 more general ones and found gene signatures for these 5 classes. Our goal was to refine this grouping, i.e., find signatures that distinguish among a broader set of classes.

Extensive preprocessing of this dataset had already been done by Professor David Dill. As in the breast cancer dataset, we normalized the data to have mean 0 and standard deviation 1. There were no missing values.

3.1 Methods

3.2 Choice of classification method

Li et al. [7] did research on multiclass classification methods when applied to microarray data. They compared a variety of combinations of feature selection and classification methods by using multiple methods on 9 different datasets. Their findings consistently pointed towards better performance of SVM when compared to J4.8 decision tree, naive Bayes and k-nearest neighbor. However, there was no clear decision as to what usage of SVM led to better results. The question remains of what generalization method (from binary to multiclass) should be used. For that reason, we only used SVM classification, and varied its parameters in search for the optimal choice. In addition to generalization method, we varied feature selection method and parameters for SVM (coefficient C of regularization term, type of kernel, class weights).

3.2.1 Feature selection and regularization

As is usually the case with microarray data, the number of features in our dataset drastically exceeded the number of training examples. Therefore we added feature selection and regularization to prevent overfitting. We used two straightforward methods: a univariate feature selection method based on χ^2 (a measure of dependence between random variables), and an l_1 regularization method that results in sparse solutions (by driving many coefficients to 0).

3.2.2 Handling multiclass classification

Multiclass classifiers fall into roughly two types: generalizations of binary classifiers versus repeated application of binary classifiers as they are. Our multiclass classifier consists of repeated application of binary classifiers. We tried three ways of building a multiclass classifier out of binary ones: one-vs-all, one-vs-one and Error-Correcting Output Coding (ECOC). In one-vs-all, a binary classifier is built for each category, each one having the positive class as that category and the negative class as the union of the rest. In one-vs-one, a binary classifier is built for each pair of classes, ignoring all the remaining categories. In ECOC, a sequence of binary classifiers are built; for each one of them, a randomly chosen subset of the categories is the positive class, and the union of the remaining categories is the negative class. In all of them, the predicted category for a new example is, briefly, the category that more closely agrees with the multiple classifications. For more precise descriptions of these generalization methods, see Li et al. [7].

3.3 Results & Discussion

3.3.1 Cross-Validation

Initially we repeatedly ran SVM with all 38 classes and we varied the kernel, C, and the feature selection and generalization methods. When trying ECOC, we also varied the code-length. We found that the optimal choice was using SVMs with a linear kernel, C = 125, equal weighting of the classes, l_1 regularization (which automatically determines the number of features) in a one-vs-all scheme

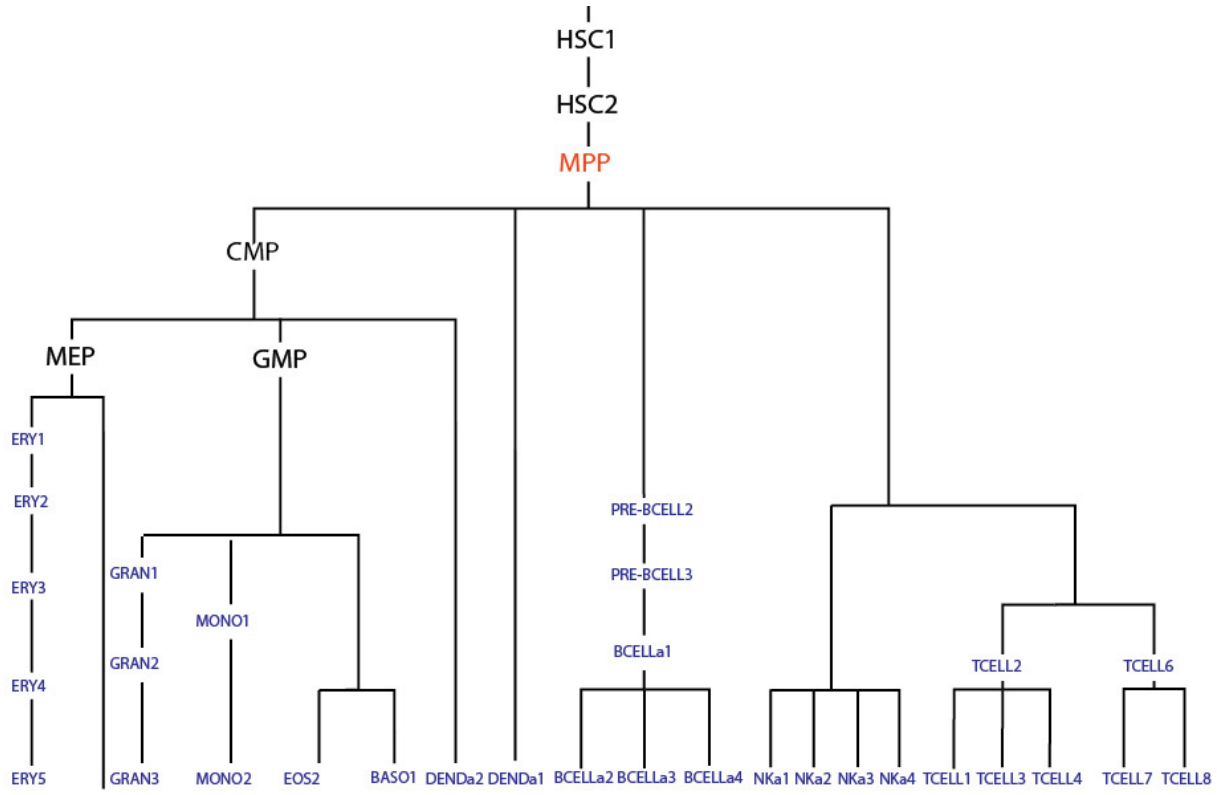

Figure 1: Hematopoietic Cell Differentiation. Black indicates early progenitors that were in our 211-array training set. Pink indicates early progenitors that were present only in a test set. Blue indicates all other cells present in our training set. HSC: hematopoietic stem cell; MPP: late multipotent progenitor; CMP: common myeloid progenitor; MEP: megakaryocyte/erythroid progenitor; ERY: erythrocyte; MEGA: megakaryocytes; GMP: granulocyte/monocyte progenitor; GRAN: granulocyte; MONO: monocyte; EOS2: eosinophil; BASO: basophil; DEND: dendritic; PRE-BCELL: early or pro-B cell; BCELL: naive, mature, able to switch, or switched B-cell; NK: natural killer; TCELL: T-cells.

(see figure 2 for comparison with other methods). This led to 78.6% accuracy with leave-one-out cross-validation. Here we observed that the most common misclassifications were ERY2, ERY3 and ERY4 mapping to each other. This reflects their biological similarity, so we turned these 3 categories into a single one (we assume that the information in our dataset is not enough to distinguish them). We did the same with MEGA1 and MEGA2. This results in 35 classes. Doing so raised the accuracy to 86.2% (see figure 3 for the confusion matrix). Li et al. [7] reported an accuracy no higher than 68% for their dataset with 14 classes, the highest number of classes they worked with.

3.3.2 Adjusting class-weights

In the standard SVM framework each class is weighted equally. Since we are worse at classifying cells that are infrequent in our training set, we tried adjusting weights on the penalty term of the SVM to be inversely proportional to class frequencies. This would cause our classifier to trade off confidence on most classes with improved classification of the rare classes. We thought that such a trade-off could result in overall better performance since our classifier did very well on the prevalent classes. We found that adjusting class weights decreased our leave-one-out cross-validation performance by 1% .

3.3.3 Test Set

We then tried to use our classifier on a test set. On two 6-array test sets, we were able to identify granulocytes correctly with our best-performing (as determined through cross-validation on the 211-array data set) one-vs-one SVM. We also attempted to classify a 27-array test set of early progenitors by using a two-stage classification. We classified all the cells using our best one-vs-all SVM. Then, we selected out all the cells classified as any one of the five early progenitors (HSC1, HSC2, MEP, CMP, GMP). We used a one-vs-all SVM trained on only the early progenitor cells to reclassify these cells classified as early progenitors. We opted for a two-layer approach because our initial classifier was particularly deficient in classifying early progenitors. Eight of those 27 cells were MPPs, a cell type not present in the 211-array test set (and thus not part of our 35 classes). We classified those 8 as either HSC1 or CMP, which are respectively 1 position and 2 positions away in the tree. Of the remaining 19 cells, 10 were classified within 1 position from its real location in

the tree. One cell was classified 2 positions away, and 8 were more than 2 positions away.

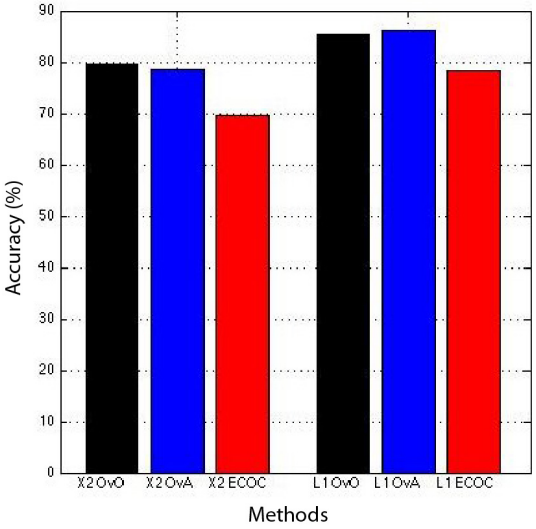

Figure 2: LOOCV Accuracy of Various Classification Methods. Using C=125, selecting 900 features with χ^2 for 1 vs 1 and 1 vs All, selecting 330 features with χ^2 for ECOC.

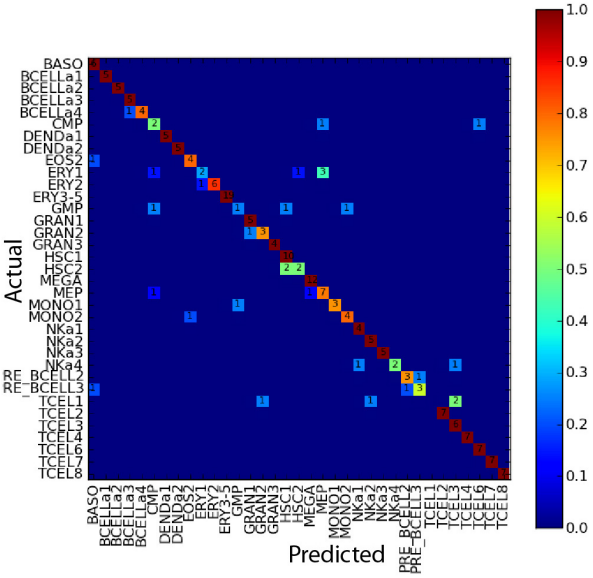

Figure 3: Confusion Matrix for 35 Classes, 1 vs All SVM with l_1 Regularization and C=125. The [0,1] on the color scale is the fraction of samples of the actual class that are classified as each of the predicted classes.

Novershtern et al. [8] group the 38 classes into 5 larger classes in order to get gene signatures. Since our classifier consistently classifies 21 classes with no false negatives, we were able to retrieve gene signatures for all of these by looking at the genes picked by feature selection by the respective classifier (one for each cell, as in one-vs-all). Some of the gene expressions in our gene signatures were as far as 2 standard deviations away from the mean (standard deviation and mean computed over all classes). The size of the gene signatures (determined by l_1 regularization) ranges from 8 to 53 genes and averages to 26. Random sets of 30 genes give an average 30% accuracy on leave-one-out cross-validation (against 86.2% for our selected genes), which gives credibility to these signatures. The following classes are predicted with no false negatives (and hence provide reasonable gene signatures): BASO, BCELLa1, BCELLa2, BCELLa3, BCELLa4, DENDa1, DENDa2, ERY3-5 (a single cluster), GRAN1, GRAN3, HSC1, MEGA1-2 (a single cluster), NKa1, NKa2, NKa3, TCEL2, TCEL3, TCEL4, TCEL5, TCEL6, TCEL7, TCEL8. See table 2 for an example of a gene signature.

Coefficient	Gene	Expression
0.58	C18orf1	0.65
0.25	SLC43A1	0.04
0.18	MFSD7	0.91
0.12	ADRA2A	1.45
0.08	PHGDH	-0.178
0.08	PARG	0.205
0.06	ZBTB1	1.32
0.05	TRAJ28	0.67

Table 2: Gene signature for BCELLa2

The following have a particularly high number of false negatives: CMP, ERY1, GMP, TCEL1.

3.4 Future Work

One of the most important problems to be addressed in future work is that of classifying cancerous cells into their stage of differentiation. This is very difficult because some genes have a distorted expression in cancerous cells.

Our work also shows the importance of collecting more microarray data, because our classifier does very well with the classes for which we have more training examples. This gives hope that obtaining more microarray data, which is currently expensive, will allow for increasingly accurate classifiers and gene signatures.

On a related note, our 35-class classifier is ineffective at distinguishing MEP, CMP, GMP and ERY1 from one another. Future work could restrict its attention to these cells.

Microarray data is susceptible to errors (in purifying, for example). One indication of this is that one of the cells labeled as GMP consistently gets mapped to T-cells and these 2 classes are very far from each other in the differentiation process. This seems not to be just a problem with the classifier since this mistake persisted even when our classifier correctly classified all the remaining GMPs. Therefore, it might be useful to apply statistical methods to identify outliers in this dataset (and make this process standard for microarray data, since they are error-prone) and ignore them.

4 Acknowledgements

Thank you to Professor David Dill (Stanford CS) for his valuable guidance and for providing us with preprocessed and ready-to-use gene expression data. Thank you to Professor Ravi Majeti (Stanford Cancer Center) for the AML test sets, to professor Stephen Boyd (Stanford EE) for a discussion on regularization, and to Yifei Meng (Computational Biology '15) for help in evaluating test set results.

References

[1] Haibe-Kains, B., Desmedt, C., Piette, F., Buyse, M., Cardoso, F., van't Veer, L., & Sotiriou, C. Comparison of prognostic gene expression signatures for breast cancer. BMC Genomics, 9:1-9. 2008. doi:10.1186/1471-2164-9-394

[2] Buyse, M., Loi, S., van't Veer, L., Viale, G., Delorenzi, M., Glas, A. M., & ... Piccart, M. J. Validation and Clinical Utility of a 70-Gene Prognostic Signature for Women With Node-Negative Breast Cancer. JNCI: Journal Of The National Cancer Institute, 98(17):1183-1192, 2006. doi:10.1093/jnci/djj329

[3] Sotiriou, C., & Pusztai, L. Gene-Expression Signatures in Breast Cancer. New England Journal Of Medicine, 360(8): 790-800, 2009. doi:10.1056/NEJMra0801289

[4] Abraham, G., Kowalczyk, A., Loi, S., Haviv, I., & Zobel, J. Prediction of breast cancer prognosis using geneset statistics provides signature stability and biological context. BMC Bioinformatics,11:277-291, 2010. doi:10.1186/1471-2105-11-277

[5] Liu, H., Li, J., & Wong, L. A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. Genome Informatics. International Conference On Genome Informatics, 13:51-60, 2002.

[6] Venet, D., Dumont, J. E., & Detours, V. Most Random Gene Expression Signatures Are Significantly Associated with Breast Cancer Outcome. Plos Computational Biology, 7(10):1-8, 2011. doi:10.1371/journal.pcbi.1002240

[7] Li, T., Zhang, C. & Ogihata, M. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. Bioinformatics 20(15): 2429-2437, 2004 doi:10.1093/bioinformatics/bth267

[8] Novershtern, N., Subramanian, A., Lawton, L. N., Mak, R. H., Haining, W.N., McConkey, M.E., Habib, N., Yosef, N., Chang, C.Y., Shay, T., Frampton, G.M., Drake, A.C., Leskov, I., Nilsson, B., Preffer, F., Dombkowski, D., Evans, J.W., Liefeld, T., Smutko, J.S., Chen, J., Friedman, N., Young, R.A., Golub, T.R., Regev, A., & Ebert, B.L. Densely Interconnected Transcriptional Circuits Control Cell States in Human Hematopoiesis. Cell 144: 296-309, 2011. doi:10.1016/j.cell.2011.01.004