Assigning B cell Maturity in Pediatric Leukemia

Gabi Fragiadakis[1], Jamie Irvine[2]
[1]Microbiology and Immunology, [2]Computer Science

Abstract

One method for analyzing pediatric B cell leukemia is to categorize each malignant cell based on its resemblance to one of 7 stages in B cell development. Researchers hand label ("gate") B cell maturity by drawing separate gates per patient. However, this process is both time consuming and must be done patient-by-patient due to the heterogeneity of each cancer. We aimed to automate this gating process with one classification algorithm for all patients. We found that a Crammer SVM algorithm classifies cells with 97% accuracy, indicating that information from other dimensions of the data can compensate for patient specific differences. This revealed that we can reliably apply a consistent framework to a diverse set of patients.

Background

A struggle in cancer cell biology is the heterogeneity of the disease: each cell in a cancer sample is different, as is each patient. Therefore amidst this chaos of cancer it is challenging to define a structure that could provide insight into cancer development, progression, and outcome.

B cell acute lymphoblastic leukemia (ALL) is a hematologic cancer where B cells (the antibody-producing cells of the immune system) become mutated and excessively proliferative during development. Current work in Dr. Garry Nolan's lab involves looking at protein expression on single cells from bone marrow from B cell ALL patients. Cells in a given sample are stained with metal-conjugated antibodies specific for certain antigens or proteins. Using a technique called mass cytometry (or CyTOF), the time of flight of the ions conjugated to those antibodies provides a readout of the levels of each protein on each cell[1]. This outputs a matrix of ~10^6 cells by ~40 parameters (proteins) that we have measured.

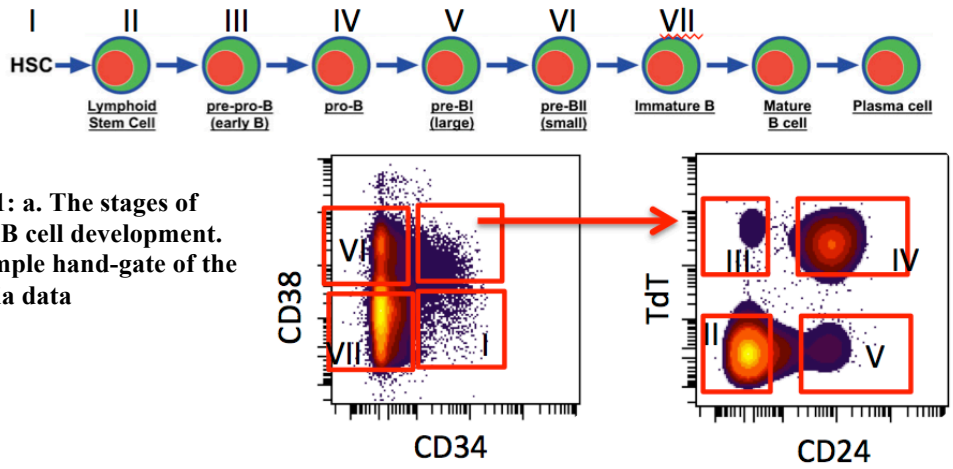

Figure 1: a. The stages of normal B cell development. b. A sample hand-gate of the leukemia data

Work in our lab has found a potential structure for understanding B cell ALL. Kara Davis and Sean Bendall found that healthy B cell development can be broken down into a progression of seven stages based on the expression of certain sets of markers (proteins) on each B cell (figure 1a). Using this definition of healthy development, Dr. Davis imposed this framework onto the

ALL samples, assigning each malignant B cell a "maturity score" based on its resemblance of that given developmental stage (figure 1b). This assignment revealed that each sample has a different composition of cells from each stage in B cell development (i.e patient A has mostly B cells in stage 1 and 2, whereas patient B has mostly B cells in stages 3 and 5 for instance). We have on the order of 10^5 cells from nine patients that have been assignment a maturity score by the researcher who has gated the data by hand.

Statement of the Problem

Assigning maturity to the cells within the cancer may provide insight into the structure and progression of the tumor, perhaps having implications for patient outcome. When hand-gating, the researcher takes into account 4 of the 33 dimensions to separate the populations, drawing a new personalized gate for each patient. However, it remains to be seen how well one classifier can perform on cells from different patients, as there is a large amount of person-to-person variation in protein expression. The goal of this project is to automate B cell maturity gating, which is time-consuming for the researcher, as well as determine how well we can impose a common framework to a diverse set of patients.

It was clear that a subset of the cells could be correctly assigned with high confidence because they have an unambiguous phenotype that matches "normal" B cell development. For cells near the stage boundaries, some degree of misclassification is unavoidable due to the unique cellular makeup of that patient. However, we hypothesized that some features we have measured on those cells could help classify them better, allowing us to determine what fraction of cells we can classify with confidence to get the distribution of B cell maturity within a given patient.

Methods and Results

Pre-processing the Data

The ion pulses from the mass cytometer are extracted as cell events and written to a .fcs file format. The data were then visualized and hand gated, first removing debris and non-B cells, by Kara Davis using the analysis software at cytobank.org. We then exported a separate .fcs file for each population for each patient and converted them to a data matrix for learning in MATLAB.

Models and Justification

For this project we needed to consider multiclass classification algorithms, as we have seven populations. We decided to test three multi-class algorithms in the LIBLINEAR package[2]:

- Softmax: the multi-class extension of logistic regression
- One vs. All Support Vector Machine: a set of binary SVM classifiers
- Crammer multi-class SVM[3]: an all-in-one SVM classifier

These algorithms were selected based on the analysis of Aly et al and Hsu et al on the performance of multi-class classifiers[4,5]. For the SVM algorithms, we used a linear kernel because it performed well and the Gaussian kernel was too computationally expensive for the size of our data.

In addition, we wanted to contrast these discriminative algorithms with a generative algorithm. Protein expression in populations of healthy cells is often considered to be normally distributed (though this may not be the case in cancer due to abnormal protein expression). Therefore we

implemented Gaussian Discriminant Analysis, modeling each population as a multivariate Gaussian (each with its own mean and variance).

Model Selection

We compared the four aforementioned algorithms using k-fold cross validation. We used a special form of k-fold cross validation where we left out all the cells from one patient each time rather than a random subset of the pooled cells. We used this method to accurately simulate evaluating cells from a new patient that was not used to train the model. This method reveals the variance of a model's performance patient to patient (i.e. how well we can trust performance on a given patient rather than the average performance). It was also informative in analyzing whether any of our patients was a strong outlier.

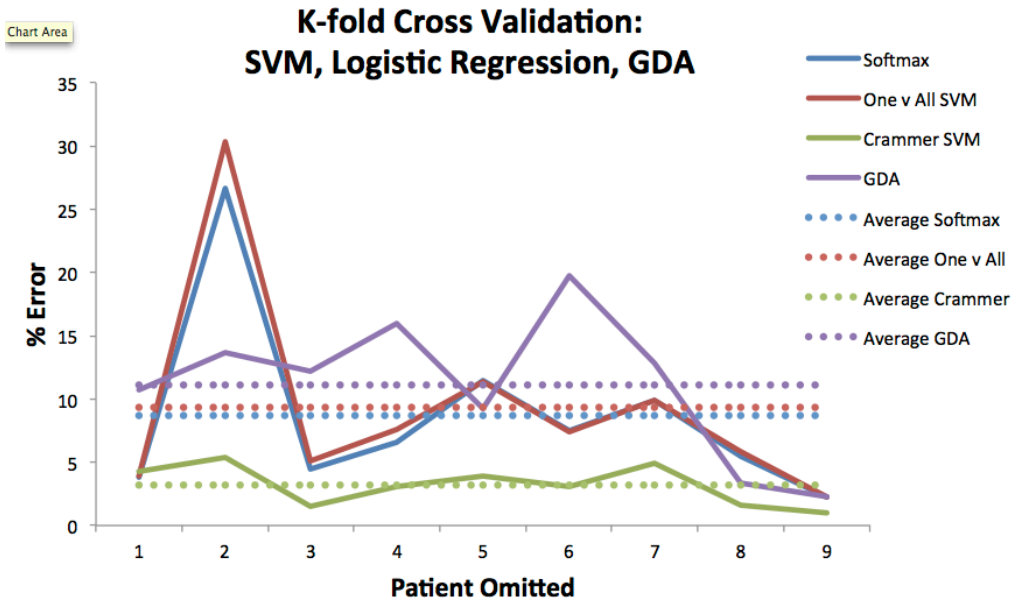

Figure 2: K-fold Cross Validation over four algorithms

As plotted in figure 2, the Crammer SVM performed best with 97.49% average accuracy. It also performed the most consistently across patients. In addition, we varied the cost parameter C on each of our SVM and softmax models as described in Hsu et al[4] and saw little change in model performance (figure 3a).

For a full analysis of the success of the Crammer SVM classifier, we generated a confusion matrix from the predictions to extract precision and recall values for each population. As shown in figure 3b, precision and recall were high across all populations, increasing our confidence in our model, never dropping below 92%.

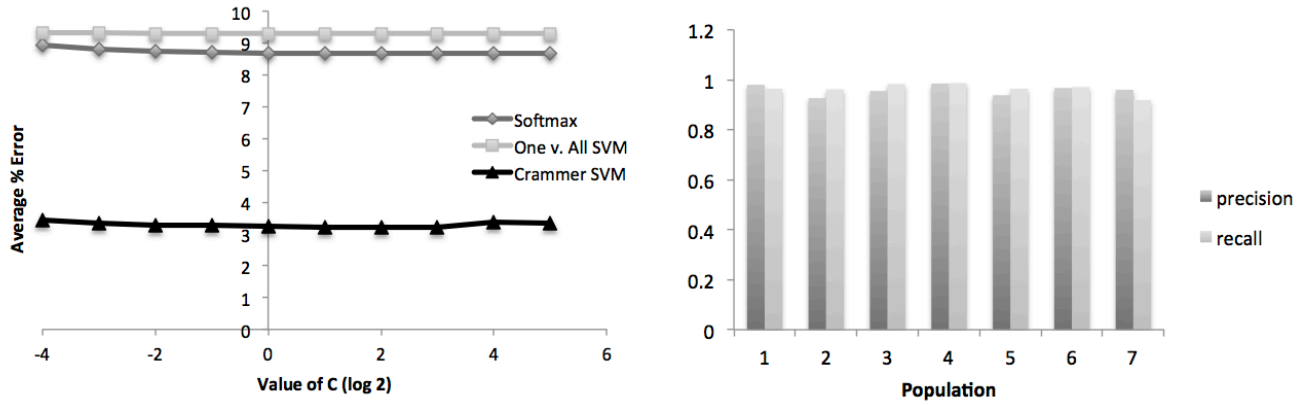

Figure 3 a. varying cost parameter; b. precision and recall for each population

Feature Selection

We were interested in running a feature selection algorithm to gain insight into which features our best classification algorithm was using. We ran a forward search to select the top ten features and, as expected, our top four features were those used by the researcher to initially gate the data: CD38, TdT, CD24, and CD34 (figure 4). Those features alone classify the data with 95% accuracy. However, taking into account additional features (CD79b, CD7, CD43, CD49d, and HLA-DR) increased accuracy by 2.5%. This indicates that other dimensions of the data can help the algorithm generalize to more than one patient (though it performs quite well looking only at the first four).

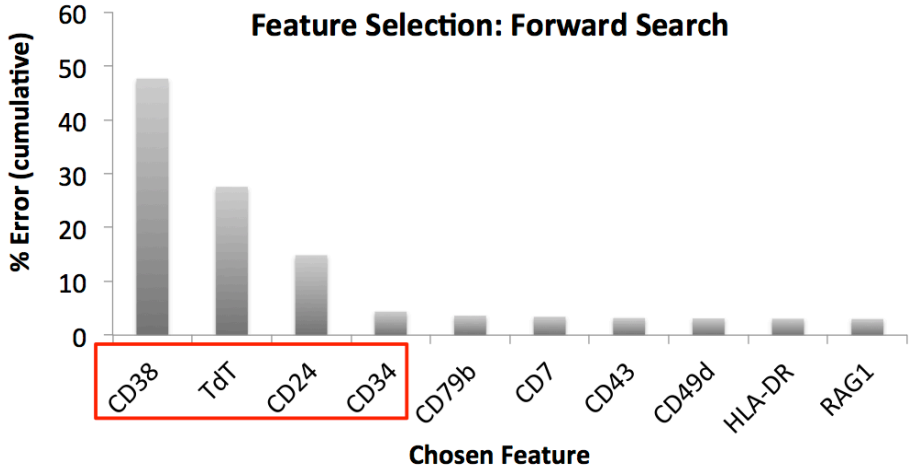

Figure 4: feature selection using forward search (features used for hand gate highlighted in red)

Discussion

In all, we found the accuracy, precision, and recall were significantly better for Crammer SVM than any of the other models, yielding an average accuracy of 97.49%. This model, along with the other SVM models, worked nearly the same with all choices of the cost parameter. As expected, most of the accuracy could be attained using only the four features used by the

researcher in hand-gating. However, our classifier achieved a notable gain when using more features.

The Crammer SVM model was predictably the most successful SVM because it directly optimizes the multiclass classification problem rather than approximating it with a one-vs-all method for each class. It was slower, but still classified in a reasonable time on our data. Gaussian Discriminant Analysis performed surprisingly poorly. In general, the distributions of cells within a stage of development are often considered to be Gaussian over the parameters that we considered. However, due to the abnormality of the cancer cells and the fact that their forms only approximate these stages of normal cells, the distributions are more abnormal and this likely made the Gaussian model too strong of an assumption.

Our results from feature selection were promising. The features beyond the four used by the researcher increased the accuracy of our classifier, telling us that it was working intelligently. These extra features, which do not directly relate to which stage of development a B-cell is in, help the classifier adjust the thresholds in the primary four dimensions differently for each patient. For example, a patient with particularly large cells would have a higher measure in all parameters. Thus, higher values in extra features inform the classifier to adjust the classification threshold for the primary features. This allows our classifier to work generally across patients with different cell profiles, something that a human researcher would have to do case by case.

These results are very exciting for the researcher and she plans to use this algorithm in future work. She will be analyzing many more patients (n = 100) and she now has a method for gating the developmental populations efficiently and consistently. Furthermore, she will have matched clinical data that includes treatment response, disease progression, and outcome. We will continue working with her to build predictive models for clinical data using the maturity scores across patients.

Works Cited

1. Bendall, S. C. et al. Single-cell Mass Cytometry of Differential Immune and Drug Responses Across a Human Hematopoietic Continuum. Science 332, 687–696 (2011).

2. R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A Library for Large Linear Classification, Journal of Machine Learning Research 9(2008), 1871-1874. Software available at http://www.csie.ntu.edu.tw/~cjlin/liblinear

3. K. Crammer and Y. Singer. On the Algorithmic Implementation of Multi-class SVMs, JMLR, 2001.

4. Mohamed Aly. Survey on Multiclass Classification Methods. November 2005. Trans. Neural Netw. 13, 2, 415–425.

5. Hsu, C.-W. and Lin, C.-J. 2002a. A Comparison of Methods for Multi-class Support Vector Machines. IEEE

A special thanks to Kara Davis for providing data, discussions, and a solvable problem.