Critical Review 2: DeepCyTOF

Ben Kotopka, Ali Sharafat, Archa Jain, David Liu

November 30, 2016

1 Problem Statement

Flow cytometry is an important technique in medicine and the biological sciences in which a population of cells is analyzed by quickly collecting data on each of millions of individual cells. In most flow cytometry approaches, cells pass one at a time through a set of lasers and optical detectors, which measure how light passes through the cell and detect any fluorescent dyes or proteins in the cell or on its surface. Fluorescent dyes conjugated to antibodies can bind and label target receptors on cell surfaces, allowing subpopulations of cells within the sample to be identified. However, overlap between the emission spectra of different dyes imposes an upper bound of about 15 on the number of fluorescent channels that can be interrogated at once.

Mass cytometry, or CyTOF, is a more recently developed alternative to traditional flow cytometry which labels antibodies with heavy metal ions instead of fluorescent dyes. Instead of measuring fluorescence from each cell, CyTOF instruments vaporize cells and direct the resulting plasma into a mass spectrometer, which sensitively measures levels of each heavy metal ion for each cell. In contrast to conventional flow cytometry, current CyTOF instruments measure up to 40 channels at once. This allows cell populations to be characterized in much more detail than was previously possible.

The object of most flow cytometry and CyTOF experiments is to distinguish subpopulations of cells in the original sample. The process of defining values of measured parameters that correspond to subpopulations of interest is called gating, and is typically done manually. However, manual gating makes comparing data collected on different instruments difficult, making collaborations between research groups and replication studies challenging. This is a particular problem for CyTOF data owing to its high dimensionality. As an alternative to manual gating, Li *et al.* describe a deep learning strategy for automated gating in CyTOF experiments.

2 Related Work

Traditionally, flow cytometry data processing relies on manual gating. The FlowCAP-I challenge is the first in a line of challenges to produce automatic methods for cell classification. The paper presented by Li *et al.* compares its classification accuracy to those of the competitors in this challenge. The competition presents four challenges: Challenge 1: completely automated gating algorithms for exploratory analysis. Challenge 2: semiautomated gating algorithms with manually adjusted parameters tuned for individual data sets. Challenge 3: algorithms for cases in which the number of expected populations was known. Challenge 4: supervised approaches trained using

human-provided gates, with 25% of the manual gates for each data set provided to participants for training and tuning their algorithms. Note that Li *et al.* present an algorithm for Challenge 4 [1].

The submitted algorithms were evaluated using the F-score (harmonic mean of precision and recall). The approaches used but the algorithms were fairly diverse, across supervised and unsupervised learning. For challenge 4, the winning algorithm RadialSVM used supervised training of radial SVMs using example manual gates, with other successful approaches using t mixture modeling and entropy-based merging, decision forests, and ensemble clustering.

3 Method

The authors used five FCM datasets from FlowCAP-I, two CyTOF datasets generated in the Montgomery Lab and 50 simulated sample generated using a single manually gated CyTOF datasets to conduct their experiments. They logarithmically transform and rescaled the data before using them as input to the model.

The basic neural network unit used in this paper is a stacked autoencoder. The authors trained each autoencoder from bottom-up in an iterative fashion. The first layer of the autoencoder is trained to reconstruct the input data, once this layer is trained, its output representation is used to train the following layer, this process continues until the last layer. For each stacked autoencoder, the authors chose to use three fully connected hidden layers with sigmoid nonlinearity, the number of hidden nodes in each layer has been set to 12, 6, and 3.

Using stacked autoencoder as the basic building block, the authors developed DeepCyTOF, a deep learning domain adaptation model. The DeepCyTOF model is trained on data from a source domain with a given distribution and applied to target domain with related but different distributions. The biological and technical variations in the FCM and CyTOF experiments made automatic gating very challenging, and it is desirable to avoid gating each dataset separately; this is the main motivation for employing the domain adaptation approach.

To create training data for the DeepCyTOF autoencoder units, the authors divided the N samples in the baseline condition into one source reference sample, $N-1$ target samples and $N-1$ mixture samples (50% of the target sample cell mixed with 50% of the reference sample cell). This data division can be represented in a generalized star-like topology as shown in figure 1. To select the most suitable candidate for reference sample, the authors first computed a $d \times d$ covariance matrix for each of the N samples (d is the dimensionality of each dataset associated with these samples), then for each pair of samples, they computed the Frobenius norm of the difference between the covariance matrix; the sample with the smallest average distance to all other samples is selected as the reference sample.

The authors trained one autoencoder using the reference sample, $N-1$ autoencoders using target samples, and $N-1$ autoencoders using mixture samples. The resulting $1+2(N-1)$ autoencoders is combined into a single large neural net, where two stacked autoencoders in each branch of the star structure in figure 1 are connected, and each branch is also connected to the reference autoencoder. A softmax layer is added on top of the combined structure as the final prediction layer. This final DeepCyTOF structure is then fine-tuned using labeled data obtained by manual gating from subject in the reference sample only.

Figure 1: A star graph representing the sample division. Autoencoders are constructed for the reference sample (r), target samples shown in the outer circle ($i = 2, \ldots, N$), and mixture samples shown in the inner circle

4 Discussion

The authors train their model on one of the 5 datasets from the 4th challenge of the FlowCAP-I competition and benchmark their results against the competition's winners for each dataset. It should be noted that only 25% of the data points are labeled from the training data set and the rest are unlabeled. The presented model matches or outperforms the winners by an accuracy margin of up to 6%. This is impressive since each dataset corresponds to one cell type and the semi-supervised training is done on only one of the cell types. This is an indicator that the model is robust for other cell types not covered in this paper. The authors also benchmark the performance of their model to whether or not the input data is normalized and they find that its F-score is close to 1 regardless of normalization. The only criticism is that they only compare their model against a vanilla softmax regression, which is very accurate itself with an F-score of 0.97.

The authors also benchmark their model against noisy inputs by adding simulated inputs that is noisy in one dimension and essentially using a denoising auto-encoder. They do not mention the SNR of the added noise or why they choose to add noise in one dimension only. Despite this, they demonstrate that their model is robust against noise by achieving 95% accuracy.

5 Conclusion and Future Work

The presented work is a novel approach to classification of CyTOF data and is the first deep learning approach to this area. While the results are impressive, it seems that some design choices were not justified or fully explained. For instance, the use of sigmoid function as nonlinearity is arbitrary when ReLU has generally been the norm in neural nets. Also, there has been no mention of hyperparameter optimization. The architecture of the net and the tuning parameters are either arbitrary or missing. The simulated noisy data set lacks justification as well since the SNR is missing and there is no reasoning behind why noise is only added in one dimension. For future work, it may be interesting to try a variational auto encoders as well.

References

[1] Nima Aghaeepour, Greg Finak, Holger Hoos, Tim R Mosmann, Ryan Brinkman, Raphael Gottardo, Richard H Scheuermann, FlowCAP Consortium, DREAM Consortium, et al. "Critical assessment of automated flow cytometry data analysis techniques". *Nature methods*, 10(3):228?238, 2013.