Automated Cell Classification of Mass Cytometry Data by Deep Learning and Domain Adaptation: A Review for CS273B

Amr Mohamed [1], Wisam Reid [2], and Irán Román [2]

I. Overview

The authors propose a novel and automated approach to analize Flow Cytometry (FCM) and Mass Cytometry (CyTOF) data. Current methods for the identification and classification of raw data from FCM and CyTOF consist of sequences of manual steps carried out by humans [1]. Thus, the slow pipelines and the variability among different individuals carrying out these sequences of steps result in a data-labeling process that hinders large-scale studies in a field that is rich in data. Motivated by recent advances classifying big-datasets using deep neural networks [2], the authors develop a deep learning framework to extract features from their data of interest. Their approach is further motivated by advances in the biosciences using deep neural networks [3], where the number of samples is often smaller than the number of variables. When dealing with FCM or CyTOF data, however, the number of samples is five or six orders of magnitude larger than the number of variables. Thus using deep neural networks to automatically classify and extract features from CyTOF and FCM datasets emerges as a viable alternative to current manual methods.

The authors identify an important problem and provide a rationale that attempts to justify their method of choice as an alternative to manually processing CyTOF and FCM data. However, their motivations and interpretations of the methods are a bit lacking and the paper can benefit from a more rigorous and extended discussion. Justifying deep learning as the tool of choice by referencing the availability of a big dataset does not withstand scrutiny. More specifically, the authors fail to mention why the problem could not have been solved by other machine learning or clustering techniques. Additionally, the neural network architectures that they propose consist of simple feedforward neural networks. A better approach for the extraction of features could have been a convolutional network that allows them to identify abstract features through the analysis of the convolutional filters after training. Another major drawback of their approach is that they train and assess the performance of their neural network by comparing with manually-labeled data, which they claim is variable and could contain misleading results due to human error. Finally, the authors only compare the performance of their neural network architectures against other supervised methods, thus avoiding comparison with other state-of-the art unsupervised methods.

II. Dataset

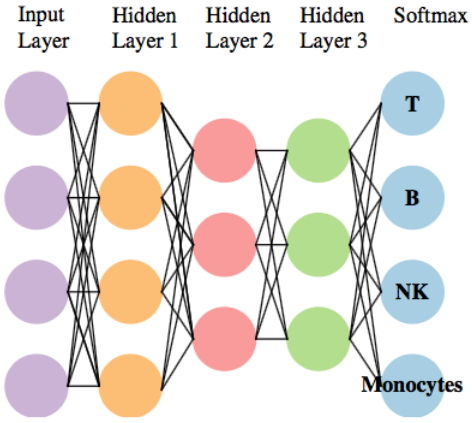
Their dataset consists of FCM and CyTOF data. The FCM data is an aggregated collection of five datasets associated with five different phenotypes: B-cell lymphoma, West Nile virus (WNV), healthy individuals, Hematopoietic stem cell transplant, and Graft-versus-host disease. CyTOF data consists of blood cells from healthy individuals and individuals affected by WNV. After CyTOF, data from cells could correspond to six different classes: unlabeled cells, B cells, CD4+ T cells, CD8+ T cells, Monocytes, and Natural killer cells. Because CyTOF requires the instrumentation to be calibrated everytime it is used, and because calibration can vary between readouts, the researchers also simulated CyTOF data capturing the variability associated with different calibrations of CyTOF tools. All data was preprocessed by applying a

[1]Department of Computer Science
[2]Center for Computer Research in Music and Acoustics

logarithmic transformation and a rescaling to fit between the values of zero and one.

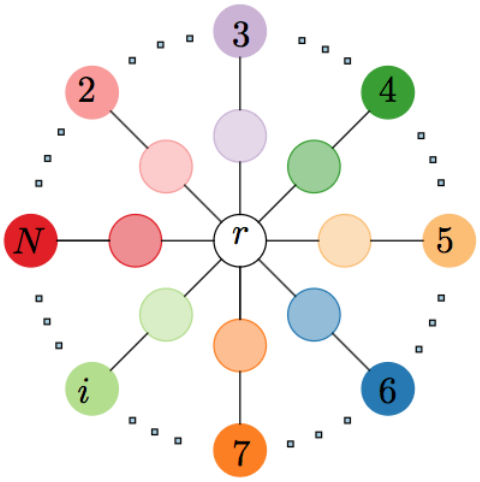
III. Methods

The authors use stacked autoencoders. Stacked autoencoders can be thought about as non-linear dimensionality reduction methods. They learn low dimensional representations of a higher dimensional input and reconstruct the input (decoding) from the representations. The authors use a model consisting of three fully connected hidden layers with sigmoid non- linearity. They add a softmax regression layer, a popular choice, on top. Also, they fine-tune the entire neural net using standard backpropagation. The number of hidden nodes in each layer is set to 12, 6, and 3. Here is a picture of a stacked autoencoder:



The autoencoder in the figure can be used for gating cell populations. The model consists of 3 hidden layers and a final softmax layer with four classes: T cells, B, NK, and Monocytes.

Not only do the authors exploit stacked autoencoders, but they also utilize domain adaptations which when trained on data from a source domain with a given distribution, can be applied to a target domain with a related but not equivalent distribution. They use a generalized version of this approach with intermediate datasets, known as DLID [4]. The model uses one source distribution (reference sample) and multiple target distributions (the remaining samples). In this generalization, the autoencoders of the source reference sample, target samples, and mixtures of the target samples with the reference sample can be represented in a generalized star-like topology. Here is a figure of the star graph representing multiple autoencoders with the outer nodes encoding the samples in the study:



The authors combine these $1 + 2(N - 1)$ autoencoders to a single large neural net, add a softmax layer on top and fine-tune the net using labeled data obtained by manual gating from subject r only. During the fine-tuning step, the two stacked autoencoders in each branch of the star are connected; each branch is also connected to the stacked autoencoder of the reference sample, so that the star graph functions as a single network, with a classifier as its upper layer.

A. Criticism

It is unclear how the authors arrived at the architecture of each stacked autoencoder and parameters. They mention that their architecture worked well in practice but they say nothing about how they found it. Did they do grid search? linear search over certain key hyperparameters? or was it a lucky shot in the dark? The authors should explain.

Regarding using domain adaptations, the authors did not establish that the domain and target distributions are strongly related which is crucial for the model to work well [5]. Finally, in the original DLID paper, the approach discussed is formulated for the case of two distributions. The

authors generalized DLID to an arbitrary number of datasets without any formal justification of why it would generalize well. To the readers and the community, it is uncertain whether the generalized DLID used by the authors works as intended.

IV. Results

The authors are successful at using the stacked autoencoder on FCM and CyTOF data. They achieve a slightly better accuracy than the winners of the FlowCAP-I competition, which is a competition for FCM data classification. Similarly, their DeepCyTOF architecture only marginally improves performance over softmax. They do a good job of analyzing whether data pre-processing has an effect on their performancs and determine that it does not. Finally, as a good sanity check the authors successfully trained their model on new data with simulated calibration variance. Performance of DeepCyTOF was similar to the original data, but softmax also had comparable performance.

A. Criticism

While the authors did improve on previous methods, their results were only slightly better. Additionally, they did not sufficiently examine why their autoencoder method outperforms the different methods used to classify cells in the competition. Thus, it remains inconclusive that their model is superior and does not overfit to the specific setting and datasets used. Regarding DeepCyTOF, it is not clear whether a neural network architecture is justified as the optimal method due to the lack of rigorous comparisons with other methods. Finally, the authors do not explain to what degree their simulated data captured the variance in hand labeled data.

V. CONCLUSIONS

As cytometry analyses become widely used in research and clinical settings, automated solutions for analyzing the high dimensional datasets are urgently needed. Current practice in which samples are first subjected to manual gating are slowly substituted by automatic gating methods. The authors successfully apply bleeding edge machine learning techniques to solve the problem of cell gating automation. They present DeepCyTOF as a promising tool and show slight improvements over other state-of-the-art methods. However, the authors do not carry out an exhaustive search for the optimal hyper-parameters and do not experiment with different architectures that may perform better. Also, they generalized DLID without establishing that the generalization works as intended. These drawbacks raise a natural question: Could the model be improved by varying autoencoder architecture? Could it be improved by using a different topology instead of the star graph? Investigating these questions would be a needed first-order extension of the paper.

References

[1] Chris P Verschoor, Alina Lelic, Jonathan L Bramson, and Dawn ME Bowdish. An introduction to automated flow cytometry gating tools and their implementation. *Frontiers in immunology*, 6, 2015.

[2] Li Deng and Dong Yu. Deep learning. *Signal Processing*, 7:3–4, 2014.

[3] Michael KK Leung, Hui Yuan Xiong, Leo J Lee, and Brendan J Frey. Deep learning of the tissue-regulated splicing code. *Bioinformatics*, 30(12):i121–i129, 2014.

[4] Sumit Chopra, Suhrid Balakrishnan, and Raghuraman Gopalan. Dlid: Deep learning for domain adaptation by interpolating between domains. *ICML workshop on challenges in representation learning*, 2, 2013.

[5] Hal Daume III and Daniel Marcu. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, page 101126, 2006.