

SWIFT: Scalable Weighted Iterative Flow-clustering Technique

Iftexhar Naim¹, Gaurav Sharma^{1,3}, Suprakash Datta⁴, James S. Cavanaugh², Jyh-Chiang E. Wang², Jonathan A. Rebhahn², Sally A. Quataert², and Tim R. Mosmann²

¹Department of Electrical and Computer Engineering, ²David H. Smith Center for Vaccine Biology and Immunology, ³Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY 14627, USA

⁴Department of Computer Science and Engineering, York University, Toronto, ON, M3J 1P3, Canada

Recent advances in data generation platforms and staining reagents for flow cytometry (FC) result in massive datasets containing high-dimensional measurements for millions of cells. The large size, dimensionality, and overlapping nature of cell populations pose a significant challenge to the traditional manual data analysis via 'manual gating' and highlight the need for automated, objective, multivariate clustering. Several statistical model-based flow-clustering methods have been proposed, but none of them are scalable to large FC datasets due to their high computational complexity. We propose a scalable clustering framework SWIFT based on weighted iterative sampling (see [Naim 2010]) that scales existing statistical model-based clustering methods to large FC datasets and allows detection of small populations that are frequently of interest. The proposed clustering algorithm SWIFT has three stages. In the first stage, SWIFT introduces a novel weighted iterative sampling framework for efficient Gaussian mixture modeling by the Expectation-Maximization (EM) algorithm. We show that the proposed weighted iterative sampling, not only increases scalability, but also facilitates better resolution of small clusters. In the second stage (after mixture model fitting), SWIFT examines each of the Gaussian clusters and splits any clusters that are bimodal along any dimensions or principal components. This stage is often crucial for high-dimensional clustering where small discrimination in any one dimension can be obfuscated by close similarity in other dimensions. Finally in the third stage, a graph-based merging is applied to fuse strongly overlapping Gaussians based on a normalized overlap measure and an entropy-based stopping criterion. This allows the method to represent skewed clusters frequently seen in FC data. A major contribution of SWIFT is its scalability to larger datasets. SWIFT provides significant speed-up for model-based clustering methods. Moreover, SWIFT shows excellent results in resolving overlapping and small populations. We designed experiments to generate large high-dimensional data with ground truth and performed validation of clustering results. Currently we are working to improve the stability and reproducibility of model-based clustering methods that will facilitate robust inference across multiple datasets.

Supported in part by a CEIS (NYSTAR) award, by a NSERC Discovery grant, and by a NIH Grant R24 A1054953

References:

Naim, I., Datta, S., Sharma, G., Cavanaugh, J. S., and Mosmann, T. R. (2010), SWIFT: Scalable weighted iterative sampling for flow cytometry clustering, in Proc. IEEE Intl. Conference on Acoustics, Speech and Signal Processing, Dallas, TX, 16-19 Mar, 2010, pp.509-512.