Proceedings of the Twenty-Seventh Conference on Innovative Applications of Artificial Intelligence

Time-Varying Clusters in Large-Scale Flow Cytometry

Jeremy Hyrkas and **Daniel Halperin** and **Bill Howe**
University of Washington

Abstract

Flow cytometers measure the optical properties of particles to classify microbes. Recent innovations have allowed oceanographers to collect flow cytometry data continuously during research cruises, leading to an explosion of data and new challenges for the classification task. The massive scale, time-varying underlying populations, and noisy measurements motivate the development of new classification methods. We describe the problem, the data, and some preliminary results demonstrating the difficulty with conventional methods.

Flow Cytometry in Science

A *flow cytometer* is an instrument used in the biological sciences to measure cell properties such as size and composition. A flow cytometer measures cells in solution by advancing cells single file through a thin capillary and analyzing its optical properties using laser light. The scattering patterns of the laser can be used to infer cell size, and the wavelengths of light the cell absorbs and re-emits are used to infer the concentration of various pigments that correspond to biological structures of interest.

Flow cytometry was developed for use in medical applications, most commonly for immunophenotyping. The usual operating conditions include a cytometer set up in a clean laboratory room with distilled water and carefully curated samples. As cytometry samples became more voluminous, manual classification by direct inspection of scatter plots ("gating") became a bottleneck and began to be replaced by automated learning algorithms (Aghaeepour et al. 2013).

In the environmental sciences, researchers increasingly use flow cytometry to analyze the microbial populations present in samples of air, water, or soil. Biological and climate scientists are particularly interested in the dynamics of phytoplankton, which produce about half of the oxygen on earth and are foundational to the oceanic food web.

Challenge: Continuous Measurement of Time-Varying Populations

The SeaFlow instrument, a new environmental flow cytometer developed at the University of Washington (Swalwell, Ribalet, and Armbrust 2011), is designed to be deployed on an oceanographic research vessel and operated continuously during multi-week cruises as the vessel moves through different environmental conditions and therefore different microbial populations. Ocean water flows continuously through the instrument, generating a time series of the population breakdown of small phytoplankton. Obtaining this data depends on the ability to classify the particles based on their optical properties.

Each particle is measured by its forward scatter at different orientations (which indicate particle size), fluourescence associated with the protein phycoerythrin (orange light), and two different wavelengths of fluorescence associated with chlorophyll (green light). Certain species of microbes exhibit different properties in different environmental conditions; they may tend to grow or shrink, or produce more or less fluourescent proteins. As a result, the location and shape of the clusters in this 6-dimensional space will vary over time as the vessel moves through different water masses.

The instrument records data in three-minute windows consisting of millions of particles each. After several weeks, there are thousands of windows to be classified. In practice, scientists apply quality control filters to eliminate noisy observations; however, even after heavy filtering this can still lead to hundreds of GB of data to classify. Moreover, we are currently working with a single instrument, but the project includes plans to deploy instruments on a significant number of vessels simultaneously, exacerbating the scale problem; as a result, automated learning methods are required for classification.

There are two obvious naïve choices for formulating a learning problem: clustering the entire dataset as one unit, or clustering each window as a separate dataset. In the first case, the problem is that the vessel passes through different water masses with very different properties; the prior populations being sampled are known to be different, so a "global" clustering approach will perform poorly and not be particularly meaningful. At the other extreme, restricting the clustering to short windows ignores opportunities to model populations that persist across multiple windows.

The uncertainty associated with these measurements is also very high relative to traditional flow cytometry applications, given the vessel's movement, the systematics and calibration issues of the instrument itself, and the diversity

of organisms and particlate matter in the water.

Overall, issues of scale, time-variance, and noise together make this use case an interesting challenge problem for the AI community. Variants of the problem include unsupervised learning (identifying and tracking the population clusters as the evolve and change over time), supervised learning (using manually-gated clusters as labeled data), and semi-supervised learning (using a small, high-quality labeled dataset and a large unlabeled dataset to improve accuracy). We also are considering formulating a structured learning problem, where the layout of the clusters in 6D space is learned directly from the set of windows rather than from the classification of individual particles.

Artificial Intelligence Solutions

In the past decade, there has been a rich literature on automatic clustering for flow cytometry methods. A recent literature review (Aghaeepour et al. 2013) examines 77 algorithms, running the gamut from parameter-less to highly tuned, from heuristic-based to statistically-sophisticated, and from unsupervised learning to supervised learning.

The study, called the FlowCAP project, showed that there are a number of automated methods that perform reasonably well on some medical cytometry applications. In the case of unsupervised classification, a handful of algorithms perform well against manual labels. Some – such as ADICyt (Adinis s.r.o.), which performs best on all datasets– have prohibitively slow runtimes without special hardware. flowMeans (Aghaeepour et al. 2011) performed second-best overall and has a reasonable running time; multiple other algorithms also performed accurately and quickly. FlowCAP shows that in many medical cytometry applications, there are a handful of automated clustering algorithms that give results of very high quality in relation to manual labeling.

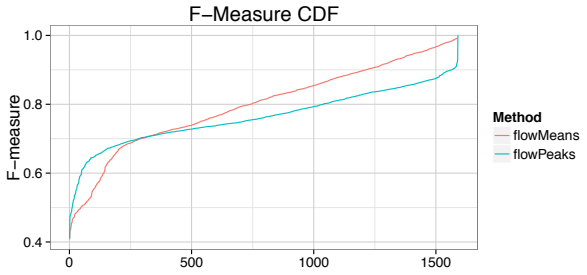

Figure 1: Cumalative Density Function of the F-measures on a SeaFlow data set using flowMeans(Aghaeepour et al. 2011) and flowPeaks(Ge and Sealfon 2012). The F-Measure is the quality metric used in the FlowCAP project based on recall and precision; it is always between 0 and 1, with 1 being a perfect score. These F-measures are obtained by applying each algorithm to each window independently and comparing against expert labeling. Both flowMeans and flowPeaks achieved average F-measures of above .9 in the FlowCAP experiments; they score much lower here.

All of these algorithms are intended to process one sample at a time; the time-varying nature of the SeaFlow data is not necessarily compatible. For example, Figure 1 shows the cumulative density function of the F-measures of two algorithms that performed well in medical settings, but perform much more poorly on SeaFlow data. Furthermore, none of these algorithms can be deployed in data-parallel, distributed environments that are important for working with 100+ GB datasets from even a single instrument and crucial for future plans involving processing data from many instruments and vessels simultaneously.

(Demers et al. 1992) explored using Gaussian Mixture Models on ocean environmental cytometry; it was posited that related samples could be analyzed as one data set to better classify populations. However, to our knowledge, such a method has never been applied on an application such as SeaFlow cytometry data.

Available Data

The University of Washington has 17 SeaFlow datasets available. The datasets all include millions of optimally aligned particle data sets measured in six dimensional space and classification for each particle. Three data sets (Thompson cruises 0, 1, and 11) are highly curated and provide the best approximation for ground truth. These three cruises are relatively small and have been made collected and made available for this challenge(Armbrust Lab).

References

ADICyt. http://www.adinis.sk/en/products/bioinformatics-and-data-processing/adicyt.html.

Aghaeepour, N.; Nikolic, R.; Hoos, H. H.; and Brinkman, R. R. 2011. Rapid cell population identification in flow cytometry data. *Cytometry Part A* 79(1):6–13.

Aghaeepour, N.; Finak, G.; Hoos, H.; Mosmann, T. R.; Brinkman, R.; Gottardo, R.; Scheuermann, R. H.; Consortium, F.; Consortium, D.; et al. 2013. Critical assessment of automated flow cytometry data analysis techniques. *Nature methods*.

Armbrust Lab. IAAI SeaFlow Challenge Data Sets. https://github.com/jhyrkas/iaai-challenge-data.

Demers, S.; Kim, J.; Legendre, P.; and Legendre, L. 1992. Analyzing multivariate flow cytometric data in aquatic sciences. *Cytometry* 13(3):291–298.

Ge, Y., and Sealfon, S. C. 2012. flowPeaks: a fast unsupervised clustering for flow cytometry data via K-means and density peak finding. *Bioinformatics* 28(15):2052–2058.

Swalwell, J.; Ribalet, F.; and Armbrust, E. 2011. SeaFlow: A novel underway flow-cytometer for continuous observations of phytoplankton in the ocean. *Limnology & Oceanography Methods* 9:466–477.