Model-free Expectation Maximization for Divisive Hierarchical Clustering of Multicolor Flow Cytometry Data

Başak Esin Köktürk , Bilge Karaçalı
Department of Electrical and Electronics Engineering
İzmir Institute of Technology
İzmir, Turkey
Email:basakkokturk,bilge@iyte.edu.tr

***Abstract*—This paper proposes a new method for automated clustering of high dimensional datasets. The method is based on a recursive binary division strategy that successively divides an original dataset into distinct clusters. Each binary division is carried out using a model-free expectation maximization scheme that exploits the posterior probability computation capability of the quasi-supervised learning algorithm. The divisions are carried out until a division cost exceeds an adaptively determined limit. Experiment results on synthetic as well as real multi-color flow cytometry datasets showed that the proposed method can accurately capture the prominent clusters without requiring any knowledge on the number of clusters or their distribution models.**

I. Introduction

Flow cytometry (FCM) is a powerful laser-based multi parametric analysis technique for characterizing individual cells within a heterogeneous population. It measures the physical, chemical and biological characteristics of each cell and uses them for cell counting, sorting and biomarker detection. FCM is used in research labs to distinguish different cell types from each other as well as in clinical labs for disease diagnosis and monitoring disease progression following therapy [1]. In FCM experiments, cells are incubated with fluorochrome-conjugated antibodies. The characteristic of the emitted light from the cells under laser excitation then allows assessing the relative abundance of the targeted biomarkers in each cell. Several biomarkers can be investigated simultaneously in multi-color flow cytometry experiments by increasing the number of fluorochromes. Currently, the FCM technology allows investigating cells for the presence and abundance of up to 20 biomarkers [2].

The increase on the number of parameters generates complex high dimensional datasets. Analysing these high dimensional datasets using standard gating methods that rely on operator-drawn regions on two-dimensional scatter plots is laborious and time-consuming. Furthermore, there are concerns over the reproducibility of the results even by the same group on the same flow data [3]. Consequently, there is a considerable demand for automated methods to address these challenges, particularly for multi-color flow data analysis.

One of the primary objectives in computational analysis of flow cytometry data is automated identification of cell subsets. To this end, several methods have been proposed in the literature to model cell population characteristics. Pyne et al. developed a direct multivariate finite mixture modelling approach that fits skew and heavy-tailed distributions to cell subpopulations in high-dimensional FCM data [4], while Aghaeepour et al. proposed an automated cell identification method based on k-means clustering that can capture concave cell populations using multiple clusters [1]. The FlowClust algorithm proposed by Lo et al. fits a t-mixture model following a Box-Cox transformation on multicolored FCM data [3]. Finak et al. modified the FlowClust algorithm by introducing a merging step once all the subpopulations are identified to check for unwarranted cluster divisions [5]. Most of the clustering methods in FCM data analysis applications use one of Bayesian information criteria (BIC), Akaike information criteria (AIC) or entropy to determine the number of clusters. This means that the clustering algorithm is run several times for varying number of clusters and the clustering result that achieve the optimal separation according to the criterion of choice is presented as the final output.

In FCM data analysis across different flow datasets, the identification of analogous cell groups is carried out also through gating. There are several supervised and unsupervised algorithms proposed in the literature for automated gating of FCM data. Supervised algorithms [6], [7] are problematic because they need a training dataset that must be created by an expert for a specific system configuration, which is not necessarily applicable for datasets collected under a different configuration, due to variations in cell preparation and flow instrument parameters. Unsupervised techniques include variations of the mixture modelling approach [8], [9], model-based clustering [10], [11] and density-based clustering [4]. Compared to the supervised methods, the unsupervised methods can offer greater automation as they do not require a universally-valid training dataset. On the other hand, their performance is hindered when the assumed models fail to match the specifics of the flow dataset at hand.

In this paper, we propose a model-free divisive procedure for automated identification of cell subgroups in multicolor flow datasets. The proposed method starts by dividing the whole dataset into two groups using an expectation maximization procedure that relies on a model-free calculation of the group posterior probabilities. The method then continues to divide the cell subgroups obtained by previous divisions until a stopping condition that detects superfluous divisions is met

978-1-4799-5669-2/14/$31.00 ©2014 IEEE

expressed through a division cost. This allows cell subgroup identification without making assumptions on the shape of cell distributions, and deduces the number of prevalent cell subgroups adaptively from the flow dataset.

This paper is organized as follows. The mathematical description of the proposed method is presented in Section 2. The results of the proposed method on synthetic datasets as well as a comparative benchmark performance evaluation on real flow cytometry datasets are presented in Section 3. Concluding remarks are presented in Section 4.

II. Methods

In this section, we first describe the quasi-supervised learning algorithm that allows a model-free estimation of group posterior probabilities at each sample [12]. We then review the expectation maximization algorithm as described by Dempster and Schafer [13], [14] followed by the proposed modification that replaces model-based posterior probabilities with those obtained using the quasi-supervised learning algorithm. The section concludes with a detailed description of the proposed model-free automated cell subpopulation identification method for multi-color flow cytometry datasets.

A. Posterior Probability Estimation Using the Quasi-supervised Learning Algorithm

The quasi-supervised learning algorithm exploits an asymptotic property of nearest neighbor classification over randomly chosen reference sets. Given a reference set R containing labeled points representing the classes C_0 and C_1, a nearest neighbour classifier $F(x;R)$ for an unknown sample $x \in X$ is defined by

$$F(x;R) = y^\star \tag{1}$$

with $y^\star$ representing the class label of the point $x^\star$ providing

$$d(x,x^\star) = \min_{x\prime \in R} d(x,x\prime) \tag{2}$$

and $d(.,.)$ denoting the metric on X. Now, letting $\mathbf{R_n}$ denote the random variable of such reference sets containing n points from each class with a probability density function $p_{\mathbf{R_n}}(R_n)$, it can be shown that for sufficiently large n, the posterior probability $P(C_0|x)$ of the class C_0 at x is approximately equal to the expected value of $F(x;R_n)$ over R_n,

$$P(C_0|x) \simeq \int_{R_n} \mathbf{1}(F(x;R_n)=0) p_{\mathbf{R_n}}(R_n) dR_n \tag{3}$$

where the indicator function $\mathbf{1}(\cdot)$ returns 1 when its argument holds, and zero otherwise [12]. In practice, the expectation integral above cannot be carried out because the probability density $p_{\mathbf{R}}(R)$ is unknown. What is available, however, are data points $\{x_i, y_i\}$ with $x_i \in X$ and their respective class labels $y_i \in \{0,1\}$ for classes C_0 and C_1, respectively, for $i = 1,2,\ldots,\ell$. Then, the expectation integral above can be approximated by the average number of times x is assigned to C_0 using all distinct reference sets R_n that can be constructed from the available samples, via

$$P(C_0|x) \simeq f_0(x) = \frac{1}{M} \sum_{R_n \subset \{x_i,y_i\}} \mathbf{1}(F(x|R_n)=0) \tag{4}$$

where M denotes the number of distinct reference sets that can be constructed using the dataset $\{x_i, y_i\}$ containing n points of each class. The posterior probability $P(C_1|x)$ can also be written in a similar fashion as

$$P(C_1|x) \simeq f_1(x) = \frac{1}{M} \sum_{R_n \subset \{x_i,y_i\}} \mathbf{1}(F(x|R_n)=1) \quad . \tag{5}$$

The quasi-supervised learning algorithm computes the averages above using a practical algorithm that avoids carrying out M separate nearest neighbor classifications. Furthermore, the ratio of $f_0(x)$ and $f_1(x)$ taken to the natural logarithm approximates the log likelihood ratio of classes C_0 and C_1 at x via

$$L(x) = \frac{p(C_0|x)}{p(C_1|x)} \simeq \frac{f_0(x)}{f_1(x)} \tag{6}$$

since the class priors are set to be equal to 0.5 due to the presence of an equal number of C_0 and C_1 points in the reference sets R_n. Finally, the optimal number of points n_{opt} to be included in the reference sets for the best learning is determined adaptively from the available data [12].

Posterior probability estimation using the quasi-supervised learning algorithm is illustrated in Figure 1. The datasets under consideration consist of points drawn randomly from two Gaussian distributions (upper row). Results show that the posterior probability estimates match the theoretical values around the well populated regions for varying dataset sizes, while the accuracy declines over the less populated regions, due to the lack of adequate representations of the underlying probability distributions with fewer points (lower row).

B. Expectation-Maximization Algorithm

In this work, we combine the expectation-maximization (EM) algorithm with the posterior probability estimation method described above and design a non-parametric model-free expectation-maximization algorithm. The expectation maximization algorithm aims to fit a distribution model to a specified dataset following an iterative procedure [13], [14], [15]. Briefly, given a set of observed data points x_i, $i = 1,2,\ldots,\ell$, assumed to be drawn from a mixture model, such as a mixture of k Gaussians, the EM algorithm estimates the parameters of each component.

Let θ_j be the parameter for the j'th component, that can be defined as

$$\theta_j = (\mu_j, \Sigma_j)$$

if the Components are Gaussian, for $j = 1,2,\ldots,k$. The objective, then, is to determine the distribution parameters θ_j. The likelihood function for each θ_j over an observation space can be expressed as

$$l_x(\theta_j; x_1, x_2, \ldots, x_l) = f(x_1, x_2, \ldots, x_\ell|\theta_j) = \prod_{i=1}^{\ell} f(x_i|\theta_j) \tag{7}$$

since the points are assumed to have been drawn independently. The maximum-likelihood estimate of θ_j is then given by θ that maximizes the likelihood function above,

$$\theta_j^{ML} = \arg\max_{\theta} l_x(\theta) \tag{8}$$

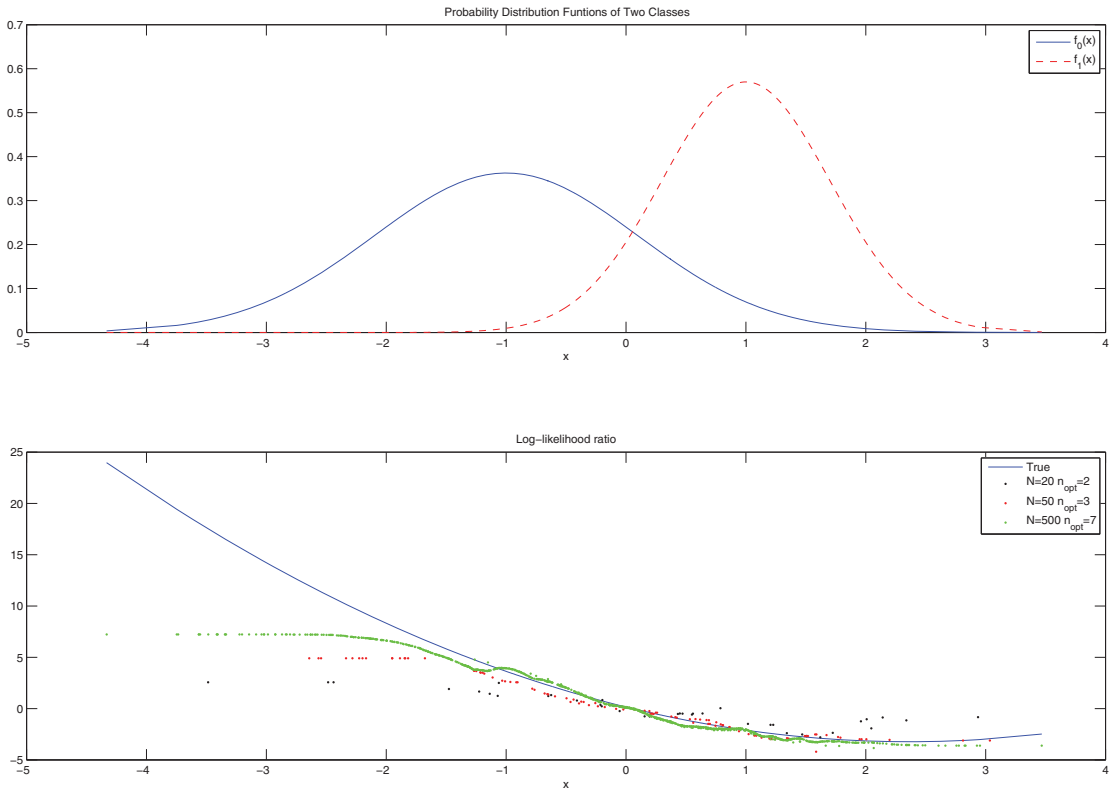

Fig. 1. True and Estimated Results of Log-likelihood Ratio for Different Sample Size .

or equivalently,

$$\theta_j^{ML} = \arg\max_{\theta} \log l_x(\theta) \tag{9}$$

as the natural logarithm function is monotonically increasing and maximizing the likelihood is equivalent to maximizing the log-likelihood.

At the expectation step, for each x_i, the method calculates a responsibility value $r_{i,j}$ defined by

$$r_{i,j} = \frac{p(x_i, \theta_j)}{\sum_{m=1}^{k} p(x_i|\theta_m)} \tag{10}$$

that expresses the likelihood of the i'th point to belong to the j'th component. The parameters θ_j are then revised in the subsequent maximization step using a maximum likelihood procedure that takes the responsibility values into account. A notable distinction between different expectation maximization procedures arise from the use of the responsibility values in the maximization step: In one alternative, the responsibility values can be used to associate each x_i with only one component by seeking the component achieving the maximum among $\{r(i,1), r(i,2), \ldots, r(i,k)\}$ for each i. In the other alternative, the parameters θ_j are estimated in a way to make the estimation uses all points simultaneously, but in a way to be influenced more by the points x_i for which $r(i,j)$ are larger and less by the others.

C. The Proposed Divisive Binary Clustering Method

The proposed method begins with an initial random assignment of points into two clusters C_0 and C_1, followed by an expectation maximization cycle that first computes the posterior probability of each of the two components at every point, and re-assigns the points to the cluster whose posterior is larger. This process is repeated until convergence.

First step : Expectation step
The posterior probabilities of each class C_0 or C_1 are computed at each point using the model-free posterior probability estimation method.

Second step : Maximization step
The class labels of each point is updated according to the maximum a posteriori classification rule via

$$\begin{aligned} C_0 &\leftarrow \{x|f_0(x) \geq 0.5\} \\ C_1 &\leftarrow \{x|f_1(x) < 0.5\} \end{aligned} \tag{11}$$

Note that the procedure above produces two distinct clusters starting with a single one, regardless of whether the resulting clusters are distinct enough to merit separation. In order to determine the distinctness of the resulting clusters, we have defined a division cost $c(C_0, C_1)$ to be calculated using

$$c(C_0, C_1) = \frac{1}{N_0} \sum_{x_i \in C_0} f_1(x_i) + \frac{1}{N_1} \sum_{x_i \in C_1} f_0(x_i) \tag{12}$$

with N_0 and N_1 denoting number of samples assigned to clusters C_0 and C_1 respectively. In this paper, we have treated the division cost as the criterion for accepting or rejecting the obtained clustering, with the rejection acting as the stopping condition for any further division of the original cluster. To this end, we have compared the division cost $c(C_0, C_1)$ with the division cost derived from the clustering that produced the parent cluster $C_0 \bigcup C_1$. The block diagram of the proposed clustering method is shown in Figure 2.

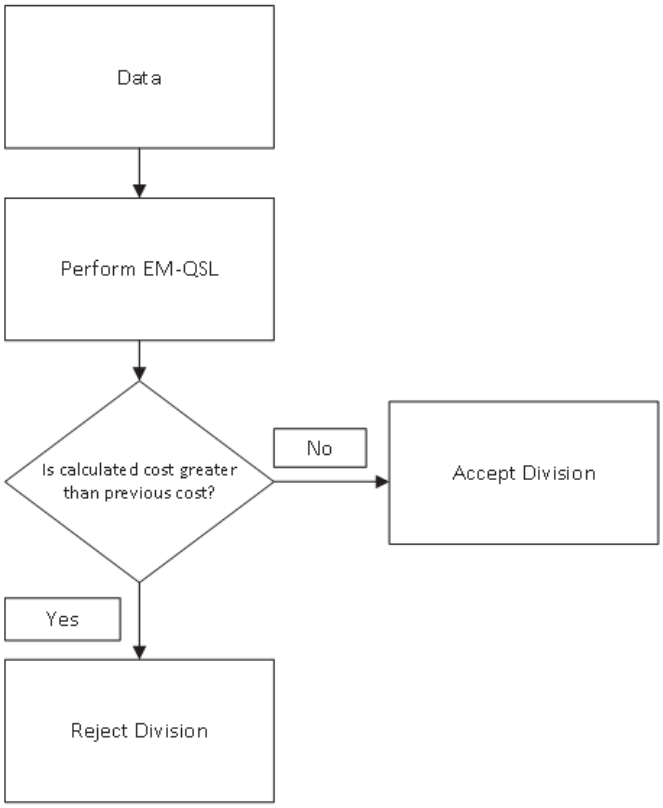

Fig. 2. Block Diagram of the Proposed Method.

Once all the binary divisions are finalized, we have used a post-processing step to check whether the union of any two of the resulting clusters forms a single coherent cluster. To that end, we have calculated the division cost between all resulting cluster pairs, and merged the clusters for which the division cost is larger than all previously accepted costs.

RESULTS

The proposed method was applied to synthetically generated datasets as well as datasets acquired from real multi-color flow cytometry experiments. The synthetic dataset contained three distinct clusters, each modeled using a two-dimensional Gaussian distribution with unit covariances but with different means, set at $[4\ 8]^T$, $[4\ 4]^T$ and $[8\ 4]^T$, respectively. The experiments consisted of generating a dataset of points drawn from this mixture with different priors and carrying out automated clustering using the proposed method as well as the conventional expectation maximization routine for Gaussian mixture fitting within the same binary division scheme for estimating the posterior probabilities from a model-based perspective (Figure 3).

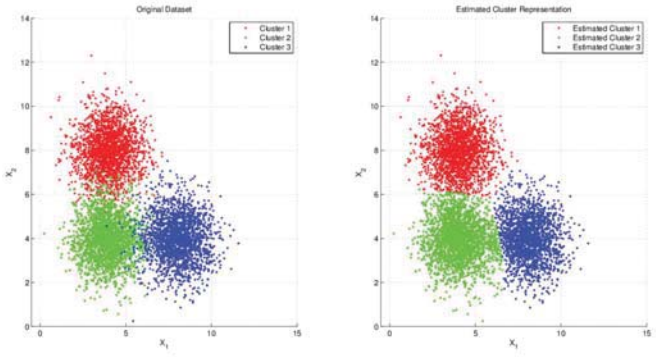

Fig. 3. Original and Estimated Clusters on Synthetic Gaussian Mixture Model.

The accuracy of the resulting clustering was evaluated using a confusion matrix-based approach. To this end, the resulting clusters were matched to the original clusters in a way to maximize the agreement between them, and a measure of clustering performance was calculated by the fraction of the points along the main diagonal to the total number of points.

The clustering performances for different sample sizes are shown in Table I. Generally, the clustering results obtained by the proposed algorithm matched the results obtained by the original expectation maximization procedure that used Gaussian models to characterize the clusters, without the benefit of a model assumption that fit the data, especially in cases where the second cluster is well populated. In cases where the second cluster had significantly fewer points, the statistical evidence in the dataset was too weak to warrant a separate cluster for those points, causing the algorithm to miss the second cluster.

TABLE I. ALGORITHM PERFORMANCE ON SYNTHETIC DATASET FOR DIFFERENT SAMPLE SIZES IN CLUSTERS

N_1	N_2	N_3	accuracy using iterative QSL	accuracy using conventional EM
500	500	1000	0.9637	0.9640
500	1000	2000	0.9473	0.9526
1000	500	1000	0.9623	0.9537
1000	1000	1000	0.9592	0.9682
2000	500	2000	0.8889	0.9419

After testing our proposed algorithm on synthetic datasets, we applied it to real multi-color flow cytometery (FCM) datasets. The FCM datasets that were used in these experiments were obtained from FlowCap-I Challange intended to comparatively evaluate automated clustering methods for FCM datasets. From this collection, we have used a human dataset diffuse large B-cell lymphoma (DLBCL) (containing 12369 samples in three clusters) and a mouse hematopoietic stem cell transplant dataset (HSCT) (containing 8914 samples in four clusters) that were publicly available with corresponding labels obtained via manual gating [16]. The manual gating procedure used to label the cells involved creating two-dimensional scatter plots of all possible parameter (fluorochrome) pairs (FL1vs2, 1vs3, 1vs4, 2vs3, 2vs4, 3vs4) and choosing the one in which the distinctions between the different clusters is most conspicious for manual gating. In accordance with this

approach, we have also used the same parameter pairs in the datasets to carry out the clustering experiments.

The manual gating results on the diffuse large B-cell lymphoma (DLBCL) dataset are presented in Figure 4. The proposed method accurately identifies two of the three clusters while missing the other one, due to its very small sample size, containing only 25 samples. This results in a clustering accuracy of 0.9045 (Figure 5). The conventional expectation maximization algorithm produced a similar two-cluster division with a similar accuracy at 0.9040 (Figure 6).

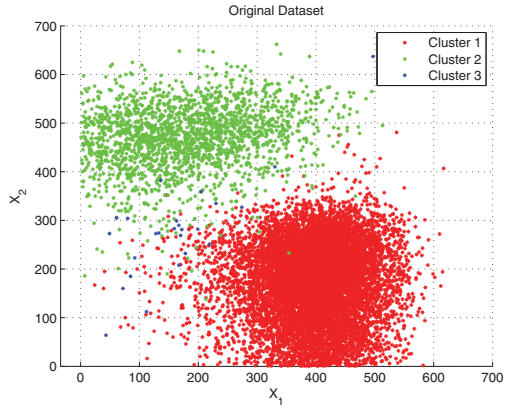

Fig. 4. Manual Gating Results for The Diffuse Large B-cell Lymphoma (DLBCL) Dataset

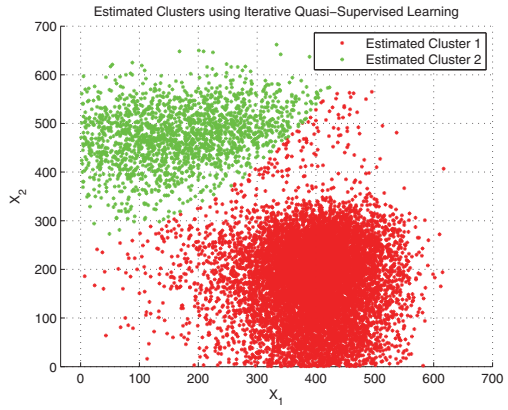

Fig. 5. Automated Gating Results for The Diffuse Large B-cell Lymphoma (DLBCL) Dataset Using Iterative Quasi-Supervised Learning

The manually gated clusters of the mouse hematopoietic stem cell transplant (HSCT) dataset are shown in Figure 7. As in the case of the earlier dataset, the proposed algorithm accurately identified three of the four clusters while missing the last one due again to its small sample size of 100, producing an overall accuracy of 0.8106 (Figure 8). Iterative binary disivison using the conventional expectation maximization algorithm identified only two cell clusters at an accuracy 0.6904. The estimated clusters are shown in Figure 9. Carrying out the original expectation maximization algorithm outside of the binary division framework assuming four clusters failed since the algorithm couldn't capture the smallest cluster. Thus, assuming three distinct clusters produced a better clustering with an accuracy of 0.9706.

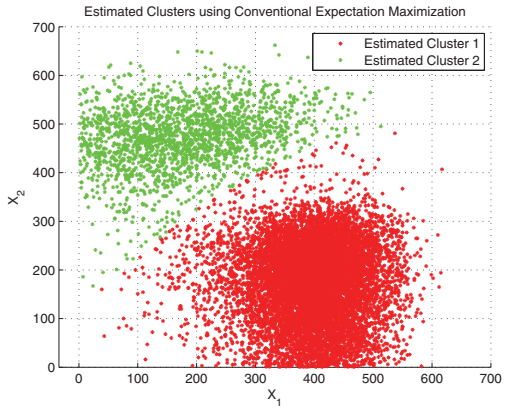

Fig. 6. Automated Gating Results for The Diffuse Large B-cell Lymphoma (DLBCL) Dataset Using Conventional Expectation Maximization

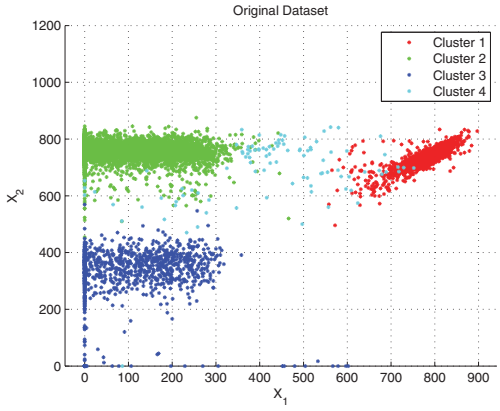

Fig. 7. Manual Gating Results for The Hematopoietic Stem Cell Transplant (HSCT) Dataset

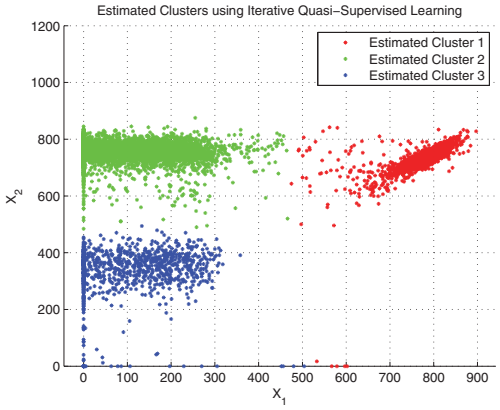

Fig. 8. Automated Gating Results for Hematopoietic Stem Cell Transplant (HSCT) Dataset using Iterative Quasi-Supervised Learning

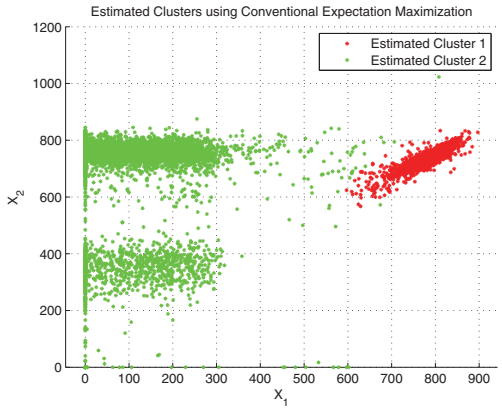

Fig. 9. Automated Gating Results for Hematopoietic Stem Cell Transplant (HSCT)) Dataset Using Conventional Expectation Maximization

III. Conclusion

We have proposed a recursive binary division method for data clustering without requiring the knowledge of the number of clusters or the parametric models that govern the individual components. The method divides each cluster into two daughter clusters using a model-free expectation maximization routine until the cost of separating the initial cluster into two daughter clusters exceeds the division cost acquired when forming the parent cluster. The model-free expectation maximization exploits the posterior probability estimation capability of the quasi-supervised learning algorithm, and avoids making assumptions on the distributions of the unknown data components.

In experiment results, the proposed method accurately identified the clusters of interest both on synthetic datasets as well as datasets derived from real multi-color flow cytometry experiments. The results also showed that the ability to identify the distinct clusters, however, depended on sample sizes as well as the distinctness of the different clusters. Consequently, clusters with too few samples were missed due to a lack of sufficient statistical evidence as assessed by the proposed method for their distinct presence. Work is currently in progress to enhance the proposed method to accurately capture small clusters.

References

[1] N. Aghaeepour, G. Finak, H. Hoos, T. R. Mosmann, R. Brinkman, R. Gottardo, R. H. Scheuermann, F. Consortium, D. Consortium *et al.*, "Critical assessment of automated flow cytometry data analysis techniques," *Nature methods*, vol. 10, no. 3, pp. 228–238, 2013.

[2] E. Lugli, M. Roederer, and A. Cossarizza, "Data analysis in flow cytometry: the future just started," *Cytometry Part A*, vol. 77, no. 7, pp. 705–713, 2010.

[3] K. Lo, R. R. Brinkman, and R. Gottardo, "Automated gating of flow cytometry data via robust model-based clustering," *Cytometry Part A*, vol. 73, no. 4, pp. 321–332, 2008.

[4] S. Pyne, X. Hu, K. Wang, E. Rossin, T.-I. Lin, L. M. Maier, C. Baecher-Allan, G. J. McLachlan, P. Tamayo, D. A. Hafler *et al.*, "Automated high-dimensional flow cytometric data analysis," *Proceedings of the National Academy of Sciences*, vol. 106, no. 21, pp. 8519–8524, 2009.

[5] G. Finak, A. Bashashati, R. Brinkman, and R. Gottardo, "Merging mixture components for cell population identification in flow cytometry," *Advances in Bioinformatics*, vol. 2009, 2009.

[6] J. Quinn, P. W. Fisher, R. J. Capocasale, R. Achuthanandam, M. Kam, P. J. Bugelski, and L. Hrebien, "A statistical pattern recognition approach for determining cellular viability and lineage phenotype in cultured cells and murine bone marrow," *Cytometry Part A*, vol. 71, no. 8, pp. 612–624, 2007.

[7] K. Lo, F. Hahne, R. R. Brinkman, and R. Gottardo, "flowclust: a bioconductor package for automated gating of flow cytometry data," *Bmc Bioinformatics*, vol. 10, no. 1, p. 145, 2009.

[8] M. J. Boedigheimer and J. Ferbas, "Mixture modeling approach to flow cytometry data," *Cytometry Part A*, vol. 73, no. 5, pp. 421–429, 2008.

[9] H. Wang and S. Huang, "Mixture-model classification in dna content analysis," *Cytometry Part A*, vol. 71, no. 9, pp. 716–723, 2007.

[10] H. Mucha, U. Simon, and R. Brüggemann, "Model-based cluster analysis applied to flow cytometry data of phytoplankton," *Weierstraß-Institute for Applied Analysis and Stochastic, Technical Report No*, vol. 5, 2002.

[11] S. Demers, J. Kim, P. Legendre, and L. Legendre, "Analyzing multivariate flow cytometric data in aquatic sciences," *Cytometry*, vol. 13, no. 3, pp. 291–298, 1992.

[12] B. Karaçalı, "Quasi-supervised learning for biomedical data analysis," *Pattern Recognition*, vol. 43, no. 10, pp. 3674–3682, 2010.

[13] G. Shafer *et al.*, *A mathematical theory of evidence*. Princeton university press Princeton, 1976, vol. 1.

[14] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–38, 1977.

[15] T. K. Moon, "The expectation-maximization algorithm," *Signal processing magazine, IEEE*, vol. 13, no. 6, pp. 47–60, 1996.

[16] FlowSite. (12001) Flowcap-i challange dataset. [Online]. Available: http://flowcap.flowsite.org