Annealing-based model-free expectation maximisation for multi-colour flow cytometry data clustering

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

Abstract: This paper proposes an optimised model-free expectation maximisation 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 maximisation scheme that exploits the posterior probability computation capability of the quasi-supervised learning algorithm subjected to a line-search optimisation over the reference set size parameter analogous to a simulated annealing approach. The divisions are continued until a division cost exceeds an adaptively determined limit. Experiment results on synthetic as well as real multi-colour flow cytometry datasets showed that the proposed method can accurately capture the prominent clusters without requiring any prior knowledge on the number of clusters or their distribution models.

Keywords: expectation maximisation algorithm; quasi-supervised learning algorithm; clustering; gating of flow cytometry data; simulated annealing algorithm; flow cytometry data analysis; bioinformatics; data mining.

Reference to this paper should be made as follows: Köktürk, B.E. and Karaçalı, B. (2016) 'Annealing-based model-free expectation maximisation for multi-colour flow cytometry data clustering', *Int. J. Data Mining and Bioinformatics*, Vol. 14, No. 1, pp.86–99.

Biographical notes: Başak Esin Köktürk, MSc, received both her Bachelor of Science and Master of Science in Electrical and Electronics Engineering from the İzmir Institute of Technology, İzmir, Turkey. She is currently a PhD candidate in the same department. Her current research focuses on the automated clustering methods and statistical biomedical signal and image processing. She is currently working as research assistant at the Biomedical Information Processing Laboratory in İzmir Institute of Technology.

Bilge Karaçalı, PhD, earned his Doctor of Philosophy degree in 2002 at the Electrical and Computer Engineering Department of the North Carolina State University in Electrical Engineering, with a minor in Mathematics. He has worked on biomedical image analysis as a research fellow at the Radiology Department of the University of Pennsylvania, before joining the Center for Integrated Bioinformatics of Drexel University in 2005 as a research assistant professor and Assistant Director of Bioimaging. Since 2008, he is with the Electrical and Electronics Engineering Department of İzmir Institute of Technology as Professor and Director of the Biomedical Information Processing Laboratory.

This paper is a revised and expanded version of a paper entitled 'Model-free expectation maximization for divisive hierarchical clustering of multicolor flow cytometry data' presented at the 'IEEE International Conference on Bioinformatics and Biomedicine (BIBM)', Belfast, UK, 2–5 November 2014.

1 Introduction

Flow cytometry (FCM) is a powerful laser-based multi parametric analysis technique for characterising 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. The measured properties include an individual particle's relative size, relative granularity, internal complexity and relative fluorescence intensity (BD Biosciences, 2000; Parks et al., 1989). FCM is used in research applications to distinguish different cell types from each other as well as in clinical applications for disease diagnosis, especially blood cancers, and monitoring disease progression following therapy (Aghaeepour et al., 2013).

In FCM experiments, cells are incubated with fluorochrome-conjugated antibodies. Fluorochromes are attached to an antibody that binds specifically to a target protein in or on the cell. Since each fluorochrome has a specific peak wavelength in its emission spectrum, the characteristics of the emitted light from the cells under laser excitation allows assessing the relative abundance of the targeted biomarkers. Several biomarkers can be investigated simultaneously in multi-colour flow cytometry experiments by increasing the number of fluorochrome-antibody pairs. Currently, the FCM technology allows investigating cells for the presence and abundance of up to 20 biomarkers (Lugli et al., 2010). However, increasing the number of parameters inevitably increases the dataset dimension and complexity, and creates new challenges in FCM data analysis.

Identification of subpopulations using standard methods based on manual gating is laborious and time-consuming. Furthermore, obtaining matching results for the same flow data is difficult even by same expert (Lo et al., 2008). Consequently, there is a considerable demand for automated methods to address these challenges, particularly for multi-colour flow data analysis.

Several methods for automated identification of cell subsets have been proposed in the literature by modelling cell population characteristics. For instance, Aghaeepour et al. (2013) developed an automated method for cell subtype identification in high dimensional FCM data based on *k*-means clustering, while Pyne et al. (2009) proposed a skew and heavy tailed distribution fitting approach. The FlowClust algorithm, proposed by Lo et al. (2008), aims to fit a t-mixture model to FCM data after the Box-Cox transformation. The FlowClust algorithm was later modified by Finak et al. (2009) by introducing a merging step to avoid unwarranted cluster divisions. 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 unknown number of distinct clusters. This means that the clustering algorithm is to be run several times for varying number of clusters and the clustering result that achieves the optimal separation according to the criterion of choice is to be taken as the final output.

In the literature, there are several supervised and unsupervised methods for automated gating of FCM data. Supervised algorithms (Lo et al., 2009; Quinn et al., 2007) are generally ill-suited for FCM data analysis because they need training datasets that must be created beforehand by an expert for a specific experiment configuration. This means that when system settings, including cytometer options as well as cell preparation protocols, change, the algorithm requires new training data representing the final configuration. Unsupervised techniques include variations of the mixture modelling approach (Boedigheimer and Ferbas, 2008; Wang and Huang, 2007), model-based clustering (Mucha et al., 2002; Demers et al., 1992) and density-based clustering (Pyne et al., 2009). Since unsupervised methods do not require training datasets, they offer greater applicability than the supervised methods. On the other hand, unsupervised methods tend to perform poorly when the assumed model does not match the actual distribution or when the adjustable parameters are not chosen correctly (Bashashati and Brinkman, 2009).

In this paper, we propose an optimised version of the model-free expectation maximisation division algorithm for FCM data presented earlier (Kokturk and Karacali, 2014). The new method also starts by dividing the whole dataset into two groups, but this time, using an optimised implementation of the original expectation maximisation procedure that relies on a model-free calculation of the group posterior probabilities. The optimisation entails carrying out a line search procedure that is analogous to annealing for an energy functional evaluating the quality of the cluster separation. The method then continues to divide the cell subgroups obtained by previous divisions until a stopping condition that detects superfluous divisions is met, expressed through a non-parametric division cost. This allows cell subgroup identification without making any assumptions on the shape of the cell subtype distributions and by deducing the number of prevalent cell subgroups adaptively from the flow dataset.

This paper is organised 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.

2 Methods

In this section, we firstly describe the quasi-supervised learning algorithm that estimates the posterior probabilities of a given pair of clusters at each sample (Karaçalı, 2010). Then, we summarise the expectation-maximisation algorithm as described by Dempster et al. (1977) and Shafer (1976) followed by the proposed modification that replaces model-based posterior probabilities with model-free posterior probabilities estimated via the quasi-supervised learning algorithm. The section concludes with a detailed description of the proposed model-free automated cell population identification method for multi-colour flow cytometry datasets along with the introduced optimisation for cluster distinctness.

2.1 Posterior probability estimation using the quasi-supervised learning algorithm

The quasi-supervised learning algorithm exploits an asymptotic property of nearest neighbour classification over randomly chosen reference sets. For an unknown sample $x \in X$, a nearest neighbour classifier $F(x;R)$ over a given reference set *R* is defined by

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

with y^* representing the class label of the point x^* satisfying

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

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

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

where the indicator function $\mathbf{1}(\cdot)$ returns 1 when its argument holds, and zero otherwise (Karaçalı, 2010). However, since $p_{R_n}(R_n)$ is unknown, the integral in equation (3) cannot be carried out in practice, but can be approximated by calculating the average number of times x is assigned to C_0 via $F(x;R)$ using all reference sets R_n that can be formed using the available data $\{x_i, y_i\}$, $x_i \in X$, $y_i \in \{0,1\}$ for $i = 1,2,\ldots,l$.

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

where M denotes the number of distinct reference sets. The posterior probability $P(C_1 \mid x)$ can also be written in a similar fashion as

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

The quasi-supervised learning algorithm computes the averages above using a practical approach that avoids carrying out M separate nearest neighbour classifications for a given choice of n. 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

$$L(x) = \log\left(\frac{p(C_0 \mid x)}{p(C_1 \mid x)}\right) \cong \log\left(\frac{f_0(x)}{f_1(x)}\right) \tag{6}$$

since the class priors are set at 0.5 in the calculation above by including an equal number of samples from C_0 and C_1 into R_n. The optimal number of points n_{opt} in the reference set for best learning is determined adaptively from the available data by minimising a cost function $E(n)$ defined by

$$E(n) = 4\sum_i f_0(x_i) * f_1(x_i) + 2n \tag{7}$$

that strikes a balance between good separation and generalisability, since large n produces overly flexible decision regions that decrease the algorithm accuracy (Karaçalı and Karim, 2003).

2.2 Expectation-maximisation algorithm

The conventional expectation maximisation algorithm aims to fit a mixture distribution model to a specified dataset (Shafer, 1976; Dempster et al., 1977; Moon, 1996). Let θ_j be the parameter for the *j*-th component of the mixture for $j = 1, 2, \ldots, k$. If the components are taken to be of the Gaussian form as is generally the case, θ_j can be defined as

$$\theta_j = (\mu_j, \Sigma_j) \tag{8}$$

where μ_j and Σ_j denote the means and the covariance matrices of the corresponding components. The objective, then, is to determine the distribution parameters θ_j to fit the available data $x_1, x_2, \ldots, x_\ell$. The likelihood function for each θ_j can be expressed as

$$L_x(\theta_j; x_1, x_2, \ldots, x_\ell) = f(x_1, x_2, \ldots, x_\ell \mid \theta_j) = \prod_{i=1}^{\ell} f(x_i \mid \theta_j) \tag{9}$$

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

$$\theta_j^{ML} = \arg\max_{\theta} L_x(\theta) \tag{10}$$

or equivalently,

$$\theta_j^{ML} = \arg\max_{\theta} \log L_x(\theta) \tag{11}$$

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

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

$$r_{i,j} = \frac{p(x_i, \theta_j)}{\sum_{m=1}^{k} p(x_i \mid \theta_m)} \tag{12}$$

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 maximisation step using a maximum likelihood procedure that takes the responsibility values into account. A notable distinction between different expectation maximisation procedures arises from the use of the responsibility values in the maximisation 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* and using only these points to estimate the corresponding model parameter. In the other alternative, the model parameters θ_j are estimated in a way that uses all points simultaneously, but in a way to be influenced more by the points x_i for which $r(i, j)$ are greater and less by the others.

2.3 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-maximisation cycle that begins with a large value for the reference set size parameter *n* and computes the posterior probability of C_0 and C_1 at each sample. The algorithm proceeds by re-assigning the points to the cluster whose posterior is larger and iterates until convergence. After convergence, the procedure is re-applied to the data starting with the latest cluster assignments using a smaller *n*. The optimal cluster assignments are selected by tracking the cost function in equation (7) as *n* decreases to 1, and identifying the level for which $E(n)$ is minimal. The block diagram of the proposed method is shown in Figure 1. The modified expectation-maximisation procedure that forms the basis of the proposed clustering method is summarised below:

1: **for** $i = n_{max} : -1 : 1$ **do**
2: **Expectation Step:**
Calculate $P(C_0 \mid x)$ and $P(C_1 \mid x)$
3: **Maximisation Step:**
Update class labels
$C_0 \leftarrow \{x \mid f_0(x) \geq 0.5\}$
$C_1 \leftarrow \{x \mid f_1(x) < 0.5\}$

Note that an analogy can be formed between the proposed clustering algorithm and a simulated annealing procedure, as simulated annealing aims to find the global minimum of a cost function by decreasing the energy level of a system gradually as it converges to the desired solution (Kirkpatrick et al., 1983). In the proposed method, *n* represents the system energy as large *n* produces a more flexible learning system, and $E(n)$ measures the complexity of the clustering obtained for a given *n*. At the level where $E(n)$ is minimal, the algorithm produces the best clustering result where the clusters exhibit the smallest overlap.

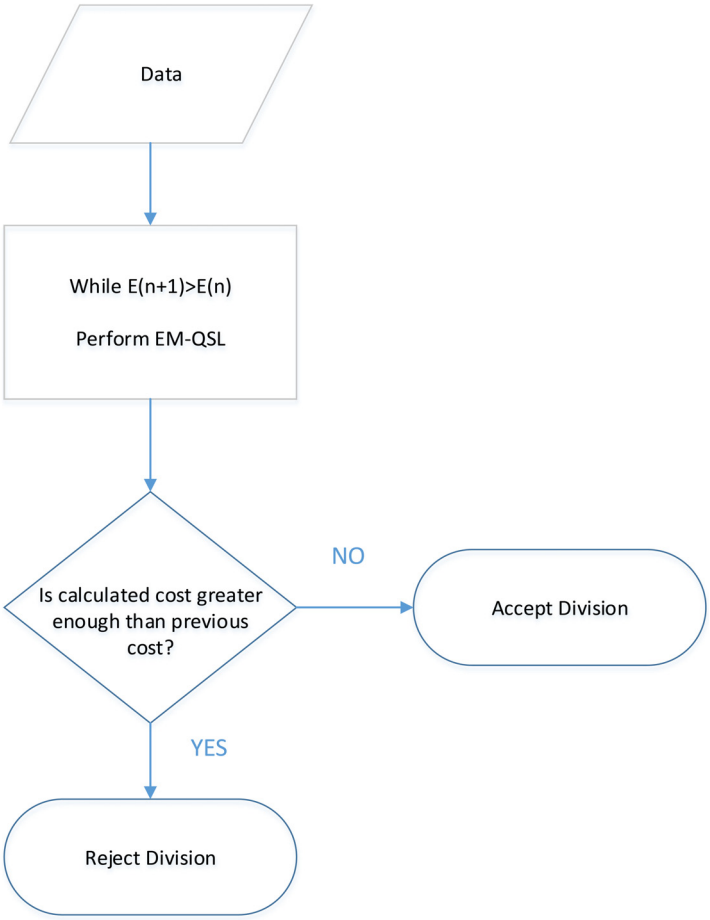
Note also that the procedure above produces two distinct clusters starting with a single one, regardless of whether the resulting clusters are distinct enough to merit separation. In order to evaluate the distinctness of the resulting clusters, we have defined a division cost $c(C_0, C_1)$ by

$$c(C_0, C_1) = \frac{1}{N_0} \sum_{x_i \in C_0} f_1(x_i) + \frac{1}{N_1} \sum_{x_i \in C_1} f_0(x_i) \tag{13}$$

with N_0 and N_1 denoting the number of points 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 of the earlier clustering that produced the parent cluster $C_0 \bigcup C_1$: If the division cost exceeds the parent cluster original division cost by 0.03, the algorithm stops and rejects the division. This amount was determined empirically to provide good clustering results on a variety of datasets.

As the last step once all the binary divisions are finalised, we have used a post-processing operation to revert unwarranted cluster divisions by evaluating whether the union of any two of the resulting clusters forms a single coherent cluster. To this end, we have combined all resulting clusters in groups of two and calculated the division cost between all resulting cluster pairs; merging the clusters for which the division cost is larger than all previously accepted division costs of their parent clusters.

Figure 1 Block diagram of the proposed method (see online version for colours)



3 Results

The proposed method was applied to synthetically generated datasets as well as datasets acquired from real multi-colour flow cytometry experiments. The synthetic dataset contained three distinct clusters, each modelled using a two-dimensional Gaussian distribution with identity 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 earlier version that does not involve optimising with respect to n (Kokturk and Karacali, 2014) and the conventional expectation-maximisation routine for two-component Gaussian mixture fitting within the same binary division scheme for estimating the posterior probabilities from a model-based perspective.

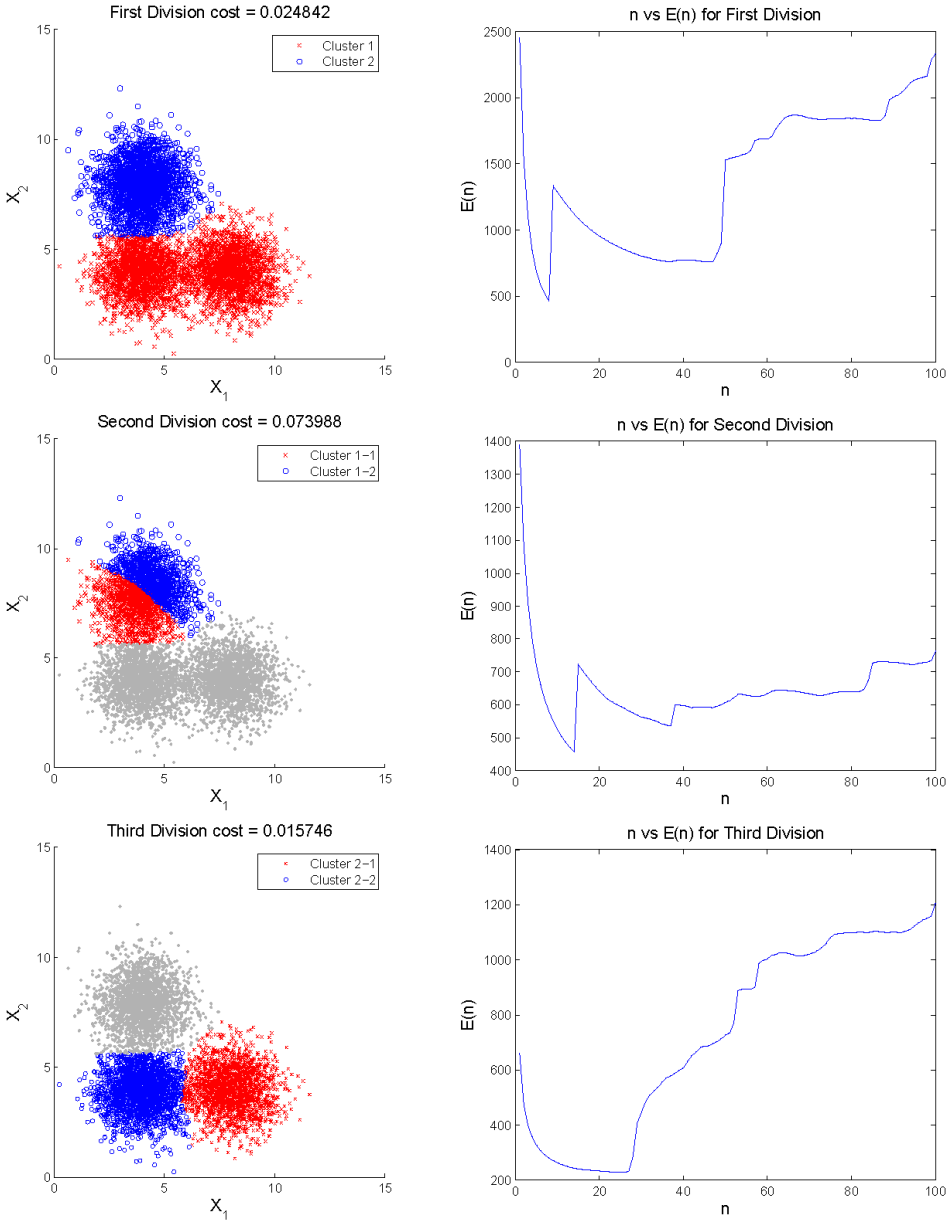
Illustrative results obtained by the proposed method for the above datasets with N_1, N_2, N_3 number of points in the three clusters are presented in Figure 2 along with the associated cost functions $E(n)$ and the division costs $c(C_0, C_1)$ for each successive binary clustering. The behaviour of $E(n)$ shows that as n decreases, the clustering becomes strained resulting in a gradual increase in $E(n)$, and resolves to a more suitable configuration after n passes a critical level, resulting in a dramatic decrease. The optimal clusterings are observed following one such decrease. Note also that the cost of separation remains relatively high in the division shown in the second row compared to the valid divisions in the first and third rows. Different colours represent the two clusters achieved at the optimal separation, while the points of the original dataset not included in the division process are shown in grey.

The accuracy of the clustering results for comparative evaluation purposes was measured using a confusion matrix-based approach by the fraction of the points along the main diagonal to the total number of points. The clustering results obtained for varying sample sizes using the proposed algorithm that involves carrying out the line-search optimisation with respect to n, along with the earlier version that uses a fixed n at 1 and the conventional expectation-maximisation routine are shown in Table 1. In general, the clustering performance of the proposed algorithm is greater than the other alternatives. Even more significantly, the proposed method compares equally or favourably against the implementation using the conventional expectation-maximisation method based on Gaussian model components that is expected to perform near-optimally due to the Gaussianity of the clusters in the actual data.

Table 1 Comparative performance evaluation results on synthetic datasets for varying sample sizes in the three clusters. The numbers represent the average accuracies for the corresponding algorithms over 20 independent repeats

N_1	N_2	N_3	*Accuracy using iterative QSL (n=1)*	*Accuracy using iterative QSL (optimised n)*	*Accuracy using conventional EM*
500	500	1000	0.9637	0.9691	0.9640
500	1000	2000	0.9473	0.9724	0.9526
1000	500	1000	0.9623	0.9689	0.9537
1000	1000	1000	0.9592	0.9424	0.9682

Figure 2 The data division is illustrated on the left column and respective $E(n)$ functions are given in the right column. The data divided into two clusters (upper row) and division cost is decided. Then same procedure applied on daughter clusters (second and third row) (see online version for colours)



After testing our proposed algorithm on synthetic datasets, we applied it to real multi-colour flow cytometry (FCM) datasets. The FCM datasets used in these experiments were obtained from FlowCap-I Challenge intended to comparatively evaluate automated clustering methods for FCM datasets. From this collection, we have used a human dataset

of diffuse large B-cell lymphoma (DLBCL) (containing 12,369 samples divided in three clusters) which consists of lymph node biopsies from patients were histologically confirmed to have DLBCL and treated at the British Columbia Cancer Agency between 2003 and 2008 and a mouse haematopoietic stem cell transplant dataset (HSCT) (containing 8914 samples divided in four clusters) derived from HSCT experiments done in the Terry Fox Laboratory. Suspensions were produced from bone marrow cells and they were depleted of erythroid precursors by immunomagnetic removal of biotin conjugated anti-Ter119-labeled cells using EasySep reagents (Aghaeepour et al., 2013). The true cluster assignments of both datasets were assigned via manual gating and provided along with the fluorescence data (FlowSite, 2014). The manual gating procedure used to label the cells involved creating two-dimensional scatter plots of all possible parameter (fluorochrome) pairs (FL1 vs. FL2, FL1 vs. FL3, FL1 vs. FL4, FL2 vs. FL3, FL2 vs. FL4, FL3 vs. FL4) and choosing the one in which the distinctions between the different clusters was most conspicuous and suitable for manual gating. We have evaluated the performance of our algorithm on these datasets by comparing the resulting cluster labels with the manual gated labels. In our experiments, we have used all fluorochromes to carry out clustering even though the resulting cluster assignments are shown on 2D scatter plots.

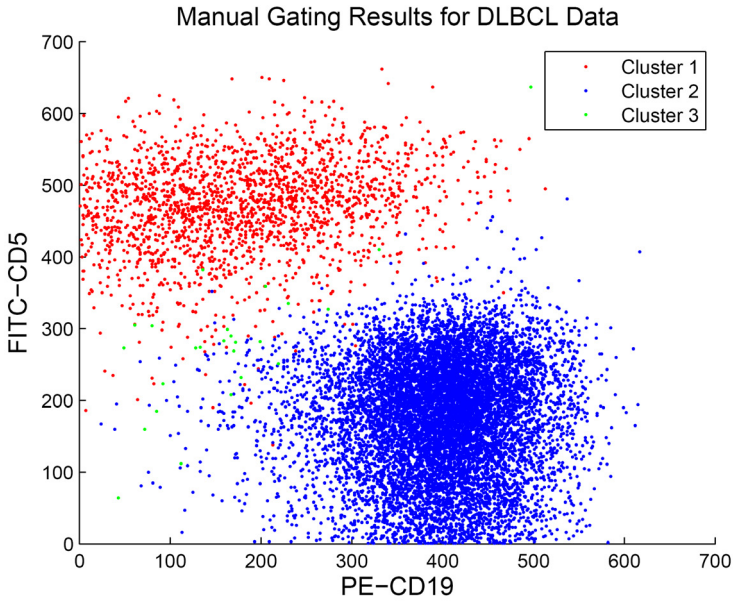
The actual labels of the cells in the diffuse large B-cell lymphoma (DLBCL) datasets are presented in Figure 3a. The earlier method in Kokturk and Karacali (2014) had identified only two of three clusters while missing the other one which has only 25 samples with an overall accuracy of 0.9045. The optimised model-free expectation maximisation algorithm proposed here also missed the third cluster again due to the absence of statistical significance of the small cluster, but it identified the other two clusters samples with an increased overall accuracy of 0.9959 (Figure 3b).

The manually gated clusters of the mouse haematopoietic stem cell transplant (HSCT) dataset are shown in Figure 4a. 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. The overall accuracy of the earlier algorithm on this dataset was 0.8106. The proposed algorithm achieved a dramatic increase in accuracy, reaching a level of 0.9874 (Figure 4b).

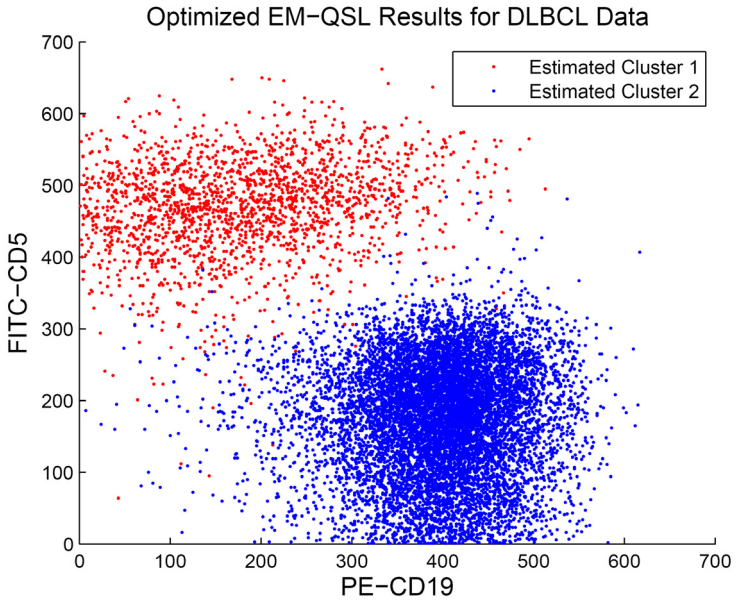
4 Conclusion

We have proposed a recursive binary division algorithm for unsupervised clustering of vector-valued data that does not require any knowledge about the underlying data distribution such as the number of clusters, distribution models or model parameters. The method operates by dividing the original dataset into two daughter clusters using the posterior probability estimates provided by the quasi-supervised learning algorithm in an expectation-maximisation framework while optimising the reference set size parameter n. The same procedure is then applied to the daughter clusters themselves and their daughter clusters and so on, until the division cost of the daughter clusters exceeds the division cost of the parent cluster by a significant margin. As the procedure relies on model-free posterior probability estimation, the proposed method avoids the pitfalls associated with making incorrect or unsuitable assumptions on the underlying distributions of the unknown data components.

Figure 3 Clusters obtained from manual gating (a) and proposed method (b) on DLBCL dataset (see online version for colours)

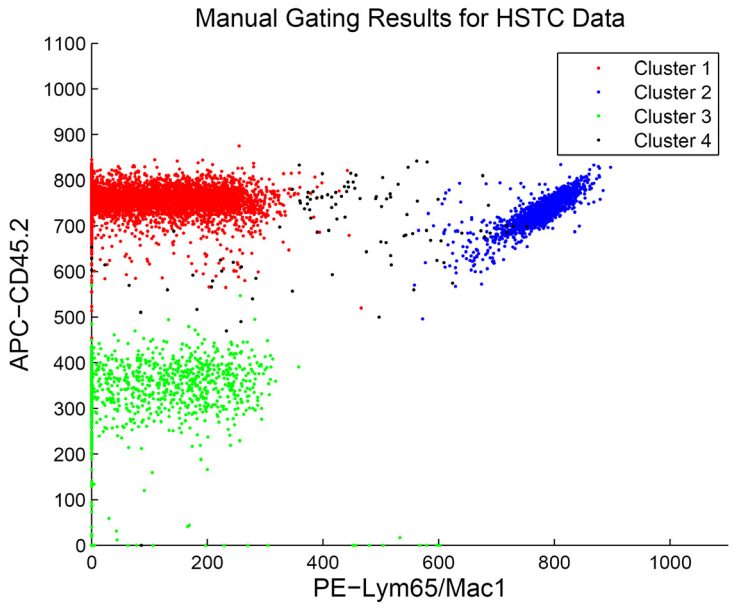


(a)

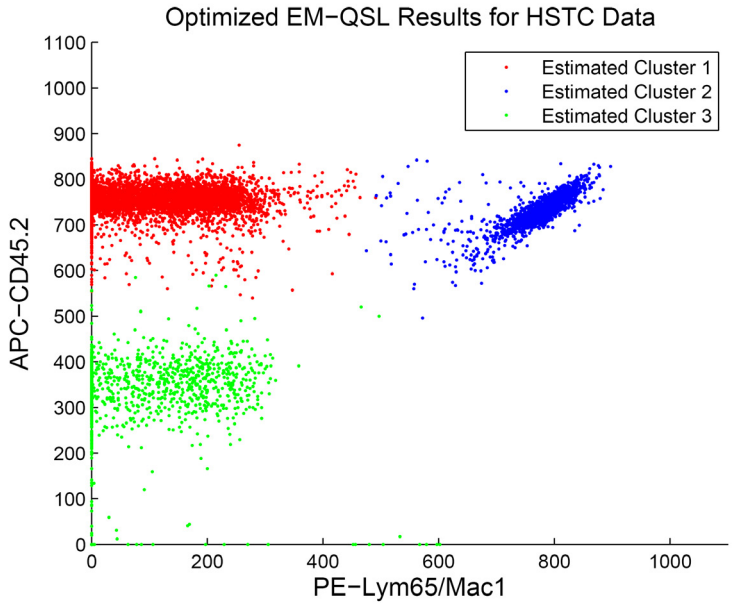


(b)

Figure 4 Clusters obtained from manual gating (a) and proposed method (b) on HSCT dataset (see online version for colours)



(a)



(b)

In experiment results, the proposed method accurately identified the clusters of interest both on synthetic datasets as well as datasets collected from real multi-colour flow cytometry experiments. The experiments also showed that clusters with too few samples were at risk of being not recognised as separate clusters. This may be a general issue with the specific case of flow data clustering as such clusters are often determined by manual gating over cells that exhibit a specific combination of fluorochrome intensities that are known to be associated with a distinct cell type, regardless of the small number of cells that fit those characteristics. From a statistical standpoint, it is not surprising that such small clusters are missed due to insufficient representation within the overall dataset. However, work is currently under way to incorporate flow data-specific priors into the algorithm by defining a more suitable division cost function for improved sensitivity to detect small clusters.

As the proposed method realises an expectation maximisation procedure for hierarchical clustering of datasets of unknown distribution characteristics, it can be applied to all clustering problems in which the conventional expectation maximisation procedure using Gaussian or other model components has offered good solutions. Even more significantly, as the quasi-supervised learning algorithm providing the model-free posterior probability estimates operates on distances between data points, the method does not require the data to be presented in vector-valued form to be applicable. This may particularly be useful for clustering datasets where distances between samples can be derived in the absence of a vector-space representation of the data via expectation-maximisation such as genomic sequence datasets. This line of research is also currently under investigation.

References

Aghaeepour, N., Finak, G., Hoos, H., Mosmann, T.R., Brinkman, R., Gottardo, R. et al. (2013) 'Critical assessment of automated flow cytometry data analysis techniques', *Nature Methods*, Vol. 10, No. 3, pp.228–238.

Bashashati, A. and Brinkman, R.R. (2009) 'A survey of flow cytometry data analysis methods', *Advances in Bioinformatics*, Vol. 2009, Article ID 584603.

BD Biosciences (2000) *Introduction to Flow Cytometry: A Learning Guide*. Manual Part Number: 11-11032-01, April.

Boedigheimer, M.J. and Ferbas, J. (2008) 'Mixture modeling approach to flow cytometry data', *Cytometry Part A*, Vol. 73, No. 5, pp.421–429.

Demers, S., Kim, J., Legendre, P. and Legendre, L. (1992) 'Analyzing multivariate flow cytometric data in aquatic sciences', *Cytometry*, Vol. 13, No. 3, pp.291–298.

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

Finak, G., Bashashati, A., Brinkman, R. and Gottardo, R. (2009) 'Merging mixture components for cell population identification in flow cytometry', *Advances in Bioinformatics*, Vol. 2009, Article ID 247646.

FlowSite (2014) *FlowCap-I Challange Dataset*, 2014.

Karaçal, B. (2010) 'Quasi-supervised learning for biomedical data analysis', *Pattern Recognition*, Vol. 43, No. 10, pp.3674–3682.

Karaçali, B. and Krim, H. (2003) 'Fast minimization of structural risk by nearest neighbor rule', *IEEE Transactions on Neural Networks*, Vol. 14, No. 1, pp.127–137.

Kirkpatrick, S., Gelatt, C.D. and Vecchi, M.P. (1983) 'Optimization by simmulated annealing', *Science*, Vol. 220, No. 4598, pp.671–680.

Kokturk, B.E. and Karacali, B. (2014) 'Model-free expectation maximization for divisive hierarchical clustering of multicolor flow cytometry data', *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp.267–272.

Lo, K., Brinkman, R.R. and Gottardo, R. (2008) 'Automated gating of flow cytometry data via robust model-based clustering', *Cytometry Part A*, Vol. 73, No. 4, pp.321–332.

Lo, K., Hahne, F., Brinkman, R.R. and Gottardo, R. (2009) 'Flowclust: a bioconductor package for automated gating of flow cytometry data', *BMC Bioinformatics*, Vol. 10, No. 1, p.145.

Lugli, E., Roederer, M. and Cossarizza, A. (2010) 'Data analysis in flow cytometry: the future just started', *Cytometry Part A*, Vol. 77, No. 7, pp.705–713.

Moon, T.K. (1996) 'The expectation-maximization algorithm', *IEEE Signal Processing Magazine*, Vol. 13, No. 6, pp.47–60.

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

Parks, D., Herzenberg, L. and Herzenberg, L. (1989) 'Flow cytometry and fluorescence-activated cell sorting', in Paul, W. (Ed.): *Fundamental Immunology*, Raven, New York.

Pyne, S., Hu, X., Wang, K., Rossin, E., Lin, T., Maier, L.M. et al. (2009) 'Automated high-dimensional flow cytometric data analysis', *Proceedings of the National Academy of Sciences*, Vol. 106, No. 21, pp.8519–8524.

Quinn, J., Fisher, P.W., Capocasale, R.J., Achuthanandam, R., Kam, M., Bugelski, P.J. and Hrebien, L. (2007) '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.

Shafer, G. (1976) *A Mathematical Theory of Evidence*, Vol. 1, Princeton University Press, Princeton.

Wang, H. and Huang, S. (2007) 'Mixture-model classification in DNA content analysis', *Cytometry Part A*, Vol. 71, No. 9, pp.716–723.