FLOPTICS: A NOVEL FLOW CYTOMETRY DATA AUTOMATED GATING TECHNIQUE BY USING OF OPTICS CLUSTERING TECHNIQUE

Wiwat Sriphum[1], Gary Wills[1] and Nicolas Green[1]
[1] *School of Electronics and Computer Science, University of Southampton, Southampton, UK*
{ws5n17, gbw, ng2}@soton.ac.uk

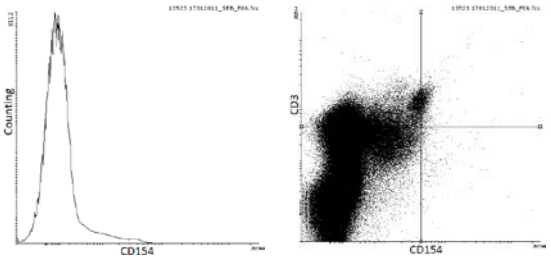
Keywords: flow cytometry, automated gating, density based clustering, optics clustering

Abstract: Gating is a process of cell differentiation of the data obtaining from flow cytometry technique. Flow cytometry (FCM) is a fluorescence concept technique that provide characteristics of individual cells in blood sample. A multidimensional dataset obtained from this method is large and complex so it is difficult to manually analysis and time consuming. Gating is the first step of FCM data analysis and highly subjective. Although many research attempted to reduce this subjectivity, more standard and faster gating technique is still needed. Some existing automated gating technique need many user-defined parameters that can lead to different results for different parameter values. Some techniques have a trouble with time consuming. FLOPTICS is a novel automated gating technique that is a combination of density-based and grid-based clustering algorithms. FLOPTICS has an ability to classify cell on FCM data faster and less user-defined parameter than some of state of art technique such as FlowGrid, FlowPeaks, and FLOCK.

1 INTRODUCTION

Flow cytometry is a well-known method that is used to identify characteristics of cells in a blood sample by using the concept of cell-scatter measurement and light emission after receiving a laser beam stimulation (Bio-Rad, 2018). It has been widely applied in medical research, especially in haematology and immunology and broadly adopted in the clinical environment in order to diagnose and monitor treatments, such as: leukaemia diagnosis, chemical healing responsiveness, and stem cell transplantation monitoring (Jahan-Tigh et al., 2012). The process provides multi-dimensional data including relative size, relative granularity, and relative fluorescence intensity (BD-Biociences, 2002). The data is so complicated that it is difficult to manually analyse.

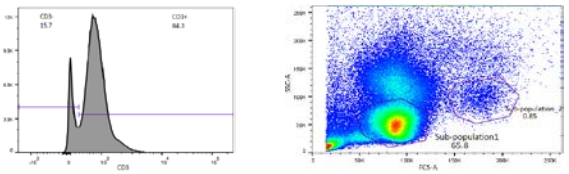
Flow cytometry (FCM) data is a multidimensional dataset and the data generally can be displayed in one or two parameters (Moloney and Shreffler, 2008). For one parameter, it can be displayed in histogram in which a feature of cell is shown on x-axis while the number of cell is shown on y-axis, as can be seen in Figure1-a. For two parameters, it is displayed simultaneously in a scatter plot which the first parameter is displayed on x-axis and another one is displayed on y-axis, as can be seen in Figure1-b. The feature of cell is able to provide up to 50 features (Lee et al., 2017), however, the number of feature depends on flow cytometer and experiment design. In Figure1-a, it represents a relation between the number of cell and intensity that responses to CD154 marker. In Figure1-b, quadrants are drawn on the scatter plot, it can notice that most cells are plotted in bottom-left area it means that most cells have low intensity of responding to both CD154 and CD3. After obtaining the data, the expertise have to first diagnose by selecting interested cells also known as gating process.



a. histogram *b. scatter plot*

Figure 1 Flow cytometry data representation

Gating is a process of selecting homogenous cells by drawing shapes around populations, which represent in 1D (histogram) or 2D (scatter plot), in order to group them together (Bashashati and Brinkman, 2009), as can be seen in Figure 2, which are manual gating examples. The expertise have to know about the characteristics of interested cells and discover sub-populations of cells as much as they can before start analysis. Manually gating process is, therefore, highly subjective and time consuming that lead to applying of machine learning for supporting this process (Lo, Brinkman and Gottardo, 2008).



a. gating of 1D histogram *b. gating of 2D scatter plot*

Figure 2 Manual gating examples

The FCM data is so large and complex that it is difficult to analyse without computational tools. The main problems of FCM analysis; firstly, manual gating (the that identifies interesting cells) is highly subjective (Lo, Brinkman and Gottardo, 2008); secondly, sometimes the number of key events is very low (Groeneveld-Krentz et al., 2016), which makes them harder to detect and may result in a false positive; thirdly, manually gating is a time consuming process (Rahim et al., 2018), especially when the number of parameters and blood cells are large. Although some applications have been developed to help clinical experts, these tools still have limitations, as mentioned before. This research aims to apply a machine learning technique to implement a novel automated gating technique which can provide appropriate clustering of cells in blood samples.

2 METHOD

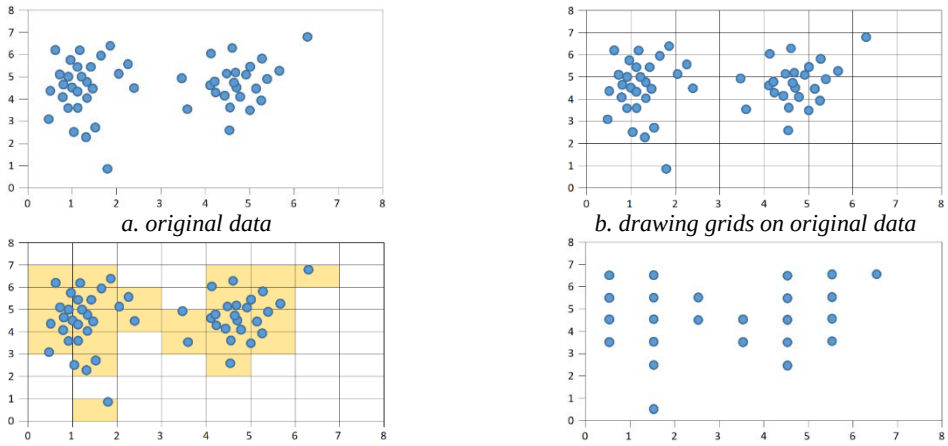
Ye and Ho, proposed FlowGrid in 2018 and they claimed that FlowGrid provides high accuracy and better time efficiency compared with flowPeaks (Ge and Sealfon, 2012), FlowSOM (Van Gassen et al., 2015), and FLOCK (Qian et al., 2010). However, FlowGrid still has the problem with requirement of too many user-defined parameters. Therefore, the aim of our method is to improve the performance of a state of art automated gating technique, FlowGrid, by reducing time process and user-defined parameters. The FLOPTICS algorithm start by partitioning data into equal size of grids for each dimension, which is called bins, then only non-empty bins are process as a data point. The example of partitioning 2-dimensional data has been shown in Figure 3. Partitioning data is not appropriate for low density dataset but FCM data is always high density, as can be seen in Figure 1 and 2, so the accuracy result of gating is acceptable and running time is faster than many state of art techniques.

After the data is partitioned into equal-sized bin, the data will be clustered by using of OPTICS algorithm (Ankerst et al., 1999), which is a density-based clustering method. Basically, radius distance called epsilon (ε) has to be defined by user such as DBSCAN (Ester et al., 1996) algorithm but OPTICS can provide the optimal value of ε by showing the structure of data that allow user can see and select the optimal ε. Therefore, the number of user-defined parameters for FLOPTICS is less than FlowGrid, which is based on DBSCAN. According to the method above, FLOPTICS algorithm can be summarized as follow:

Step1: All data points are partitioned into equal sized bins for each dimension
Step2: Only non-empty bins are processed
Step3: Read an un processed bin (*b*) from dataset obtaining from Step2
Step4: If b is a core-bin, update core-distance of b
for each $a \in N\varepsilon(b)$
update reachability of object *a*
update OrderSeeds list, which contains the objects ordered by reachability-distance from smallest
Mark b as processed
Step5: Read an unprocessed bin b from OrderSeeds list if the list is not empty, otherwise, read next unprocessed bin from dataset
Step6: Repeat Step3-Step5 until the end of dataset

Key terms involved in the algorithm above are defined below.

- Nε(b) is a set of neighbours of bin b regarding the radius distance value ε, which is identified by user
- Core-bin is the bin that its number of neighbour (regarding the radius ε) more than or equals to MinPts, which is identified by a user
- Core-distance is the minimum distance that lead the number of neighbour of bin b reach MinPts
- Directly connected
Bin a is directly connected to Bin b if $Distance(a,b) \leq \varepsilon$



a. original data *b. drawing grids on original data*

c. selecting non-empty bins *d. transformed data*

Figure 3: Partitioning 2-dimensional data into equal-sized bin for each dimensions

- Reachability-distance
  Let b and o be bins in the grid space, ε the distance between two bins, Nε(b) a set of neighbour of bin b, and MinPts the minimum number of a neighbour. Then reachability - distanceε,MinPts(o,b) is equal to
  - Infinity or undefined, if the number of member in Nε(p) less than MinPts
  - Max (core - distance (b), distance (o, b)), otherwise
    - core-distance (b) is the minimum distance from b that lead the number of neighbour of bin b reach *MinPts*
- OrderSeeds list is the list (queue) of bin in grid space ordered by reachability distance

3 RESULT AND DISCUSSION

This section provides results of applying DBSCAN, OPTICS, FlowGrid and FLOPTICS to synthesis dataset that is mimic generated regarding real FCM dataset. The experiment was conducted on a computer with specification as follows:

- Intel(R) Core(TM) i7-6700 CPU @ 3.40GHz
- RAM 16.0 GB
- Windows 10 Enterprise, 64 bit Operating System

Table 1: The values of arguments for each level of overlapping

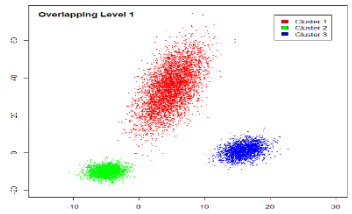
Cluster	Sigma (Covariance matrix)	N	Means (Centres)							
			Level1		Level2		Level3		Level4	
			x	y	x	y	x	y	x	y
1	[(6,15), (15,120)]	5000	5	35	3	30	1	25	-1	20
2	[(2,0.3), (0.3,5)]	2500	-5	-10	-5	-10	-5	-10	-5	-10
3	[(3,2), (2,10)]	2500	16	1	12	1	8	1	4	1

3.1 Reference Dataset

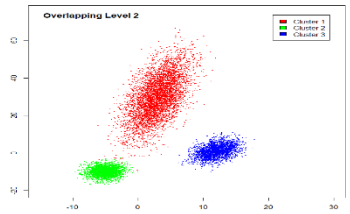
Patterns or clusters of real sample datasets obtaining from different donors might be different even they are conducted in the same flow cytometry experimental design. A real data sample distribution regarding individual markers usually follow a normal distribution different parameters. Therefore, cluster shaped can be symmetric or asymmetric depends on donors and markers used. The clusters might be, such as, circle shaped with different radius or cigar shaped with different width, height, and angle. For this reason, datasets used in these experiments are mimic generated from a real sample dataset.

3.2 Generation of imitative datasets

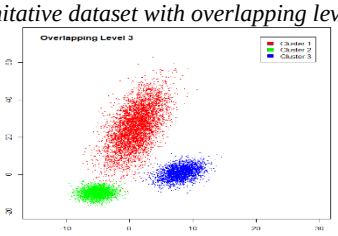
The imitative datasets used in the experiment are 2-dimensional datasets generated by function `rmvnorm (n, mean, sigma)` in RStudio 3.5.2; this function randomly generates data from a multivariate normal distribution, which are often found on FCM data. For this function, three arguments are required: the number of data points (`n`), an average of the data (`mean`), and a covariance matrix (`sigma`). The structure of the imitative datasets consisted of three clusters for each dataset. The number of data points in Clusters 1, 2 and 3 were 5000, 2500 and 2500 respectively. They were generated with four different argument sets, which mean four different overlapping, as shown in Table 1, and generated three times for each set of arguments, so there are 12 datasets used in the experiment. The datasets are divided into four levels of overlapping among each



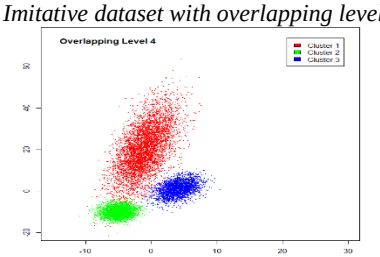
a Imitative dataset with overlapping level 1



b. Imitative dataset with overlapping level 2



c. Imitative dataset with overlapping level 3



d. Imitative dataset with overlapping level 4

Figure 4: Scatter plot of imitative datasets

clusters; examples of these imitative datasets can be seen in Figure 4.

3.3 Result

The imitative datasets are clustered using DBSCAN, OPTICS, FlowGird, and FLOPTICS with user-defined parameters, as can be seen in Table 2. The value of ε for DBSCAN and FlowGrid that provide the best average accuracy result are selected, which are 0.8 and 6.0 respectively.

Table 2: The parameter values for individual techniques

DBSCAN	OPTICS	FlowGrid	FLOPTICS
ε = 0.8 MinPts =10	ε = optimal value MinPts = 10	ε = 6 MinDenB = 3 MinDenC = 40 bin_size = 100	ε = optimal value MinPts = 10 bin_size = 100

All techniques are implemented and run on RStudio 3.5.2 by using the computer with specification mentioned before. The result are presented in Table 3 and some result of FLOPTICS applying to the dataset are shown in Figure 5.

Table 3. The result of applying clustering techniques to imitative datasets

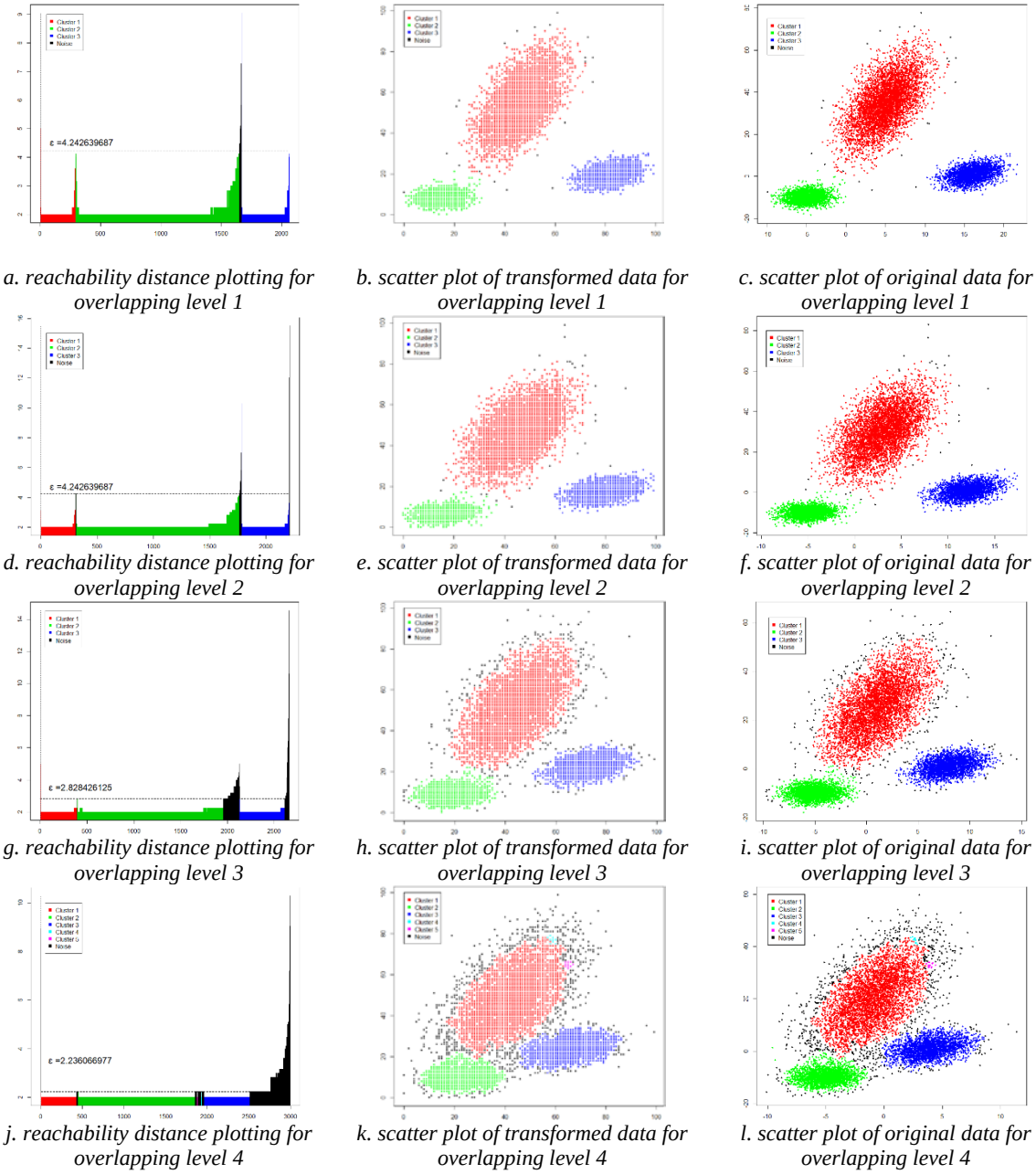
Techniques	Average accuracy (%)				Overall average accuracy (%)	Average Runtime (milli-second)
	Overlapping dataset					
	Level1	Level2	Level3	Level4		
DBSCAN	94.99	94.70	94.80	62.70	86.80	6,728.60
OPTICS	99.93	99.75	98.60	92.07	97.59	1,961.12
FlowGrid	96.00	95.75	95.01	93.59	95.09	723.80
FLOPTICS	99.87	99.49	97.60	90.45	96.85	265.88

4 CONCLUSIONS

According to the result from Table 3, OPTICS provides the best average accuracy, which is 97.59%. FLOPTICS gives higher accuracy result than DBSCAN and FlowGrid. Although OPTICS gives the highest accuracy, it is approximately 7.4 times slower than FLOPTICS technique. FLOPTICS is the fastest technique applying to imitative dataset comparing with DBSCAN, OPTICS, and FlowGrid. In term of the number of user-defined parameters, FLOPTICS require two parameters, which are MinPts and bin_size, while, FlowGrid require four parameters, which are ε, bin_size, MinDenB, and MinDenC. Therefore, FLOPTICS has better performance than the state of art automated gating technique as the FlowGrid technique.

5 FUTURE WORKS

Although FLOPTICS algorithm provide high accuracy and fast running, it can be improved the performance by adding some parameters. In the process of partioning data into equal size of bin, only non-empty bin are process but both of high density bin and low density bin are treated equally. Moreover, core points are identified by consideration of the



a. reachability distance plotting for overlapping level 1
b. scatter plot of transformed data for overlapping level 1
c. scatter plot of original data for overlapping level 1
d. reachability distance plotting for overlapping level 2
e. scatter plot of transformed data for overlapping level 2
f. scatter plot of original data for overlapping level 2
g. reachability distance plotting for overlapping level 3
h. scatter plot of transformed data for overlapping level 3
i. scatter plot of original data for overlapping level 3
j. reachability distance plotting for overlapping level 4
k. scatter plot of transformed data for overlapping level 4
l. scatter plot of original data for overlapping level 4

Figure 5: Some result of FLOPTICS applying to imitative dataset

number of neighbour. It would be better if identification of core points can be decided by not only the number of neighbour but also the density of individual bins.

REFERENCES

Ankerst, M. et al. (1999) *'OPTICS: Ordering Points To Identify the Clustering Structure'*, in Proc. ACM SIGMOD'99 Int. Conf. on Management of Data. Philadelphia.

Bashashati, A. and Brinkman, R. R. (2009) *'A Survey of Flow Cytometry Data Analysis Methods'*, Advances in

Bioinformatics, 2009, pp. 1–19. doi: 10.1155/2009/584603.

Bio-Rad (2018) *Flow Cytometry - User Manual*, *Bioinformatics and Biomedical Engineering*, 2008. ICBBE 2008. The 2nd International Conference on. doi: 10.1002/ejoc.201200111.

Ester, M. et al. (1996) *'A Density-Based Algotithm for Discovering Clusters in Large Spatial Database with Noise'*, Comprehensive Chemometrics, 2, pp. 635–654. doi: 10.1016/B978-044452701-1.00067-3.

Van Gassen, S. et al. (2015) *'FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data'*, Cytometry Part A, 87(7), pp. 636–645. doi: 10.1002/cyto.a.22625.

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), pp. 2052–2058. doi: 10.1093/bioinformatics/bts300.

Groeneveld-Krentz, S. et al. (2016) *'The Role of Machine Learning in Medical Data Analysis. A Case Study: Flow Cytometry'*, (January 2016), pp. 303–310. doi: 10.5220/0005675903030310.

Jahan-Tigh, R. R. et al. (2012) *'Flow Cytometry'*, J. Invest. Dermatol. Nature Publishing Group, 132(10), p. e1. doi: 10.1038/jid.2012.282.

Lee, H. C. et al. (2017) *'Automated cell type discovery and classification through knowledge transfer'*, Bioinformatics, 33(11), pp. 1689–1695. doi: 10.1093/bioinformatics/btx054.

Lo, K., Brinkman, R. R. and Gottardo, R. (2008) *'Automated gating of flow cytometry data via robust model-based clustering'*, Cytometry Part A, 73(4), pp. 321–332. doi: 10.1002/cyto.a.20531.

Moloney, M. and Shreffler, W. G. (2008) *'Special Series : Basic Science for the Practicing Clinician Basic science for the practicing physician : flow cytometry and cell sorting'*, p. 2008.

Qian, Y. et al. (2010) *'Elucidation of seventeen human peripheral blood B-cell subsets and quantification of the tetanus response using a density-based method for the automated identification of cell populations in multidimensional flow cytometry data'*, Cytometry Part B: Clinical Cytometry, 78B(S1), pp. S69–S82. doi: 10.1002/cyto.b.20554.

Rahim, A. et al. (2018) *'High throughput automated analysis of big flow cytometry data'*, 135, pp. 164–176. doi: 10.1016/j.ymeth.2017.12.015.

Ye, X. and Ho, J. W. K. (2018) *'Ultrafast clustering of single-cell flow cytometry data using FlowGrid'*, BMC Systems Biology, 13. doi: 10.1186/s12918-019-0690-2.