Technical Report

TR-2008-003

Sample clustering of flow cytometry data

by

Lin Liu, Li Xiong, James J. Lu, Kim Gernert, Vicki Hertzberg

Mathematics and Computer Science

EMORY UNIVERSITY

Sample Clustering of Flow Cytometry Data

Lin Liu, Li Xiong, James J. Lu
Emory University
Department of Mathematics and Computer Science
Atlanta, GA 30322-1950, U.S.A.
lliu24, lxiong, jlu@mathcs.emory.edu

Kim Gernert
Emory University
The Biomolecular Computing Resource
Atlanta, GA 30322-1950, U.S.A.
gernert@emory.edu

Vicki Hertzberg
Emory University
Department of Biostatistics
Atlanta, GA 30322-1950, U.S.A.
vhertzb@sph.emory.edu

Abstract

Flow cytometry technique produces large, multi-dimensional datasets of individual cells that are helpful for biomedical science and clinical research. Given the size of the data, efficient computational analysis techniques are necessary to assist researchers in understanding and interpreting the data. Automatic cluster analysis of samples based on flow cytometry data is a powerful and promising technique for identifying separate sub-groups of patient samples. To overcome challenges posed by the irregularities and the high dimensions of the data, the current paper explores efficient feature reduction techniques based on regression analysis. Experiments clustering data from the Protective Immunity Project (PIP) shows the effectiveness of the approach.

1. Introduction

Flow Cytometry (FCM) is a technique used in clinical research for studying the immunological status of patients with vaccines or other immunotherapies, for characterizing cancer, for HIV/AIDS infection and other diseases, as well as for research and therapy involving stem cell manipulations. The technique measures the characteristics of single cells, determined by visible and fluorescent light emissions as liquid flow moves the suspended cells pass a laser that emits light at a particular wavelength [10]. The fluorescence emission from each cell is collected and subsequent electrical events are analyzed on a computer that assigns a fluorescence intensity value to each signal in Flow Cytometry Standard (FCS) data files. Each FCS data file thus consists of thousands to millions of multi-parametric descriptions of individual cells.

How such large sets of data points in a highly multidimensional space can be efficiently and systematically analyzed represents a basic yet important challenge. In addition, the computational analysis is often further complicated by the integration of FCS data with other datasets for clinical studies (e.g., the immunological status of patients with vaccines or other immunotherapies). Data mining techniques such as clustering and classification offer promising tools for identifying interesting distributions and patterns of patients based on the large FCM datasets and learning models for predicting patients' immunological responses.

In this paper, we explore an approach for clustering patient samples based on their FCM data. Clustering analysis divides data into meaningful clusters according to various measures of similarity. Recently, some clustering algorithms have been applied to FCM data for clustering cells into groups [18,11]. In contrast, clustering samples based on FCM data has been less explored. Such sample-based learning presents a number of challenges. First, while there may be only tens to hundreds of samples available, the space of potential features consists of thousands of millions of cell intensity data values. This induces an extraordinarily large search space for the parameters of the model. Second, the cells are not ordered uniformly across samples; they may be in any random order. This makes feature modeling a challenge as the samples are not directly comparable across cells.

To address the above challenges, we develop a set of FCM data preprocessing techniques to facilitate effective clustering of patient samples. Specific contributions are as follows. First, we model the features by converting the original cell intensity values to cell-intensity distribution so that they are comparable across samples. Second, we observe the special characteristics of the cell-intensity distribution data and perform a dimension reduction through regression analysis to reduce the number of features significantly. Finally, we evaluate the approach using a set of real data, from the Protective Immunity Project, to demonstrate the effectiveness of our approach. In particular, we show that the dimension reduction significantly improves the clustering analysis both in quality and efficiency. We conduct a comparative analysis of various clustering algorithms.

The rest of the paper is organized as follows. Section 2 describes the datasets used in our study, and our methodology including feature modeling, feature reduction, and the clustering techniques being used. Section 3 presents our experiment setup, evaluation metrics, and the experiment results. Section 4 presents a brief review of related work. Section 5 concludes the paper with a brief summary and a discussion of future directions of our research.

2. Methodology

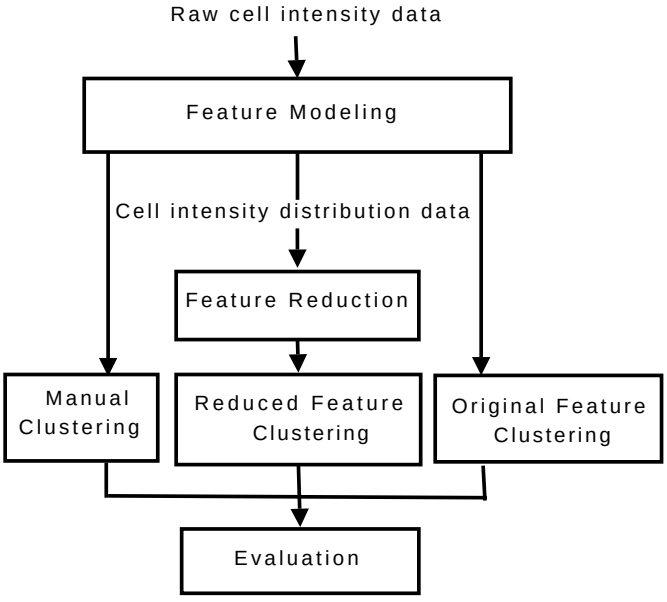

Figure 1. System Overview

Our goal is to cluster a set of samples based on their raw FCS files containing thousands to millions of cell-intensity values. We first model the features by encoding the cell-intensity values into cell-intensity distribution values so that they are comparable across samples. We then perform dimension reduction through regression analysis to reduce the large number of features. Using both the original features and the reduced features, we cluster the set of samples. For evaluation, we compare the computed results against manually clustered data based on domain knowledge and visual inspection. Manual clustering is possible here because of the relative small set of samples available in our study, while our computational technique can be applied to potentially very large sets of samples. A flow diagram of our approach is displayed in **Figure 1**. Below we describe the dataset being used in our study and elaborate on each of these steps.

2.1. Data Description

In our study, 18 samples were collected from 3 patients at different time points (6 samples per patient). Each sample was processed with 6 panels which measure the cell intensity on a set of channels. Each sample corresponds to a FCS file that contains the raw intensity value of cells for all the channels (FSC-A, FSC-H, SSC-A, SSC-H, Time, Comp-FITC-A, Comp-PE-A, Comp-PerCP-A, Comp-PE CY-7-A, Comp-Pacific Blue-A, Comp-APC-A). There are 10^5 cells and 11 channels in each data file. Hence each raw FCS data file contains a $10^5 * 11$ cell-channel intensity matrix where each value is the intensity of a certain cell at a certain channel. A snippet of the raw intensity file is shown in **Table 1**.

Table 1. Raw FCS Data (Cell Intensity)

	FCS-A	FCS-H	Comp-PE-A	...
Cell 1	634	547	1381	...
Cell 2	393	319	1465	...
Cell 3	634	537	1231	...
.....	...	...	...	...

2.2. Feature Modeling

As the cells are not uniquely identified and can be ordered randomly in a raw FCS file, the intensity values are not directly comparable across samples and can not be used directly as features for clustering. To address this, we transformed the absolute intensity values contained in the raw FCS data file into intensity distributions so that they can be compared across samples and used as features for clustering. For example, if sample 1 and sample 2 have the same or similar number of cells with each intensity values at each channel, they would be considered similar to each other. The transformed feature data now contains a $10^3 * 11$ intensity-channel distribution matrix where each value is the count of cells with a certain intensity value at a

certain channel. A snippet of the feature data is illustrated in **Table 2**.

Table 2. Feature Data (Cell Intensity Distribution)

	FCS-A	FCS-H	Comp-PE-A	...
.....	...	...	...	...
150	1	333	0	...
151	0	290	4	...
152	0	275	0	...
.....	...	...	...	...

2.3. Feature Reduction

After we converted the intensity values into intensity distribution values, each sample now contains thousands of data points (features), determined by the range of intensity values. This large number of dimensions poses a challenge for effective and efficient clustering.

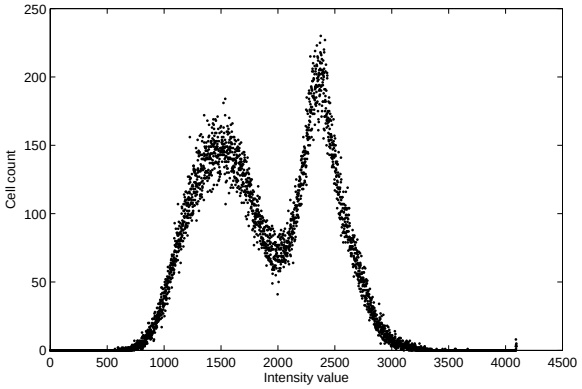

Figure 2. Sample Plot of Feature Data (Intensity-Distribution)

By plotting the intensity distribution values, we observed that all the sample files display bell shaped curves. A plot of a sample intensity distribution file is presented in **Figure 2**. This motivates us to apply regression analysis techniques, in particular, polynomial fitting, to reduce the data. By storing the parameters of the polynomial that represent the original data and discarding the original data points, we can reduce the number of features significantly.

One challenge of this approach is to determine the optimal power for the polynomial fitting. We adopt least square method and tested polynomials with different powers to uncover the optimal power based on two principles, namely, the square error and the efficiency. For example, for channel FSC-A, we discovered that the relevant errors (absolute error/average intensity distribution value) improve significantly when the highest polynomial power increases from 8 to 9 but only marginally from 9 to 10. On the other hand, the CPU time for the polynomial fitting increased marginally when the highest polynomial power increases from 8 to 9 but significantly (doubled) from 9 to 10. Therefore, we settled on 9 for the power in the polynomial fitting for FSC-A. Results of the polynomial fitting for all the channels are shown in **Table 3**. The polynomial power used for each channel is highlighted.

Table 3. Polynomial Fitting for Feature Data

channel	power	relevant error	CPU time (sec)
FSC-A	8	25%	4
	9	14%	5
	10	12%	10
FSC-H	8	30%	4
	9	15%	5
	10	13%	10
SSC-A	8	28%	4
	9	17%	5
	10	12%	10
SSC-H	7	28%	4
	8	14%	4
	9	12%	8
Time	7	37%	4
	8	28%	4
	9	25%	8
Comp-FITC-A	7	33%	4
	8	20%	4
	9	17%	8
Comp-PE-A	7	25%	4
	8	17%	4
	9	15%	8
Comp-PerCP-A	7	29%	4
	8	17%	4
	9	15%	8
Comp-PE CY-7-A	6	29%	3
	7	17%	3
	8	16%	5
Comp-Pacific Blue-A	6	31%	3
	7	16%	3
	8	15%	5
Comp-APC-A	6	25%	3
	7	19%	3
	8	17%	5

2.4. Clustering

We clustered the samples based on the reduced features as well as the original features to evaluate the effect of the dimension reduction. We compared a set of clustering algorithms implemented in Weka[1], an open source data mining toolkit. The clustering algorithms used in our experiments are briefly described next.

[1] Weka. http://www.cs.waikato.ac.nz/ml/weka/

Cobweb is an incremental system for hierarchical conceptual clustering [15]. DBScan (Density-Based Spatial Clustering of Applications with Noise) finds clusters of a minimum specified size and density [5]. EM (Expectation-Maximisation) calculates the probabilities of an instance belonging to clusters [5]. FarthestFirst algorithm is an implementation of the "Farthest First Traversal Algorithm" [7] and looks for approximate clusters [13]. FilteredClusterer offers the possibility to apply filters directly before the clusterer is learned [16]. OPTICS (Ordering Points To Identify the Clustering Structure) creates an augmented ordering of the dataset representing its density based clustering structure [1]. XMeans extends K-means with efficient estimation of the number of clusters [17]. MakeDensityBasedClusterer algorithm is a new density based clustering algorithm [14]. SimpleKMeans is a simple K-means clustering algorithm without parameter estimation [13].

3. Experimental Results

This section presents a set of experiments evaluating the feasibility, effectiveness and cost of our proposed approach. Our goal is to answer the following important questions: 1) Can the regression analysis based dimension reduction help with the clustering analysis with respect to quality and efficiency? 2) Are there differences between different clustering algorithms on our specific dataset?

3.1. Evaluation Metrics

To evaluate the quality of the clustering result, we inspected the intensity distribution curves of each sample and clustered them based on their visual similarity and our domain knowledge. We then used this visual clustering result as a reference against which our computed clusters are compared. In particular, we used the Jaccard score defined as follows.

Let T be the *true* solution and S the solution we wish to evaluate. Let n_{11} denote the number of pairs of elements that are in the same cluster in both S and T, n_{01} the number of pairs that are in the same cluster only in S, and n_{10} the number of pairs that are in the same cluster only in T. The Jaccard Score is defined in the equation below. The result score is in the range of [0,1] with higher score indicating a better clustering quality.

$$D_J(T,S) = \frac{n_{11}}{n_{11}+n_{10}+n_{01}}$$

To evaluate the efficiency of the clustering algorithm, we also measured the CPU time for the clustering process as well as the time for data preprocessing for the feature reduction.

3.2. Impact of Feature Reduction on Quality of Clustering

One goal of the experiment is to verify our hypothesis that the feature reduction will improve the clustering quality by extracting the essential features of the data. We report the quality of clustering measured by Jaccard score for the original features and the extracted features respectively. The FarthestFirst clustering algorithm is used to obtain this set of results and the differences of various clustering algorithms will be reported later.

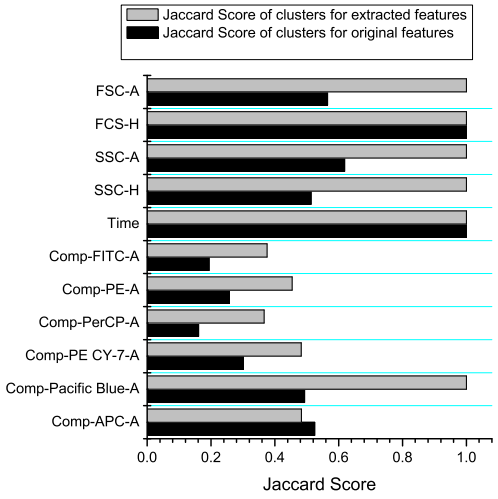

Figure 3. Comparison of Clustering Quality

Figure 3 compares the Jaccard Score of the clustering results based on extracted features and the original features at all channels. It can be observed that the clustering based on extracted parameters achieved a better Jaccard Score than the original features for most of the channels. Indeed, it achieves a perfect score for 6 channels. This verified our hypothesis that polynomial fitting based feature reduction improves the quality of clustering significantly.

3.3. Impact of Feature Reduction on Efficiency of Clustering

In addition to evaluating the impact of feature reduction on the quality of clustering, we also evaluated its impact on the efficiency of clustering through measuring the CPU time.

Figure 4 presents the average CPU time for feature extraction, and for clustering based on extracted features and original features. We observe that clustering based on extracted features significantly shortens the time for clustering (around 6.5 times). In addition, if we consider the overall time for the approach by summing the clustering and feature

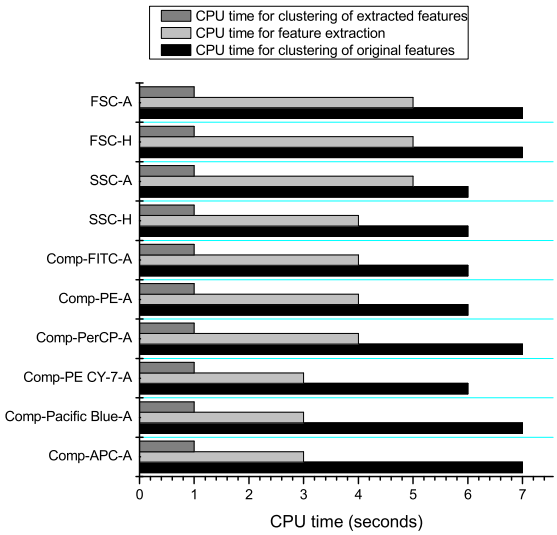

Figure 4. Comparison of Clustering Efficiency

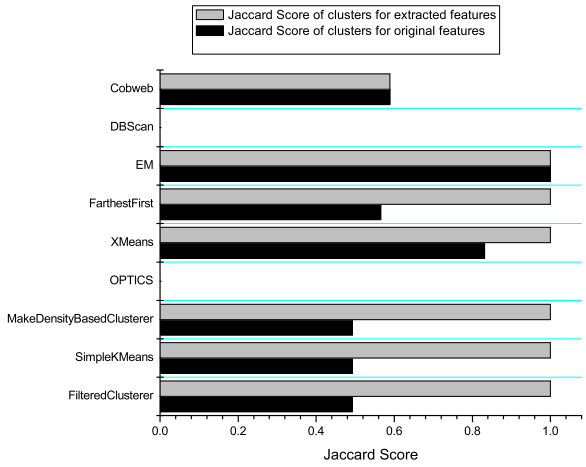

Figure 5. Comparison of Different Clustering Algorithms

extraction time, it still represents significant improvement over the original feature based clustering.

3.4. Effect of Different Clustering Algorithms

The other goal of the experiment is to uncover the differences between different clustering algorithms on the specific dataset.

Figure 5 presents the quality of clustering measured by Jaccard Score for the set of clustering algorithms we used on channel FCS-A. While the different results achieved by different algorithms using the original features warrant further investigation, it can be observed that most of the algorithms provide similar and good results using reduced features. This shows a surprising yet interesting effect of the feature reduction in reducing the sensitivity to different algorithms.

4. Related Work

Some clustering algorithms have been recently applied to FCM data for clustering cells into cell groups. Notably, the feature-guided clustering of multi-dimensional flow cytometry datasets [18] and model-based clustering analysis [11] both focus on clustering cells within FCM data based on their characteristics. Our work focuses on clustering samples based on FCM data.

Data dimension reduction has been applied in a variety of data analysis problems [6]. Principal component analysis (PCA) [8, 10, 2] is the commonly used technique to reduce multidimensional data sets to lower dimensions for analysis. It transforms the data to a new coordinate system such that the greatest variance by any projection of the data lies on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on. PCA can be used for dimensionality reduction in a data set by retaining those lower-order principal components and ignoring higher-order ones. Such low-order components often contain the most important aspects of the data. However, depending on the application this may not always be the case and for clustering analysis, PCA has several limitations [4]. The main limitation is that it ignores class separability as it does not consider the class label of feature vector by simply performing a coordinate rotation that aligns the transformed axes with the directions of maximum variance [4, 12]. Another limitation of PCA is its assumption that principal components are orthogonal with each other [3], and this assumption is also not guaranteed for our datasets. Finally, PCA uses the eigenvectors

of the covariance matrix to find the independent axes of the data under the Gaussian assumption [4, 12, 3]. However, the datasets we used in the experiment do not have Gaussian distribution feature. Based on the specific characteristics of our dataset, we used regression analysis as our data reduction technique which improved both the quality and efficiency for clustering.

5. Conclusion and Future Works

We developed and presented a framework for clustering samples based on flow cytometry data that contain cell intensity values for different channels. We experimentally show that our system produces meaningful results with good efficiency.

While our work is a convincing proof-of-concept, there are several aspects of our system that will be further explored. First, the current system uses polynomial fitting for feature reduction. We would like to adopt other parametric models as well as other dimension reduction techniques and examine their effectiveness on our dataset. Second, different clustering algorithms yielded different results with respect to our evaluation criteria, and a study of the reasons of these differences would be useful. Third, we are currently considering the channels separately for clustering, and a natural extension of the work is to consider all the channels and study their correlations and possibly further reduce the data. In addition, we are also planning to explore temporal data analysis techniques to learn the variances and evolving trend of samples along different time points. Finally, we are integrating the FCM data with clinical datasets and possibly gene expression datasets to perform supervised learning in order to predict patients' immune status.

Acknowledgements

The authors would like to acknowledge the support of the Protective Immunity Project through the NIH grant NO1-AI-50025.

References

[1] Ankerst, M., Breunig M. M., Kriegel H.P., Sander J. OPTICS: Ordering Points To Identify the Clustering Structure, *ACM SIGMOD Int. Conf. on Management of Data* 1999.

[2] Bell, A., and Sejnowski, T. The Independent Components of Natural Scenes are Edge Filters. *Vision Research* 37, 23, 3327-3338, 1997.

[3] Crichtley, F. Influence in principal component analysis. *Journal of Biometrika* 29, 5-14, 1985.

[4] Fukunaga, K. Introduction to Statistical Pattern Recognition. Academic Press. 1990.

[5] Hand, D., Mannila, H. and Smyth, P. Principles of Data Mining, MIT Press, Cambridge, MA. 2001.

[6] Hartigan, J., and Wang, M. A K-means Clustering Algorithm. *Journal of Applied Statistics* 28, 100-108, 1979.

[7] Hochbaum, D., and Shmoys, D. A best possible heuristic for the k-center problem, *Mathematics of Operations Research* 10, 2, 180-184, 1985.

[8] Jolliffe, I. Principal Component Analysis. Springer. 2nd edition. 2002.

[9] Shapiro, H. Practical Flow Cytometry, 4th ed., John Wiley & Sons, Inc., New York. 2003.

[10] Sherlock, G. Analysis of Large-scale Gene Expression Data. *Current Opinion in Immunology* 12, 201-205, 2000.

[11] Simon, U., Mucha, H., and Brggemann, R. Model-Based Cluster Analysis Applied to Flow Cytometry Data. *Innovations in Classification, Data Science, and Information Systems* 69-76, 2006.

[12] Sun, R., Tsung, F., and Qu, L. Evolving Kernel Principal Component Analysis for Fault Diagnosis. *International Conference on Computers & Industrial Engineering* 53, 2, 361-371, 2007.

[13] Tan, P., Steinbach, M. and Kumar, V. Introduction to Data Mining, Addison Wesley. 2006.

[14] Waikato. Weka data mining software. http://www.cs.waikato.ac.nz/ml/weka/

[15] Weka. Cobweb clustering algorithm. http://en.wikipedia.org/wiki/Cobweb.

[16] Weka. FilteredClusterer clustering algorithms. http://weka...sourceforge.net/wekadoc/index.php.

[17] Witten, I. and Frank, E. Data Mining: Practical Machine Learning Tools and Techniques ,Second Edition, Morgan Kaufmann, San Francisco. 2005.

[18] Zeng, Q., Pratt, J., Pak, J., Ravnic, D., Huss, H., and Mentzer, S. Feature-guided Clustering of Multi-dimensional Flow Cytometry Datasets.*Journal of Biomedical Informatics* 40, 3, 325-331. 2007.