The Challenge of Clustering Flow Cytometry Data from Phytoplankton in Lakes

Stefan Glüge[1], Francesco Pomati[2], Carlo Albert[2], Peter Kauf[1], Thomas Ott[1]

[1] ZHAW Zurich University of Applied Sciences, Switzerland
[2] Swiss Federal Institute of Aquatic Science and Technology (EAWAG), Switzerland

Abstract. Flow cytometry (FC) devices count and measure cells in fluids in an automated procedure. In this paper we present our work in progress on the clustering of FC data. We compare standard clustering algorithms such as K-means, Ward's clustering, etc., to the more advanced approach of sequential superparamagnetic clustering (SSC). We found Ward's hierarchical clustering to perform best regarding internal cluster validation measures, while SSC yielded the best results based on the visual inspection of the clustering results.

Keywords: Clustering, Sequential Superparamagnetic Clustering, Flow Cytometry, Phytoplankton

1 Introduction

Flow cytometry (FC) devices basically count/measure cells in fluids. Usually, cells pass a single wavelength laser beam and emit a specific optical signal depending on their structure and fluorescence. The area of FC applications is wide, e.g. monitoring of industrial processes [8], aquatic system monitoring [6], and medical research [4].

At the Swiss Federal Institute of Aquatic Science and Technology (Eawag) a FC approach is applied to survey phytoplankton occurrences in lakes [6]. The focus is on understanding and predicting the effects of environmental change on biodiversity. One mayor issue is the need to introduce a trait-based analysis using the characteristics of individual phenotypes rather than species. Thereby one would create a measures of functional diversity, which is a predictor of ecosystem functioning across a range of communities.

In this paper we present our work in progress on the clustering of FC data, collected in the pre-alpin Lake Zurich (Switzerland). We compare several standard clustering algorithms (e.g. K-means) to the more advanced approach of SSC [5].

The data poses several interesting challenges, that is, a large amount of unlabeled samples, clusters of different shapes and densities, and an unknown number of classes/categories. Further, the data contains a significant number of outliers.

Our results induce that SSC is a promising approach to the automatic analysis of the complex FC data. Even though, for the application in field studies some labeled data is needed for a more reliable evaluation of the clustering results.

2 Methods

2.1 Flow Cytometry Data

The phytoplankton dataset used in this study consists of water samples collected from April to December 2009 in monthly intervals in Lake Zurich at 14 different water depths from $0, \ldots, 135$m. For flow-cytometry analysis 50ml of sampled water were fixed with a filter-sterilized solution of paraformaldehyde and glutaraldehyde (0.01 and 0.1% final concentration, pH 7). A scanning flow-cytometer from Cytobuoy[3] was used for counting and characterization of phytoplankton particles. It allows the analysis of pulse-signals providing 54 descriptors of 3D structure and fluorescence (FL) profile for each particle [7]. Raw Cytobuoy data were visually inspected for the distribution of FL signals in order to set threshold levels to extract FL particles (phytoplankton) with a size larger than 1µm.

2.2 Data Preprocessing

The whole data set consists of 73055 samples, each represented by 54 features (cf. Sec. 2.1). At first the raw data was centered and normalised to its standard deviation.

Some of the features are highly correlated, or anti-correlated respectively. Therefore, we reduced the feature space by applying a principle component analysis (PCA), ending up with 20 principle components (PCs) representing 90.75% of the variance in the data. For a comparison based on visual inspection we also applied the clustering algorithms on the first 3 PCs of the data. Further, clustering was done on 5% of the data (3653 samples) that where chosen randomly from the complete data set. We had to limit ourself on this subset due to constraints on memory and computation time.

Figure 1 shows the first three PCs of the selected 5% of the data to give an impression of the clustering problem we face. Looking on the data one can identify at least three different clusters. There is a small cluster (C1) on the left at PC1 ≈ 2 and PC2 ≈ -2 separated quite well, another big dense cluster (C2) at the bottom lying more or less in the PC2-PC3–plane, and one big cluster (C3) in the middle of the plot surrounded by a rather unspecific halo that was considered to be noise. The clusters where labelled manually to create values that allowed us an external cluster validation. However, these labels are prone to errors, especially in the regions of the edges the labelling is rather subjectively.

2.3 Clustering Algorithms

For our comparative study we considered four standard clustering algorithms that are usually used in FC data analysis (cf. [1]), namely K-means, Partitioning Around Medoids (PAM), model based clustering (mclust), and Ward's hierarchical clustering (ward). Further, we applied SSC, which is a more recent and advanced approach [5].

[3] Woerden, the Netherlands; http://www.cytobuoy.com

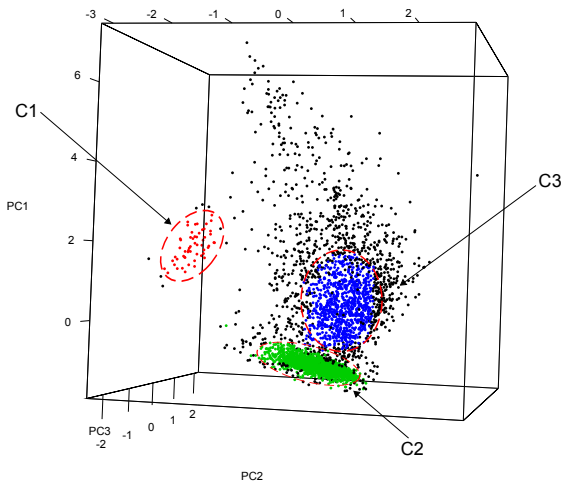


Fig. 1: 5% of the flow cytometry (FC) data projected on its first three principle components (PCs). At least three clusters can be identified by visual inspection.

K-means clustering is a method of vector quantization. It aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

PAM is quite similar to K-means. Both algorithms break the datasets into groups, and both try to minimize the error. However, PAM works with Medoids, that are an entity of the dataset that represent the group in which it is inserted, while K-means works with Centroids [2].

mclust is based on distribution models. We applied Gaussian mixture models using the expectation-maximization algorithm. To obtain a hard clustering, samples were assigned to the Gaussian distribution they most likely belong to.

ward is a hierarchical cluster method. The criterion for choosing the pair of clusters to merge at each step is based on the minimum of some inter-cluster distance measure. Ward used the minimum variance method which minimizes the total within-cluster variance [9].

SSC is based on superparamagnetic clustering enhanced by a sequential procedure to select the most natural clusters [5].

Superparamagnetic Clustering can be description via Potts spin models. For N samples to be clustered with pairwise affinities d_{ij}, an inhomogeneous grid of Potts spins is constructed in the following way: each sample i is represented by one site of the grid with Potts spin variable $s_i \in \{1, \ldots, q\}$. q is typically set to 10 or 20. Each spin is symmetrically coupled to its k nearest neighbours. The

coupling strength J_{ij} is a decreasing function of d_{ij},

$$J_{ij} = J_{ji} = \frac{1}{\hat{K}} \exp\left(\frac{-d_{ij}^2}{2a^2}\right). \tag{1}$$

$\hat{K}$ is the average number of coupled neighbours per site. a is a local length scale which is set by default to the average distance between coupled spins.

Each spin configuration is characterized by an energy expressed by the Potts spin Hamiltonian

$$H(s) = \sum_{(ij)} J_{ij} \left(1 - \delta_{s_i s_j}\right), \tag{2}$$

where the sum runs over all connections (ij) and s denotes a spin configuration. The system is considered in the canonical ensemble. The probability for a certain spin configuration is thus given by the Boltzmann/Gibbs distribution

$$p(s) = \frac{1}{Z} \exp\left(\frac{-H(s)}{T}\right), \tag{3}$$

where the partition function $Z = Z(T)$ serves as a normalization factor.

At a given temperature T, clusters are identified with the help of the pair correlation: two points i and j belong to the same cluster if the pair correlation

$$G_{ij} = \sum_{s} p(s) \delta_{s_i s_j} \tag{4}$$

exceeds a given threshold Θ.

Clear clusters express themselves as regions of order that are stable over a substantial range of T. The idea is to choose the clusters that have the largest T-range extensions (denoted by T_{cl}). Consequently, T-stability s_T of a cluster is defined as

$$s_T = \frac{T_{\mathrm{cl}}}{T_{\max}}, \tag{5}$$

where $T_{\max}$ is the temperature of the paramagnetic transition. Hence, s_T represents the stability of a cluster in relation to the stability of the whole set. However, the most natural clusters may not be the most stable ones. Different densities, shapes, and sizes of the clusters result in different ranges of temperature where they occur. Consequently, often natural clusters emerge only for short T-ranges when dense superclusters break-up at higher temperatures. The sequential procedure overcomes this problem by reclustering the most stable cluster in terms of s_T with readjusted weights. That means, the most stable cluster is clustered in two new separate sessions, and the connectivity and weights are independently redetermined for each set according to the criterion of k nearest neighbours and (1).

2.4 Cluster Validation Measures

As we worked on originally unlabeled data we relied on internal cluster validation measures. These were Connectivity and the Dunn Index because they are well known and widely used in practice.

Unfortunately, internal validation measures usually have the drawback that they can identify only well separated hyper sphere shaped clusters. This is because the indices measure the variance of the clusters around some representative points [3].

In our case the clusters are of arbitrary shape and may not have representative centre point. Therefore, we evaluated the clustering results concerning our expected result (cf. Fig. 1) using the Jaccard coefficient. Further, two qualitative criteria were evaluated, that are:

- QC1 are the *three* clusters in Fig. 1 separated (yes/no)
- QC2 is the background noise separated (yes/no)

Connectivity measures to what extent observations are placed in the same cluster as their nearest neighbours in the data space. The connectivity has a value between 0 and ∞, and should be minimized.

Dunn Index is the ratio of the smallest distance between observations not in the same cluster to the largest intra-cluster distance. The Dunn Index has a value between 0 and ∞, and should be maximized.

Jaccard Coefficient is defined as: $J_{\text{coeff}} = \frac{n_{11}}{n_{11}+n_{10}+n_{01}}$, where n_{11} is the number of observation pairs where both observations are in both clusterings, n_{10} is the number of pairs where the observations are in the first clustering but not the second, and n_{01} is the number of pairs where the observations are in the second clustering but not the first. The index takes values between 0 and 1, and is maximized when both clusterings are identical.

For K-means, PAM, mclust and Ward's clustering we evaluated the clustering according to Connectivity and Dunn Index for $k = 2 \dots 20$ clusters. Then the best results where evaluated according to the qualitative criteria and the Jaccard Coefficient. For SSC the number of clusters to be found cannot be set directly, but is controlled by several parameters. The parameters were set in trail-and-error manner to achieve the most reasonable result according to visual inspection.

3 Results

Table 1 summarises results for the different clustering algorithms. The evaluation according to the Jaccard coefficient and qualitative criteria (QC1,QC2) was done on the clustering result that provided the best Dunn Index for each method.

Figure 2 shows the results of Ward's clustering which are the best according to the internal cluster validation measures. In Fig. 2a only two clusters are

Table 1: Results for the different clustering algorithms for data samples using 3/20 PCs. Connectivity and Dunn Index are shown together with the number of clusters k. The cluster results with the highest Dunn Index are also evaluated according to Jaccard coefficient and two qualitative criteria QC1 (C1-C3 found) and QC2 (background noise detected).

	PCs	Dunn Index / k	Connectivity / k	Jaccard	QC1	QC2
K-means	3	0.013 / 3	123.00 / 2	0.520	no	no
	20	0.047 / 4	260.90 / 2	0.554	no	no
PAM	3	0.009 / 3	109.36 / 2	0.568	no	no
	20	0.041 / 4	287.76 / 2	0.561	no	no
mclust	3	0.006 / 3	166.84 / 2	0.631	no	yes
	20	0.029 / 2	857.70 / 2	0.574	no	no
ward	3	0.012 / 5	40.52 / 2	0.580	no	no
	20	0.052 / 5	148.97 / 2	0.556	yes	no
SSC	3	0.003 / 4	299.69 / 2	0.642	yes	yes
	20	0.002 / 3	1249.15 / 3	0.389	no	yes

found. These clusters are separated quite well, which leads to a low connectivity. However, we would expect at least four clusters (C1-C3 plus background noise). In Fig. 2b C1-C3 are found while especially C1 dose not separate very well. Additionally, C3 and background are subdivided in 3 separate clusters.

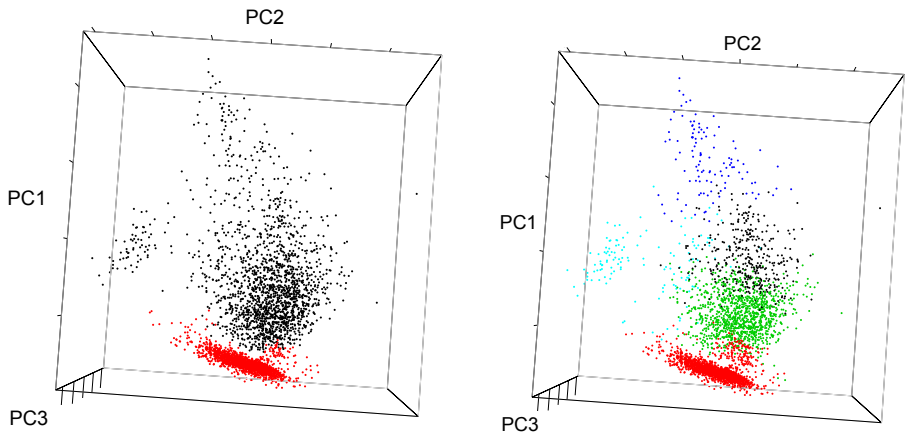
Finally, Fig. 3 shows the result of SSC which performed best regarding the Jaccard coeffients and our qualitative clustering measures QC1 and QC2. All cluster C1-C3 are identified while the remaining points are grouped together in one cluster that we consider to be noise. Obviously it is hard make a clear decision whether some specific point should be part of a cluster or rather is noise.

4 Discussion

Our study shows that the FC data presents an interesting challenge for cluster algorithms. We have to deal with a large amount of unlabeled data. The clusters C1-C3 are of different shapes and densities, and the data contains a significant number of outliers.

We can conclude that Ward's hierarchical clustering performs best regarding the internal cluster validation measures. However, those measures are of limited significance as they work only on separated hyper sphere shaped clusters (cf. Fig. 2). Therefore, we introduced qualitative measures based on the visual inspection of the clustering results and compared the clustreing against our expected clustering using the Jaccard coefficient. According to these handcrafted, but reasonable, measures SSC yielded the best results (cf. Fig. 3).

The Jaccard coefficient of the model based clustering (mclust) with three PCs is quite high with $J_{\text{coeff}} = 0.631$ (cf. Tab. 1). Cluster C1 was not found by the algorithm. As this cluster is small (cf. Fig. 1), it was of limited influence



(a) (b)

Fig. 2: Ward's clustering result with best Connectivity 40.52 (a) and best Dunn Index 0.052 (b).

on the Jaccard coefficient. However, the difference to the expected clustering is rather significant as C1 is lying far outside and should be easily separable. The results of SSC with 20 PCs is also remarkable as $J_{\text{coeff}} = 0.389$ is quite low. This is due to the fact that C2 and C3 where merged in one cluster while C1 and the background were identified well. C2 and C3 are fairly large (cf. Fig. 1) which results in a low Jaccard coefficient. Based on this observation we hypothesise that C2 and C3 lie close together in the higher dimensional space and are not as different as they appear in the three dimensional projection.

Our future work is focused on the improvement of the evaluation of the clustering results. One possibility could be the definition of new internal measures. However, this Problem is a research field on its own. Further, we seek labels based on the biological origin of the data that allow a well-grounded external evaluation. Concerning the clustering algorithms itself we are working on a better implementation of SSC, such that it becomes applicable to larger datasets with reasonable computational costs.

References

1. Boddy, L., Wilkins, M.F., Morris, C.W.: Pattern recognition in flow cytometry. Cytometry 44(3), 195–209 (Jun 2001)
2. Kaufman, L., Rousseeuw, P.: Clustering by Means of Medoids. Reports of the Faculty of Mathematics and Informatics. Delft University of Technology, Fac., Univ. (1987)
3. Legány, C., Juhász, S., Babos, A.: Cluster validity measurement techniques. In: Proceedings of the 5th WSEAS International Conference on Artificial Intelligence,

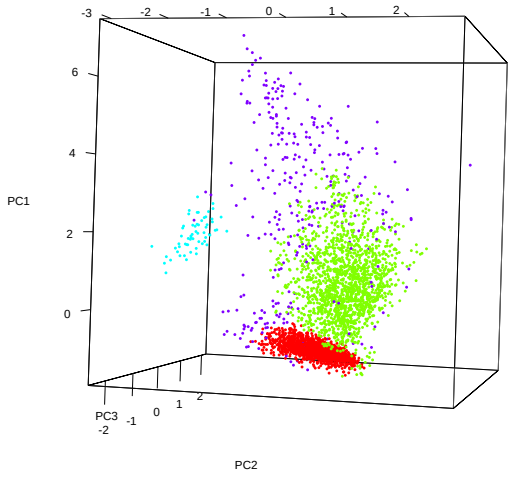


Fig. 3: SSC clustering result of flow cytometry (FC) data based on the first three principle components.

Knowledge Engineering and Data Bases. pp. 388–393. AIKED'06, World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, Wisconsin, USA (2006)
4. Mandy, F.F.: Twentyfive years of clinical flow cytometry: Aids accelerated global instrument distribution. Cytometry Part A 58(1), 55–56 (Mar 2004)
5. Ott, T., Kern, A., Steeb, W.H., Stoop, R.: Sequential clustering: tracking down the most natural clusters. Journal of Statistical Mechanics: Theory and Experiment 2005(11), P11014 (Nov 2005)
6. Pomati, F., Jokela, J., Simona, M., Veronesi, M., Ibelings, B.W.: An automated platform for phytoplankton ecology and aquatic ecosystem monitoring. Environmental Science Technology 45, 9658–9665 (Nov 2011)
7. Pomati, F., Kraft, N.J.B., Posch, T., Eugster, B., Jokela, J., Ibelings, B.W.: Individual cell based traits obtained by scanning flow-cytometry show selection by biotic and abiotic environmental factors during a phytoplankton spring bloom. PLoS ONE 8(8), e71677 (Aug 2013)
8. Urano, N., Nomura, M., Sahara, H., Koshino, S.: The use of flow cytometry and small-scale brewing in protoplast fusion: Exclusion of undesired phenotypes in yeasts. Enzyme and Microbial Technology 16(10), 839–843 (Oct 1994)
9. Ward, J.H.: Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association 58(301), 236–244 (1963)