Clustering environmental flow cytometry data by searching density peaks

Peter Rubbens PETER.RUBBENS@UGENT.BE
Willem Waegeman WILLEM.WAEGEMAN@UGENT.BE
KERMIT, Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, Coupure Links 653, B-9000, Belgium

Ruben Props RUBEN.PROPS@UGENT.BE
Nico Boon NICO.BOON@UGENT.BE
Center for Microbial Ecology and Technology (CMET), Ghent University, Coupure Links 653, B-9000, Belgium

Keywords: clustering, density estimation, dimensionality reduction, environmental microbiology, flow cytometry

Abstract

Microbial single cells can be characterized by their phenotypic properties using flow cytometry. Therefore flow cytometry can be used to analyze various aspects of environmental microbial communities. In recent years, researchers have focused on fully exploiting the multivariate data that such analyses generate. As they are interested in the diversity of an environmental sample, we need a proper estimation of the number of species and their abundances. We modified a recently published algorithm to estimate the microbial diversity based on flow cytometry data. After giving a brief sketch of the problem setup, we will review this algorithm alongside its various implementations. Moreover we will present our current implementation combined with future challenges we foresee.

1. Introduction

Scientists are leaning more and more towards to the use of flow cytometry (FCM) to analyze microbial communities in an environmental context (De Roy et al., 2012; Props et al., 2016). Using FCM, phenotypic properties of single cells can be measured using scatter signals and fluorescence intensity (Müller & Nebe-von Caron, 2010), resulting in a multiparametric description of every cell.

Microbiologists and environmentalists are interested in measuring the biodiversity of a sample, expressed by the number of species and by the evenness of a species community, which can be calculated from the relative abundances. Opposed to synthetic microbial communities, we often do not know which species to encounter in environmental samples. This means that in order to determine the diversity of an environmental community, we are left with unsupervised methods such as clustering in order to make an estimation. A variety of algorithms already exists in the FCM literature (Aghaeepour et al., 2013), however most of them are developed in a medical context. A recent study has shown that these state-of-the-art FCM clustering algorithms do not achieve optimal results when applied to time series of environmental flow cytometry data (Hyrkas et al., 2016).

Recently an algorithm able to deal with non-spherical clusters of varying sizes was proposed (Rodriguez & Laio, 2014). Although the general approach is intuitive, various implementations exist, and the most optimal implementation seems to be domain-specific. Therefore, a brief review of the algorithm will be given in the next part. Conclusively our current implementation is presented, for which preliminary results are promising.

2. Algorithm overview

The algorithm is built in terms of the local density ρ and the distance to the nearest point with a higher

Appearing in *Proceedings of Benelearn 2016*. Copyright 2016 by the author(s)/owner(s).

density δ. The basic assumption of the algorithm is that different density peaks resemble with different clusters. Peaks are identified having both large ρ and δ, quantified by a *decision function* $\gamma = \rho \times \delta$. Points are next assigned according to a *walk down the hill* principle, i.e., to those clusters to which their nearest neighbor with higher density is assigned to.

Whereas the general principles are quite clear, variations in implementation appear on two levels. First, ρ can be determined in various ways. Whereas the original implementation uses a hard threshold, alternative implementations already exist to implement a soft threshold, e.g. by using a Gaussian kernel function (Du et al., 2016; Wang et al., 2016), or by performing a kernel density estimation (KDE) (Wang & Xu, 2015).

Second, there are two main problems concerning the recognition of genuine density peaks As (Liang & Chen, 2016) note:

- There is no quantitative way of determining a 'peak-distinguishing' threshold for γ.
- γ might identify so-called 'pseudo' cluster centers, points which have a large value for γ but in fact do not constitute peaks.

Several solutions have been proposed (Wang & Xu, 2015; Chen et al., 2016; Cheng et al., 2016; Jia et al., 2016), all which seem suitable for the application the method is applied to, but these criteria do not seem sufficient or applicable to our problem set-up. In the next section we will illustrate our implementation of the algorithm and motivate choices we made along the way.

3. Current implementation & preliminary results

1. *Data preprocessing.* During the data preprocessing step, data is often transformed and normalized (O'Neill et al., 2013). We perform a hyperbolic arcsine transformation after which we standardize our data.
2. *Dimensionality reduction.* In our case we have data in 12 dimensions. In order to cope with the *curse of dimensionality*, we first perform a dimensionality reduction technique. For now we have gained the best results using *Kernel Principal Component Analysis* (Scholkopf et al., 1998), using the first three or four components.
3. *Density estimation ρ.* As a density estimation we choose to perform KDE with a Gaussian kernel in combination with a grid search in order to obtain an optimal bandwidth h. In this way we obtain a continuous density function. A future implementation might follow the method suggested by (Du et al., 2016), which uses k-nearest neighbors to incorporate more local properties when estimating ρ.

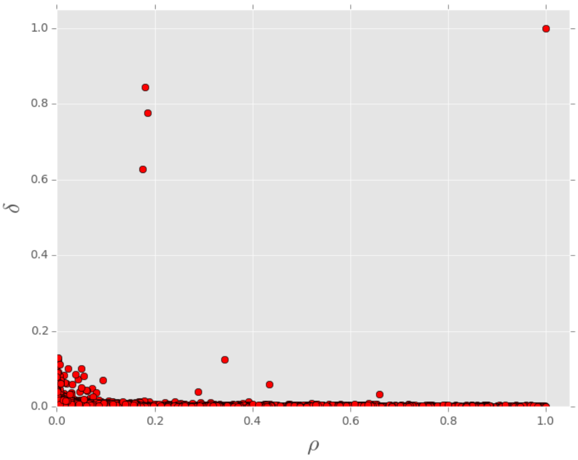

Figure 1. Decision graph $\rho - \delta$ for one sample taken from (Props et al., 2016).

4. *Calculate δ.* Having estimated ρ, it is straightforward to determine δ. For now we use the Euclidean distance to determine the nearest neighboring point with higher density, however a different distance metric might also be used. Having calculated ρ and δ (and after normalizing them), we are able to visualize the *decision graph* $\rho - \delta$, see Fig. 1.
5. *Discerning density peaks and clustering the data.* So far we have not found a suitable decision function or decision criteria. Using available metadata we can fit a threshold for δ above which points are identified as peaks. Using this approach we are able to achieve comparable results as reported by the fingerprinting method in (Props et al., 2016).

4. Future challenges

The current challenge lies in optimizing the decision boundary of the decision graph $\rho - \delta$ for the identification of actual peaks. As the number of species in an environmental sample can be bigger than 100 species, and as the abundances can vary up to three orders of magnitude, a more sophisticated implementation of the algorithm will be needed, as the goal is to retain to a fully unsupervised learning approach.

Acknowledgments

This research was supported by the Special Research Fund (BOF) from Ghent University. The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by Ghent University, the Hercules Foundation and the Flemish Government department EWI.

References

Aghaeepour, N., Finak, G., Hoos, H., Mosmann, T. R., Brinkman, R., Gottardo, R., Scheuermann, R. H., Consortium, F., & Consortium, D. (2013). Critical assessment of automated flow cytometry data analysis techniques. *Nature Methods*, *10*, 228–238.

Chen, Y., Zhao, P., Li, P., Zhang, K., & Zhang, J. (2016). Finding communities by their centers. *Scientific Reports*, *6*.

Cheng, Q., Liu, Z., Huang, J., & Cheng, G. (2016). Community detection in hypernetwork via Density-Ordered Tree partition. *Applied Mathematics and Computation*, *276*, 384–393.

De Roy, K., Clement, L., Thas, O., Wang, Y., & Boon, N. (2012). Flow cytometry for fast microbial community fingerprinting. *Water Research*, *46*, 907–919.

Du, M., Ding, S., & Jia, H. (2016). Study on density peaks clustering based on k-nearest neighbors and principal component analysis. *Knowledge-Based Systems*, *99*, 135 – 145.

Hyrkas, J., Clayton, S., Ribalet, F., Halperin, D., Armbrust, E. V., & Howe, B. (2016). Scalable clustering algorithms for continuous environmental flow cytometry. *Bioinformatics*, *32*, 417–423.

Jia, S., Tang, G., Zhu, J., & Li, Q. (2016). A novel ranking-based clustering approach for hyperspectral band selection. *IEEE Transactions on Geoscience and Remote Sensing*, *54*, 88–102.

Liang, Z., & Chen, P. (2016). Delta-density based clustering with a divide-and-conquer strategy: 3DC clustering. *Pattern Recognition Letters*, *73*, 52–59.

Müller, S., & Nebe-von Caron, G. (2010). Functional single-cell analyses: flow cytometry and cell sorting of microbial populations and communities. *FEMS Microbiology Reviews*, *34*, 554–587.

O'Neill, K., Aghaeepour, N., Spidlen, J., & Brinkman, R. (2013). Flow cytometry bioinformatics. *PLOS Computational Biology*, *9*.

Props, R., Monsieurs, P., Mysara, M., Clement, L., & Boon, N. (2016). Measuring the biodiversity of microbial communities by single-cell analysis. *Methods in Ecology and Evolotion*, n/a–n/a.

Rodriguez, A., & Laio, A. (2014). Clustering by fast search and find of density peaks. *Science*, *344*, 1492–1496.

Scholkopf, B., Smola, A., & Muller, K. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, *10*, 1299–1319.

Wang, M., Zuo, W., & Wang, Y. (2016). An improved density peaks-based clustering method for social circle discovery in social networks. *Neurocomputing*, *179*, 219–227.

Wang, X.-F., & Xu, Y. (2015). Fast clustering using adaptive density peak detection. *Statistical Methods in Medical Research*, 1–14.