Clustering genetic expression time series

Gonzalo Vidal

December 2019

1 Introduction and state of the art

Machine learning (ML) is an application of artificial intelligence (AI) which study algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions (Kubat, 2017). This provides the system the ability to automatically learn and improve from experience without being explicitly programmed. We can divide this field in supervised learning and unsupervised learning, with some special cases like semi-supervised learning and reinforcement learning.

Clustering is a group of unsupervised machine learning methods that automatically group similar objects into sets (Jain, 2010). K-means is an algorithm that clusters data by trying to separate samples in groups of equal variance, minimizing a criterion known as the inertia (within clusters sum of squares)(Hartigan Wong, 1979). The algorithm requires the number of clusters (K) to be specified. It can be seen as an special case of Gaussian mixture model with equal co-variance per component, scales well to large number of samples and has been used across a large range of application areas in many different fields.

Other clustering methods differ in the way to make clusters. Hierarchical clustering is a general family of clustering algorithms that builds nested clusters by merging or splitting them successively (Shalizi, 2009). The hierarchy clusters can be represented as a tree (dendrogram). The root represent the cluster that contain all the data and the leaves are clusters with only one sample. In this group of methods you can not set the number of clusters to be made so you need to choose based in other metrics like linkage (a distance criteria). Density-Based Spatial Clustering of Applications with Noise (DBSCAN) finds core samples of high density and expands clusters from them.It can be seen as an special case of Ordering Points to Identify the Clustering Structure (OPTICS). Good for data which contains clusters with similar density.

Genetic expression in bacteria vary over time and change depending on the phase in which they are (lag, exponential or stationary) due to the expression of different transcription factors (TF) and metabolic changes. In Synthetic Biology we use fluorescent proteins to obtain the expression rate of genes in an indirect way, that can be seen as a time series (ts). A recent work (French, Coutts, Brown, 2018) use R package MFuzz, a noise-robust soft clustering algorithm (Kumar Futschik, 2007) developed for microarray ts, to cluster folds of change of genetic expression in response to a panel of 15 antibiotics. Other group develop Dirichlet process Gaussian process mixture model (DPGP) in microarray microbial organisms models and RNA-seq from human cell lines (McDowell et al., 2018). This method effectively identified disjoint clusters of ts gene expression on intensive simulations and compare favourably to existing methods, is robust to non-Gaussian marginal observations and includes measures of uncertainty.

Evaluate the performance of a clustering algorithm is not as trivial as counting the number of errors or the precision and recall of a supervised classification algorithm.

In particular any evaluation metric should not take the absolute values of the cluster labels into account but rather if this clustering define separations of the data similar to some ground truth set of classes or satisfying some assumption such that members belong to the same class are more similar than members of different classes according to some similarity metric. There are different measurements depending if you know the ground of truth or not. Mutual information based scores is based on the knowledge of the ground of truth and our clustering algorithm assignments to the samples and is a function of the agreement of this two (Goldberger, Gordon, Greenspan, 2006). Silhouette coefficient perform the evaluation using the model itself ignoring the ground of truth (Rousseeuw, 1987). A higher Silhouette Coefficient relates to a model with better defined clusters.

2 Hypothesis

Clusters of gene expression time series will show the intrinsic expression dynamics of promoter families.

3 Objectives

3.1 General objectives

Cluster genetic expression time series

3.2 Specific objectives

Cluster genetic expression time series data
Evaluate the optimal cluster number (K)
Plot the optimal clusters with the barycenters
Assign labels to promoters
Analyze the biological and statistical meaning of clusters

4 Methods

4.1 Data

The data were compiled in R, resulting in time course matrices of raw fluorescence data. Low-span LOESS regression where fit to the data, in order to limit noise in downstream calculation. The minimal edge effects of overall fluorescence intensities were normalized out using for high-density colony array normalization. This method divides colony fluorescence by interquartile means of rows and columns across the plate, this standardized fluorescence across plates 1 (French, Coutts, Brown, 2018).

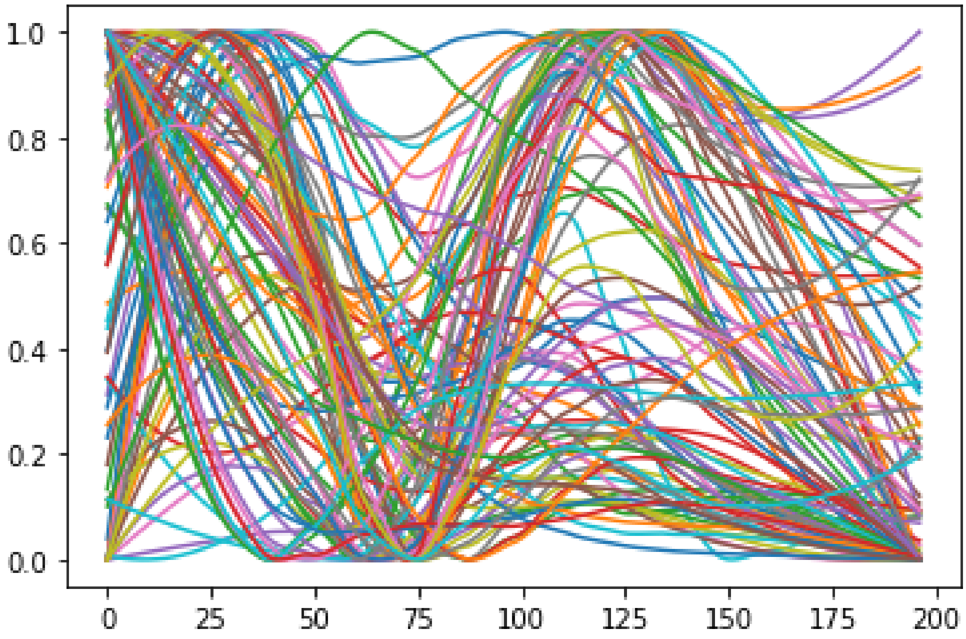

Figure 1: Promoter expression of untreated dataset. Promoters are labeled with different colors. X axis are arbitrary units of fluorescence and Y is time, and delta time is 5 minutes.

4.2 Package

tslearn is a Python package that provides machine learning tools for the analysis of time series build on scikit-learn, numpy and scipy libraries. (Tavenard, Romain and Faouzi, Johann and Vandewiele, 2017)

4.3 Clustering

K-Means clustering algorithm provides good results in a wide range of areas and can be used as an exploratory method. Euclidean distance metric was chosen over Dynamic Time Warping (DTW) and Soft-DTW because produces smoother barycenters (cluster mass center or average) and is the fastest among them (Pedregosa et al., 2011).

4.4 Validation

The Sihouette Coefficient or Silhouette score is calculated using the mean intra-cluster distance (a) and the mean nearest-cluster distance (b) for each sample using $(b - a)/max(a, b)$. The best value is 1 and the worst value is -1. Values near 0 indicates overlapping clusters. Negative values generally indicate that a sample has been assigned to the wrong cluster, as a different cluster is more similar. To be in concordance with the clustering algorithm also use Euclidean distance. One of the best validation if you do not know the ground of truth(Pedregosa et al., 2011).

5 Results

Genetic expression data were clustered with K-Means with K equal to 2, these clusters show genes that turn on and turn off over time 2. Using K equal 10 the clusters are more informative in therms of the different dynamics of expression but replicates are not always clustered together 3.

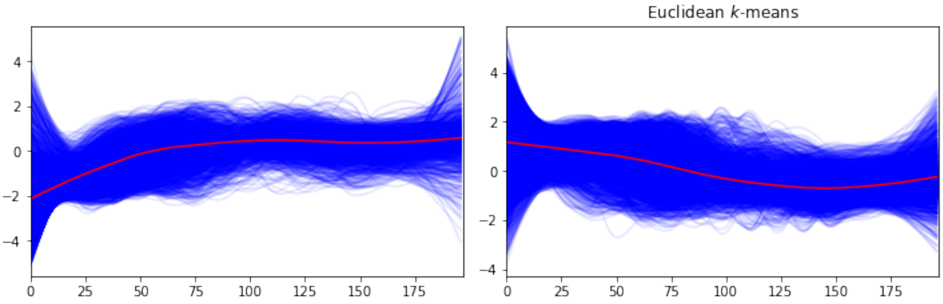

Figure 2: K-Means with 2 clusters. In blue is the data, in red the barycenter, X axis are arbitrary units of fluorescence and Y is time, and delta time is 5 minutes. The metric is Euclidean distance. The algorithm start 10 times in different random point and we show the try with less inertia.

5.1 Data description

Data consists on fluorescent measurements over time. It was obtained with PFI-box, an Open-Source high-throughput fluorescence imaging system for high-density colony array of microorganisms. The samples were placed in 6144 plate, with Luria-Bertani (LB) agar solid media, with *E. Coli* K-12 MG1655. These bacteria has a GFP promoter-reporter fusion library that contains about 1800 different promoters and can asses global transcriptional response. The data was acquired with 5 minutes temporal resolution over 18 hours of growth. From the dataset a subset of non treatment expression was taken. The number of rows or time series was 6144.

5.2 Data pre-processing

Using pandas DataFrame attribute dropna, all the rows with missing values were discarded. All the constant value data was removed. Time series were scaled to have the same mean μ in each dimension.The final number of rows is 5106

5.3 Developed model

Unsupervised cluster label was performed with K-Means. The algorithm was performed for exploratory K from 2 to 50, after this all the analysis were made with K between 2 to 15 because here are the more interesting behaviour of the metrics 4. The algorithm iterates until the inertia difference between iteration is less than

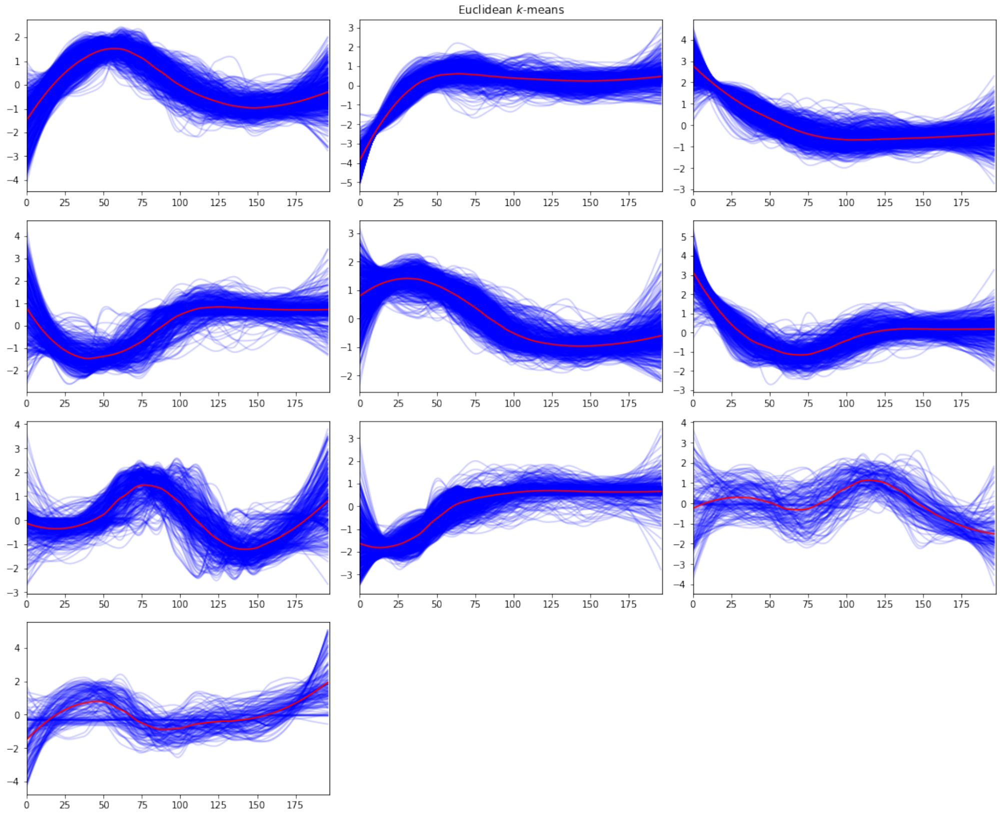

Figure 3: K-Means with 10 clusters. In blue is the data, in red the barycenter, X axis are arbitrary units of fluorescence and Y is time, and delta time is 5 minutes. The metric is Euclidean distance. The algorithm start 10 times in different random point and we show the try with less inertia.

$10^{-}6$. Because of the inertia minimization constrain the algorithm can stop in a local minimum. To avoid this the initial K were placed at random 10 times and the try with less inertia was chosen among them.

5.4 Validation

Silhouette coefficient were calculated to all the numbers of clusters in order to get the best. Inertia was also used as a validation metric as measure of how internally coherent clusters are. Some drawbacks of inertia are that it assumes that clusters are convex and isotropic, respond bad to elongated or manifolds cluster shapes and is not normalized metric. Inertia should decrease with cluster number increment, less samples per clusters makes samples more coherent to the barycenter or in other words the average represent better a cluster with less members. Silhouette Coefficient should decrease with cluster number(K) increment because genetic expression data intrinsically overlaps together so clusters will tend to do the same. Using both metrics the local maximum in 10 clusters result interesting to analyze. Meanwhile the inertia keeps decreasing, the Silhouette coefficient increases from 9 to 10 clusters. In these K the clusters are more coherent than all previous K but also they are less overlapped.

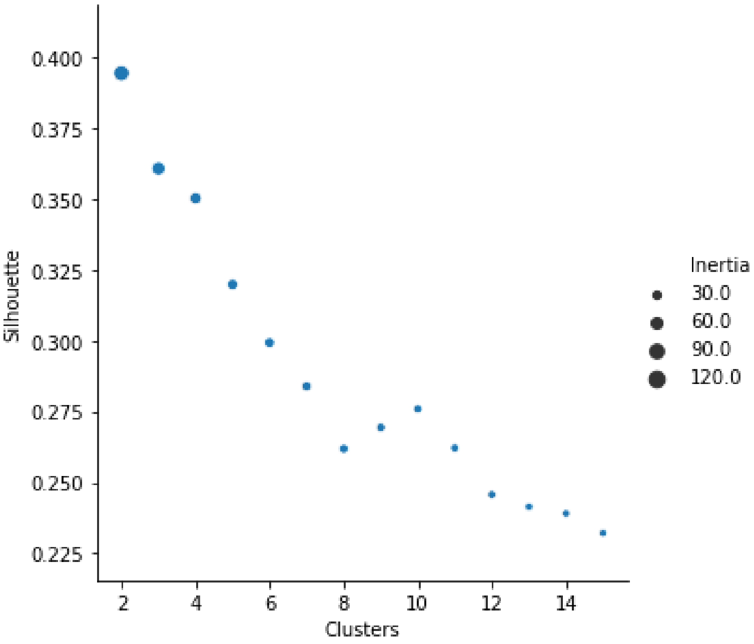

Figure 4: Silhouette Coefficient. K-Means algorithm were used with K from 2 to 15, the Silhouette Coefficient were calculated and appended. X is Silhouette Coefficient, Y is the number of clusters and the size of the dot is the minimal Inertia among the 10 iterations.

6 Conclusion

Clustering is a unsupervised machine learning set of methods that has been optimized for cartesian or plane data and there is a lot of work to do on time series data. Time series data is still an intense field of study and can be used in voice recognition, market trends, dynamic systems and gene expression for example. The K-Means assumption that all the cluster have the same variance do no fit well to biological data. Silhouette Coefficient alone also is not orientated to biological time series of gene expression because score cluster separation and the dynamics tend to be be similar and shared at some times. Even all these drawbacks K-Means used with Silhouette Coefficient and Inertia is a good exploratory method. In future more clustering algorithms developed to aim genetic expression and biological data will aid this field, even the advancing on voice recognition algorithms can apply to these problems if we found the correct way to implement them.

7 References

French, S., Coutts, B. E., Brown, E. D. (2018). Open-Source High-Throughput Phenomics of Bacterial Promoter-Reporter Strains. Cell Systems. https://doi.org/10.1016/j.cels.2018.07. Goldberger, J., Gordon, S., Greenspan, H. (2006). Unsupervised image-set clustering using an information theoretic framework. IEEE Transactions on Image Processing. https://doi.org/10.1109/TIP.2005.860593 Hartigan, A., Wong, M. A. (1979). A K-Means Clustering Algorithm. Journal of the Royal Statistical Society. https://doi.org/10.2307/2346830 Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern Recognition Letters. https://doi.org/10.1016/j.patrec.2009.09.011 Kubat, M. (2017). An Introduction to Machine Learning. An Introduction to Machine Learning. https://doi.org/10.1007/978-3-319-63913-0 Kumar, L., Futschik, M. E. (2007). Mfuzz: A software package for soft clustering of microarray data. Bioinformation. https://doi.org/10.6026/97320630002005 McDowell, I. C., Manandhar, D., Vockley, C. M., Schmid, A. K., Reddy, T. E., Engelhardt, B. E. (2018). Clustering gene expression time series data using an infinite Gaussian process mixture model. PLoS Computational Biology. https://doi.org/10.1371/journal.pcbi.1005896 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research. Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics. https://doi.org/10.1016/0377-0427(87)90125-7 Shalizi, C. (2009). Distances between Clustering , Hierarchical Clustering. Data Mining. Tavenard, Romain and Faouzi, Johann and Vandewiele, G. (2017). tslearn: A machine learning toolkit dedicated to time-series data. Retrieved from https://github.com/rtavenar/tslearn