Encoding Time Series Data for Better Clustering Results

Tomáš Bartoň and Pavel Kordík

Faculty of Information Technology
Czech Technical University in Prague

Abstract. Clustering algorithms belong to a category of unsupervised learning methods which aim to discover underlying structure in a dataset without given labels. We carry out research of methods for an analysis of a biological time series signals, putting stress on global patterns found in samples. When clustering raw time series data, high dimensionality of input vectors, correlation of inputs, shift or scaling sensitivity often deteriorates the result. In this paper, we propose to represent time series signals by various parametric models. A significant parameters are determined by means of heuristic methods and selected parameters are used for clustering. We applied this method to the data of cell's impedance profiles. Clustering results are more stable, accurate and computationally less expensive than processing raw time series data.

1 Introduction

Clustering techniques partition objects into groups of **clusters** so that objects within a cluster are *similar* to one another and *dissimilar* to objects in other clusters. Similarity is commonly defined in terms of how "close" the objects are in space, based on a *distance function* [1].

In real-world application clustering can be misled by the assumption of uniqueness of the solution. In the field of unsupervised learning a single best solution might not exists. There could be a set of equally good solutions which show a dataset from different perspectives. Because of noise, intrinsic ambiguity in data and optimization models attempting to maximize a fitness function, clustering might produce misleading results. Most clustering algorithms search for one optimal solution based on a pre-specified clustering criterion. Usually the quality of the solution can be adjusted by setting up algorithm-specific parameters [2,3].

In the end, all the user cares about is the "usefulness" of the clustering for achieving his final goal [4]. This usefulness is vaguely defined, but can be easily evaluated by an expert who has background knowledge of the dataset and is able to verify the correctness of a result. Sometimes constructive criticism is too difficult, but negative examples (marking incorrect cluster assignments) are easy to find.

2 Problem Specification

A given dataset contains time series data that reflect the biological activity of cells placed on a plastic plate with 96 separated wells. On the bottom of each well an electrode was placed, which measured impedance while a small electric current was going

Á. Herrero et al. (Eds.): Int. JointConf. CISIS'12-ICEUTE'12-SOCO'12, AISC 189, pp. 467–475.
springerlink.com

through. These datasets were measured at IMG CAS[1]. The goal is to find a grouping of samples in a way which would put similar response patterns into the same groups. The similarity is defined rather by the visual similarity of the curves than the absolute values of the curves. The measured quantity is called *Cell Index* and it is a proportional variable displaying a change of impedance compared to the initial state. The concentration of samples has a huge influence on the value of Cell Index. However the shape of curve stays the same. The signals are not periodical and all of them should converge to zero in the end. A sample input is shown at Figure 1.

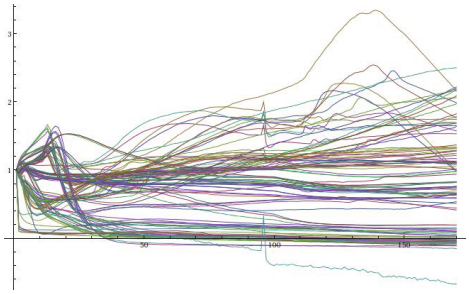


Fig. 1. This chart represents an example of an input dataset with 112 samples – on the y axis Cell Index is shown, which captures number of surviving cells in specific well. The x axis displays time in hours, at the beginning a reactive compound was added. This caused a significant decrease (cells died) in some wells, other samples survived.

The time series signal starts at the moment when a reactive compound is added. The most interesting samples are those, which are resistant to the added virus or capable of recovery after a couple of hours. The samples that are decreasing quickly could be used as a reference group but they are not really perspective for further research.

To sum up, what really matters are the global trends and possibility to discover new, unknown patterns. Which is also the reason why we used an unsupervised learning method.

3 Clustering Time Series

Time series include a huge variety of data that is processed in the areas of medicine [5], biology [6], speech recognition [7], financial market analysis and many other fields. We can distinguish several categories of time series data by their character: **horizontal** (data fluctuate around a constant value), **trend** (contains a visible global trend of increase or decrease), **seasonal** (the trend in data repeats periodically) and **cyclical** (rises and falls repeat without a fixed period).

[1] Institute of Molecular Genetics, Academy of Sciences Czech Republic
`http://www.img.cas.cz/`

Some time series datasets could include a combination of these patterns. It is quite clear that some algorithms perform better on a specific type of data. And for each category a specific encoding of information could be found.

The given time series data are not cyclical, the data might fall in a category of trend time series. However the trend of global decrease is not interesting as long as it is common for all patterns found in a dataset. Some samples have the characteristic of rapid increase and after a culmination point it is either slowly or rapidly decreasing, while others contain a wave shape which could be considered as just one period of some goniometric function. The final remarkable category of patterns are signals with exponential decrease.

There are basically three major approaches to performing clustering of time series data. You can either work directly with raw data, indirectly with features extracted from the raw data, or indirectly with models built from the raw data [8]. In the following sections we would like to compare the first two approaches.

3.1 Raw Time Series Data

Abassi et al. [9] described some biological patterns found in their data. For the clustering of impedance profiles a hierarchical clustering of all data points was used. A bottom-up approach used in agglomerative clustering works well for grouping similar responses if they are aligned in time and all sample have the same length. We applied the very same approach to a test dataset. The resulting clustering is quite good (see Figure 2a), however using all data points as an input for a clustering algorithm is not just computationally expensive but also does not represent patterns well.

3.2 Model Based Encoding

Since the impedance values change over time in a smooth fashion, we wish to fit our data to a curved function. Thus, we assume that the data can by represented by a general model:

$$m_t = f(t_t) + \varepsilon_t, \quad t = 1, 2, \ldots T$$

where the ε are the errors modelled by a Gaussian distribution $N(0, \sigma^2)$ and m_t is a value of a model in time t. While doing curve fitting we try to minimize the root mean square error, defined as:

$$E_{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{T} (y_i - f(t_i))^2}$$

The obtained parameters of the model are further used in the clustering process. A general polynomial function is defined as follows:

$$p_n(t) = a_n t^n + a_{n-1} t^{n-1} + \cdots + a_1 t_1 + a_0 \tag{1}$$

where a_n to a_0 are parameters. To fit the parameters to the data the Levenberg-Marquardt [10] method was used, which is a special type of the Newton method.

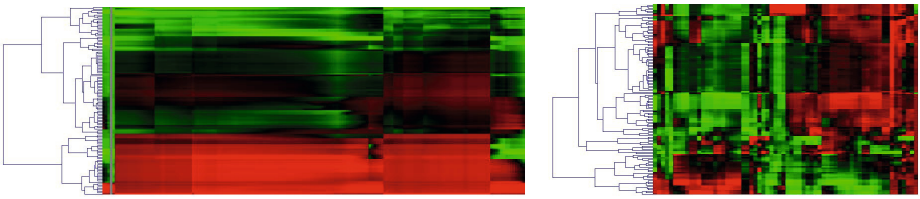
In order to get comparable parameters without missing values models must be simple and universal. Otherwise curve fitting might not converge or the RMSE would be too huge.

As an input for the clustering algorithm we used parameters from n equations. With higher degrees of polynomials the number of inputs grows and is given by formula $(n+1)(n+2)/2-1$. Polynomials of higher degrees fit more precisely to various types of data. Experimentally we found that polynomials with degree between 4 to 6 give a good trade-off between the number of input attributes and precise representation of inputs. By leaving out the last parameter of the polynomial we can easily get rid of the vertical transition of curve.

Another model is base on an exponential function defined as follows:

$$y(t) = a \cdot e^{-bt} + c \tag{2}$$

Many processes in biology have an exponential trend, also in our case some categories could be easily fitted to this model.



(a) raw time series data (b) fitted parameters

Fig. 2. Visual representation of clustering results

4 Experiments

In order to have a precise evaluation metric we have created a dataset with samples from different experiments and afterwards an expert classified the data into 5 categories (see Figure 6). To understand better the way how the experts analyse the data we run a forward selection algorithm on the dataset with labels. Surprisingly not many attributes are needed to decide such a task when you have the advantage of having labels. With only two attributes, one from the beginning and other from the end of the time series data was enough to obtain 98% classification precision (similarly the decision tree only uses 6 time points from 3 different parts of the time series for classification – see Figure 3). This result might suggest splitting the analysis of the signal into multiple parts. The selected time points are almost equally distributed which signifies that all parts of the measurement are important and we can not draw any conclusion from the analysis of just one subsequence.

Any approach based on a similar selection (or aggregation) of specific attributes might suffer from overrating absolute values of samples. This decision tree would fail on a dataset which contained similar but either horizontally or vertically shifted samples

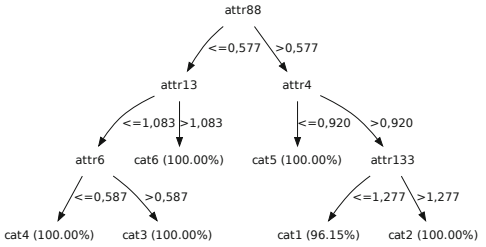


Fig. 3. Decision trees are able to decide the classification of time series with only very few attributes. At least one attribute is taken from the beginning of measurement and other from the end.

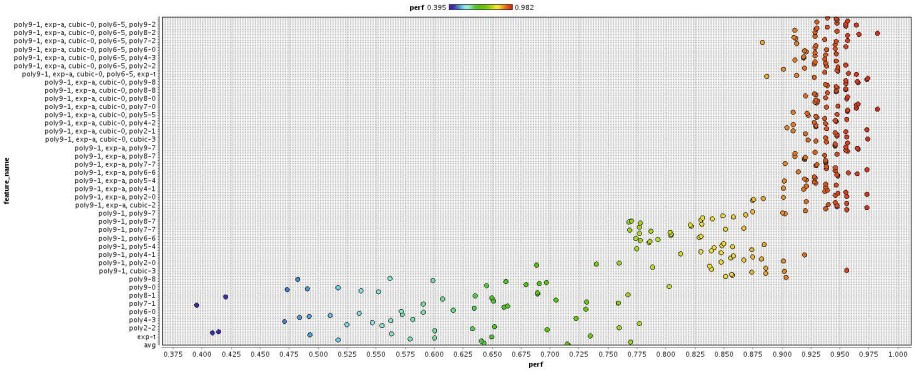


Fig. 4. Classification performance of forward selection on the testing dataset. Algorithm was given a set of 66 attributes to choose from, with 3 attributes we can get the precision above 90%, with 4 and 5 attributes the precision is even better. However using more than 5 attributes would lead to over fitting (worse results).

(where a similar effect appears either sooner or later in time). Therefore we would like to introduce an approach which is resistant to this shift and capable of finding similar patterns.

It is quite clear that including polynomials with increasing degree will introduce some duplicate information. To avoid this issue we used forward selection again with 10 fold cross-validation, to see which attributes are more significant. As you can see from Figure 4 at least three attributes are needed to obtain 90% precision in classification. This proves that the information included in our estimated parameters is more general that information in raw time series data. One of the chosen parameters is the a coefficient of an exponential model (see (2)). This parameter does characterise well the rapid increase or decrease at the beginning of curve, so including this one does make sense. Another parameter signifies horizontal movement of the curve, which seems to be important for the assignment to a category.

Table 1. Comparison of CPCC for agglomerative hierarchical clustering with different settings. CPCC closer to 1 means better clustering. It is clear to see that preprocessing of data and chosen distance metric have a significant influence on results. Best CPCC was achieved with raw data, however in average fitted parameters have better results.

Input	Standardisation	Linkage	Distance metric	CPCC
raw time series	min-max	Complete	Euclidean	0.811
	z-score	Complete	Euclidean	0.652
	maximum	Complete	Euclidean	0.806
	z-score	Average	Canberra	0.944
fitted parameters	min-max	Complete	Euclidean	0.747
	maximum	Average	Euclidean	0.778
	z-score	Average	Canberra	0.894
	z-score	Complete	Canberra	0.889

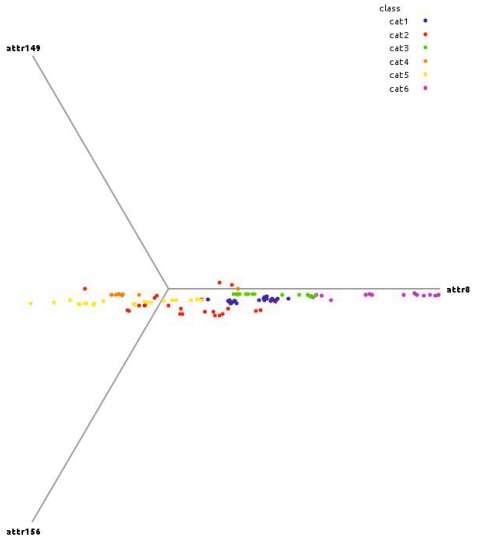
5 Evaluation of Results

Throughout the years many new clustering algorithms have been introduced, however the oldest ones like k-means [11] and agglomerative hierarchical clustering [12] tend to be the most popular methods in literature. When dealing with high dimensional time series data, k-means looks for well-separated clusters with rounded shape in n-dimensional space. That is obviously not the case of time series data and therefore k-means fails to find reasonable clustering. Agglomerative hierarchical clustering use a bottom-up approach while merging closest clusters together. The best results were obtained when average linkage was used, on the other hand the worst results were obtained with the single linkage.

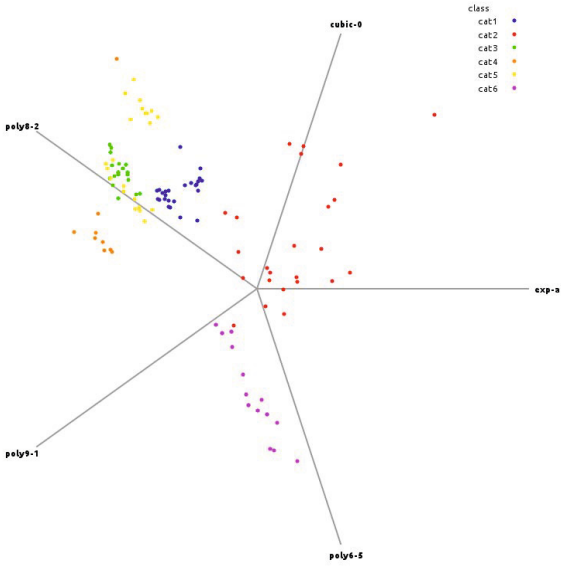
Unsupervised learning is a challenging field mainly due to the lack of a universal evaluation criterion. To deal with this problem we have chosen a semi-supervised approach.

From traditional evaluation metrics we used the Cophenetic correlation coefficient (CPCC) [13] which is one of many evaluation metrics that can be used as optimization criteria. The value of the CPCC should be maximized, however a higher value of the CPCC does not guarantee that user would prefer this result to another clustering. Also should be noted that the number of clusters does not influence the value of CPCC.

By fitting parameters we manage to lower the dimensionality of input data. Clustering is less sensitive to noise and therefore more stable. However, some attributes are more important than others. Clustering produced directly by hierarchical clustering corresponds to counting area below a curve. Therefore patterns with rapid decrease are always together. This might not be the case of the other approach, nevertheless this is just a question of proper weighting and selection of input attributes, which should be done by user.

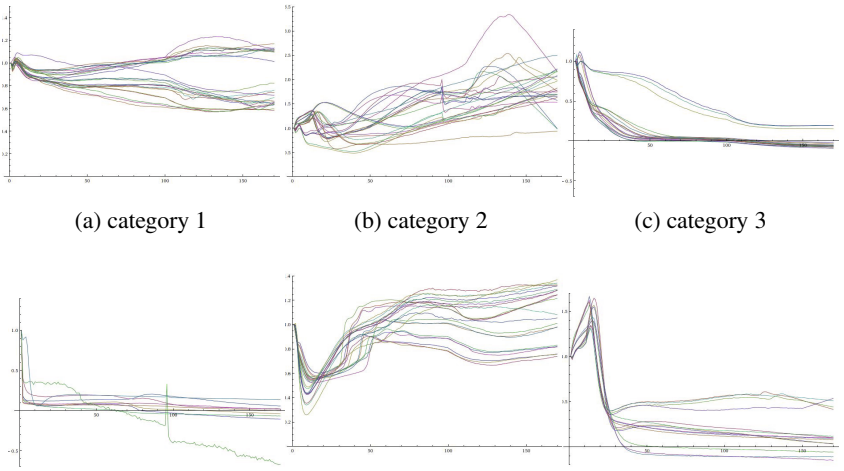


(a) raw time series data



(b) fitted parameters

Fig. 5. Visualization of important attributes for both approaches, which were selected by an evolutionary optimization. It is clear to see that raw time series data does not form separable nor compact clusters. A linear projection was used for both visualizations.



(a) category 1 (b) category 2 (c) category 3

(d) category 4 (e) category 5 (f) category 6

Fig. 6. Manual classification of patterns into 5 categories which was done by an expert

Using similar models is problem specific, and a generalized approach that would work on different types of data is hard to find. Our method manages to capture some patterns found in our datasets, however to fulfil the user's expectations some training in clustering algorithms might be needed.

6 Conclusion

In this contribution show the advantage of representing time series signals by parameters of fitted functions for clustering results. We use a heuristic algorithm to select a subset of parameters for better cluster separation. We would like to focus on the interactive evolution of clustering with multi-objective criterion optimization. The fitted parameters proved to produce good results on the annotated dataset, however, we need to take into account the user's expectation and domain knowledge. This could be done by running an evolutionary algorithm for choosing the best parameters while the user is iteratively annotating the data and visually checking the clustering results on a different set of parameters (individuals).

Acknowledgements. Our research is partially supported by the Novel Model Ensembling Algorithms ($SGS10/\ 307/OHK3/3T/181$) grant of the Czech Technical University in Prague.

We would like to thank Petr Bartůněk, Ph.D. and Antonio Pombinho, M.Sc. from the IMG CAS institute for supporting our research, providing the data and letting us publish all details of our work.

References

1. Han, J., Kamber, M.: Data mining: concepts and techniques. Morgan Kaufmann Publishers Inc., San Francisco (2000)
2. Bifulco, I., Fedullo, C., Napolitano, F., Raiconi, G., Tagliaferri, R.: Global optimization, meta clustering and consensus clustering for class prediction. In: Proceedings of the 2009 International Joint Conference on Neural Networks, IJCNN 2009, pp. 1463–1470. IEEE Press, Piscataway (2009)
3. Caruana, R., Elhawary, M., Nguyen, N., Smith, C.: Meta clustering. In: Proceedings of the Sixth International Conference on Data Mining, ICDM 2006, pp. 107–118. IEEE Computer Society, Washington, DC (2006)
4. Guyon, I., Luxburg, U.V., Williamson, R.C.: Clustering: Science or art. In: NIPS 2009 Workshop on Clustering Theory (2009)
5. Golay, X., Kollias, S., Stoll, G., Meier, D., Valavanis, A., Boesiger, P.: A new correlation-based fuzzy logic clustering algorithm for fmri. Magnetic Resonance in Medicine 40(2), 249–260 (1998)
6. Möller-Levet, C.S., Klawonn, F., Cho, K.-H., Wolkenhauer, O.: Fuzzy Clustering of Short Time-Series and Unevenly Distributed Sampling Points. In: Berthold, M., Lenz, H.-J., Bradley, E., Kruse, R., Borgelt, C. (eds.) IDA 2003. LNCS, vol. 2810, pp. 330–340. Springer, Heidelberg (2003)
7. Wilpon, J.G., Rabiner, L.R.: A modified k-means clustering algorithm for use in speaker-independent isolated word recognition. The Journal of the Acoustical Society of America 75, S93 (1984)
8. Liao, T.W.: Clustering of time series data – a survey. Pattern Recognition 38(11), 1857–1874 (2005)
9. Abassi, Y.A., Xi, B., Zhang, W., Ye, P., Kirstein, S.L., Gaylord, M.R., Feinstein, S.C., Wang, X., Xu, X.: Kinetic cell-based morphological screening: prediction of mechanism of compound action and off-target effects. Chem. Biol. 16(7), 712–723 (2009)
10. Marquardt, D.W.: An algorithm for least-squares estimation of nonlinear parameters. Journal of the Society for Industrial and Applied Mathematics 11(2), 431–441 (1963)
11. MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: Cam, L.M.L., Neyman, J. (eds.) Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297. University of California Press (1967)
12. Lance, G.N., Williams, W.T.: A general theory of classificatory sorting strategies. The Computer Journal 9(4), 373–380 (1967)
13. Sokal, R.R., Rohlf, F.J.: The comparison of dendrograms by objective methods. Taxon 11(2), 33–40 (1962)