Spike train clustering using a Lempel-Ziv distance measure

Stoop Group / Institute of Neuroinformatics

uni | eth | zürich

Winterthurerstrasse 190, CH-8057 Zurich, Tel: (+41) (0)1 635 30 52, http:// www.ini.ethz.ch

Markus Christen (markus@ini.phys.ethz.ch), Thomas Ott and Ruedi Stoop

Multi-electrode array recordings reveal complex structures in the firing of spatia ly distributed neurons. The analysis of this neuronal network activity demands a classification of neurons according to sim iarities in their firing.

If similar sp▮e patterns do not occur synchronously, but have unknown delays within sp▮e trains, this processing step is difficult. To solve this problem, we introduce a Lempel-Ziv comnplexity based distance measure.

1. The Problem

Multi-electrode arrays probe the spatio-temporal activity within a neural net. Up to now, mostly synchronized activity has been investigated. However, due to the complex neuronal connectivity, the same pattern may occur at different times in different neurons. Such neurons may be assumed to be receiving similar input or/and performing similar computations. A method to classify spike trains should thus group together neurons with similar temporal structure, irrespective of the delays between the patterns (Fig. 1). A classification of spike trains requires the choice of a distance measure and a clustering algorithm. For the first step, we introduce a novel Lempel-Ziv based distance measure [1]. For clustering, we use th sequential superparamagnetic clustering algorithm [2].

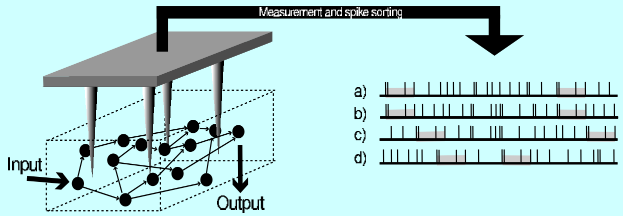

Figure 1: Spike trains obtained by array recordings: Correlation-based measures group together trains with synchronous patterns (a,b), but fail to recruit delayed patterns (c,d).

2. The Lempel-Ziv distance

Algorithm: 1) Translate spike trains into bitstrings $X_n = x_1 \ldots x_n$ using n bins of $\Delta\tau = 1$ ms such that $n\Delta\tau = T$ (T: spike train duration) and set $x_i = 1$ if a spike falls into the i-th bin (otherwise $x_i = 0$).

2) Let $X_n(i,j)$ be a phrase (substring of X_n) starting at position i and ending at position j, P_{X_n} be the set of phrases generated by parsing X_n, and $c(X_n) = |P_{X_n}|$. Lempel-Ziv parsing procedure [3]: Assume that X_n has been parsed up to position i. For the next phrase $X_n(i+1,j)$, increase j until $X_n(i+1,j) \notin P_{X_n(1,i)}$ and add this phrase to $P_{X_n(1,i)}$.

3) Calculate the pairwise LZ-distance between all binary-coded spike trains under investigation t obtain the distance matrix (= input for clustering).

LZ-complexity: $K(X_n)$ of a string X_n:

$$K(X_n) = \frac{c(X_n)\log c(X_n)}{n} \qquad (1)$$

$K(X_n|Y_n)$ of a string X_n given a string Y_n of equal length n is defined accordingly, whereas $c(X_n|Y_n)$ is th siz of the differenc set $P_{X_n} \setminus P_{Y_n}$.

LZ-distance: $d(X_n, Y_n)$ of two strings o length n:

$$d(X_n, Y_n) = 1 - \min\left\{\frac{K(X_n) - K(X_n|Y_n)}{K(X_n)}, \frac{K(Y_n) - K(Y_n|X_n)}{K(Y_n)}\right\} \qquad (1)$$

We have shown, that the definition satisfies the axioms of a metric. The LZ-distance compares the set of phrases generate dy a LZ parsing procedure of two bitstrings originating from corresponding spike trains. A large number of similar patterns appearing in both trains should lead t o large overlap of the sets of phrases. We predict, that distances between trains with similar patterns are small, whereas distances between trains with different patterns are large.

3. Test cases

Figure 2.a: The test file consists of 5 classes (9 train per class) : A) Poisson spike trains with refractory period. B) Poisson spike trains with refractory period driven by a step function o f2.5 Hz. C) Noisy burst-pattern spike trains. D) Spike train sf a complex cell driven by drifting grating sf 6.25 Hz. E) Spike trains of a simple cell driven by drifting gratings of 12.5 Hz. The dendrogram shows, that classes B, C and E have been separated in the first run, but classes A and D are only separated incompletely, indicating that the firing behavior of th complex cell is (in a first approximation) properly modelled by the model.

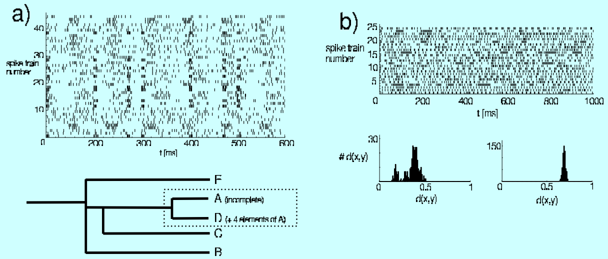

Figure 2: Clustering files with synchronized structures (a, below: dendrogram) and files with delayed spike patterns (b, below: dynamic range of distances). Courtesy of data [4].

Figure 2.b: The test file consists of 5 classes (5 trains per class) , characterized by the ISI-patterns A: (4,4), B: (13,13,13), C: (5,20,3), D: (3,16,3,16), and E: (1,4,7,2,6,11) immersed into a homogeneous Poisson process (50%). After clustering using the LZ-distance and a correlation-based distance, only the LZ-distance allowed a classification of the trains. The dynamic range of the distances (below) explains why: For the LZ-distance (left), the range is ~ 0.4 with a multimodal distribution (indicating the structure within the dataset), whereas for the correlation-based distance, the range is ~ 0.1 with a unimodal distribution (i.e. rescaling is useless for a performance improvement).

Figure 3: Classification of multi-electrode recordings in the olfactory bulb of the rat (64 neurons). a) Neurons are ordered according to the firing rate before stimulus presentation. Dendrograms indicate the result of clustering. Clusters are color-coded, borders are indicated by lines. b) Classification during stimulus presentation. Arrows indicate neurons belonging to different clusters. Striped bars indicate neurons belonging to no cluster. LZ-distance based clustering thus takes account of the changes in the temporal structure of the firing of neurons before and during stimulus-presentation.

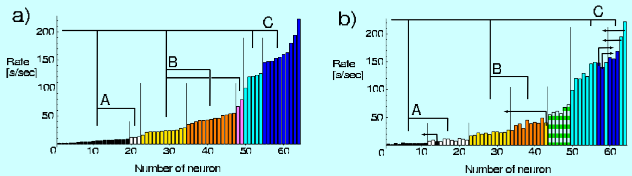

Figure 3: Olfactory bulb data. Spike trains with firing rates > 150 spikes/sec are suspected to result from imperfect spik sorting. Courtesy of the data: [5].

5. Conclusions

Our method is able to group spike trains with similar, but not necessarily synchronous, patterns. It therefore has a broader rang of applications than other distance measures, does not rely on prior information, is easy to implement, and computationally cheap. We thus can: 1) identify neurons with similar firing behaviors in a fast and unbiased way; a refined analysis [6] can then be restricted to one representative of each class. 2) assess the degree of precision to which neuron models reproduce the temporal structure of a biological neuron. 3) measure neuronal firing reliability (and compar ith other neurons).

6. References

[1] A. Lempel, J. Ziv, "On the complexity of finite sequences", *IEEE Trans. Inform. Theory*, IT-22, pp.75–81, 1976. [2] T. Ott, A. Kern, A. Schuffenhauer, M. Popov, P. Acklin, E. Jacoby and R. Stoop, "Sequential Superparamagnetic Clustering for Unbiased Classification of High-dimensional Chemical Data", *J. Chem. Inf. Comput. Sci.*, 44(4), pp.1358–1364, 2004. [3] J. Ziv, A. Lempel, "Compression of individual sequences by variable rate coding", *IEEE Trans. Inform. Theory*, IT-24, pp.530–536, 1978.
[4] We thank Adam Kohn (Center of Neural Science, New York University) for macaque monkey V1 data used in this analysis. [5] We thank Alister Nicol (Cognitive & Behavioural Neuroscience Lab, The Babraham Institute, Cambridge) for rat olfactory bulb data used in this analysis.
[6] M. Christen, A. Kern, A. Nikitchenko, W.-H. Steeb, R. Stoop, "Fast spike pattern detection using the correlation integral", *Phys. Rev. E*, 70, pp.011901-1-7 (2004). [Selected for the July 15, 200 issue of Virtual Journal of Biological Physics Research.]