Soft Pattern Mining in Neuroscience

Christian Borgelt

Abstract While the lower-level mechanisms of neural information processing (in biological neural networks) are fairly well understood, the principles of higher-level processing remain a topic of intense debate in the neuroscience community. With many theories competing to explain how stimuli are encoded in nerve signal (spike) patterns, data analysis tools are desired by which proper tests can be carried out on recorded parallel spike trains. This paper surveys how pattern mining methods, especially soft methods that tackle the core problems of *temporal imprecision* and *selective participation*, can help to test the *temporal coincidence coding hypothesis*. Future challenges consist in extending these methods, in particular to the case of *spatio-temporal coding*.

1 Introduction

Basically all information transmission and processing in humans and animals is carried out by the *nervous system*, which is a network of special cells called *neurons* or *nerve cells*. These cells communicate with each other by electrical and chemical signals. While the lower-level mechanisms are fairly well understood (see Section 2) and it is widely accepted in the neuroscience community that stimuli are encoded and processed by cell assemblies rather than single cells [17, 23], it is still a topic of intense ongoing debate how exactly information is encoded and processed on such a higher level: there are many competing theories, each of which has its domain of validity. Due to modern multi-electrode arrays, which allow to record the electrical signals emitted by hundreds of neurons in parallel [9], more and more data becomes available in the form of (massively) *parallel spike trains* that can help to tackle the challenge of understanding higher-level neural information processing.

European Centre for Soft Computing, Edificio de Investigación, c/ Gonzalo Gutiérrez Quirós s/n, 33600 Mieres, Asturias, Spain, e-mail: christian@borgelt.net

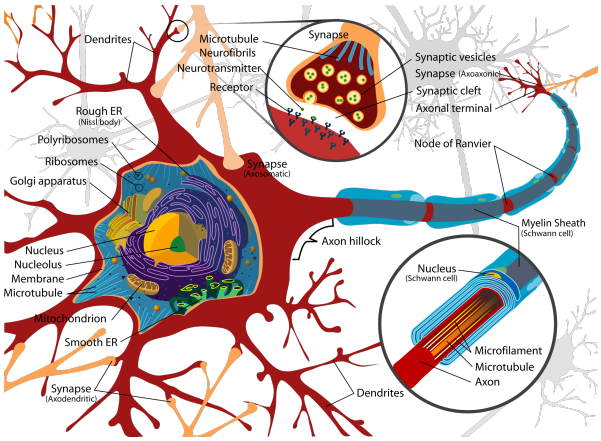

Fig. 1 Diagram of a typical myelinated vertebrate motoneuron (source: Wikipedia, [27]), showing the main parts involved in its signaling activity like the *dendrites*, the *axon*, and the *synapses*.

After reviewing some of the main competing models of neural information coding (Section 2), this paper focuses on the *temporal coincidence coding hypothesis*. It explores how pattern mining methods can help in the search for synchronous spike patterns in parallel spike trains (Section 3) and considers, in particular, soft methods that can handle the core problems of *temporal imprecision* and *selective participation* (Section 4). The paper closes with an outlook on future work, especially tackling the challenge of identifying *spatio-temporal patterns* under such conditions (Section 5).

2 Neural Information Processing

Essentially, neurons are electrically excitable cells that send signals to each other. The mechanisms are well understood on a physiological and chemical level, but how several neurons coordinate their activity is not quite clear yet.

Physiology and Signaling Activity. Neurons are special types of cells that can be found in most animals. They connect to each other, thus forming complex networks. Attached to the *cell body* (or *soma*) are several arborescent branches that are called *dendrites* and one longer cellular extension called the *axon*. The axon terminals form junctions, so-called *synapses*, with the dendrites or the cell bodies of other neurons (see Figure 1) [14].

The most typical form of communication between neurons (this is a *very* simplified description!) is that the axon terminals of a neuron release chemical substances, called *neurotransmitters*, which act on the membrane of the connected neuron and change its polarization (its electrical potential). Synapses that reduce the potential difference between the inside and the outside of the membrane are called *excitatory*, those that increase it, *inhibitory*. Although the change caused by a single synapse is comparatively small, the effects of

multiple synapses accumulate. If the total excitatory input is large enough, the start of the axon becomes, for a short period of time (around 1ms), depolarized (i.e. the potential difference is inverted). This sudden change of the electrical potential, called *action potential*, travels along the axon, with the speed depending on the amount of *myelin* present. When this nerve impulse reaches the end of the axon, it triggers the release of neurotransmitters. Thus the signal is passed on to the next neuron [14]. The electrical signals can be recorded with electrodes, yielding so-called *spike trains*.

Neural Information Coding. It is widely accepted in the neuroscience community that stimuli and other pieces of information are not represented by individual neurons and their action potentials, but that multiple neurons work together, forming so-called *cell assemblies*. However, there are several competing theories about how exactly the information is encoded. The main models that are considered include, but are not limited to the following [23]:

- **Frequency Coding** [29, 12]
  Neurons generate spikes trains with varying frequency as a response to different stimulus intensities: the stronger the stimulus, the higher the spike frequency. Frequency coding is used in the motor system, which directly or indirectly controls muscles, because the rate at which a muscle contracts is correlated with the number of spikes it receives. Frequency coding has also been shown to be present in the sensory system.
- **Temporal Coincidence Coding** [21, 30, 19, 25]
  Tighter coincidence of spikes recorded from different neurons represent higher stimulus intensity, with spike occurrences being modulated by local field oscillation [23]. A temporal coincidence code has the advantage that it leads to shorter "switching times," because it avoids the need to measure a frequency, which requires to observe multiple spikes. Therefore it appears to be a better model for neural processing in the cerebral cortex.
- **Delay Coding** [18, 8]
  The input stimulus is converted into a spike delay (possibly relative to some reference signal). A neuron that is stimulated more strongly reaches the depolarization threshold earlier and thus initiates a spike (action potential) sooner than neurons that are stimulated less strongly.
- **Spatio-Temporal Coding** [1, 2]
  Neurons emit a causal sequence of spikes in response to a stimulus configuration. A stronger stimulus induces spikes earlier and initiates spikes in other, connected cells. The sequence of spike propagation is determined by the spatio-temporal configuration of the stimulus as well as the connectivity of the network [23]. This coding model can be seen as integrating the temporal coincidence and the delay coding principles.

Among other models a spatio-temporal scheme based on a frequency code [28] is noteworthy. In this model the increased spike frequencies form specific spatio-temporal patterns over the involved neurons. Thus it can be seen as combining spatio-temporal coding with frequency coding.

mathematical problem	market basket analysis	spike train analysis
item	product	neuron
item base	set of all products	set of all neurons
— (transaction id)	customer	time bin
transaction	set of products bought by a customer	set of neurons firing in a time bin
frequent item set	set of products frequently bought together	set of neurons frequently firing together

Table 1 Translation of basic notions of frequent item set mining to market basket analysis (for which it was originally developed) and to spike train analysis.

3 Detecting Synchronous Activity

This paper focuses on the temporal coincidence coding hypothesis and thus on the task to detect unusual synchronous spiking activity in recorded parallel spike trains, where "unusual" means that it cannot be explained as a chance event. In addition, we do not merely consider whether a parallel spike train contains synchronous spiking activity (e.g. [31]) or whether a given neuron participates in the synchronous spiking activity of a cell assembly (of otherwise unknown composition) (e.g. [4]). Rather we concentrate on the most complex task of identifying specific assemblies that exhibit(s) (significant) synchronous spiking activity (e.g. [13, 3]). Tackling this task is computationally expensive for (massively) parallel spike trains due to a combinatorial explosion of possible neuron groups that have to be examined.

Other core problems are *temporal imprecision* and *selective participation*. The former means that it cannot be expected that spikes are temporally perfectly aligned, while the latter means that only a subset of the neurons in an assembly may participate in any given synchronous spiking event, with the subset varying between different such events. Note that both may be the effect of deficiencies of the spike recording process (the spike time or even whether a spike occurred is not correctly extracted from the measured profile of the electrical potential) or may be due to the underlying biological process (delays or even failures to produce a spike due to lower total synaptic input, as neurons may receive signals coding different information in parallel).

The most common (or even: the almost exclusively applied) method of handling temporal imprecision is time binning: given a user-specified bin width, a spike train, which is originally a (continuous) point process of spike times, is turned into a binary sequence: a 1 means that the corresponding neuron produced a spike and a 0 that there is no spike in the corresponding time bin. In this way the problem is essentially transformed into a frequent item set mining problem [3]. The translation of the relevant notions to market basket analysis (for which frequent item set mining was originally developed) and to spike train analysis is shown in Table 1. Clearly, the problems are structurally equivalent and thus can be attacked with the same means.

The standard problem of frequent item set mining—namely that a huge number of frequent item sets may be found, most of them *false discoveries*—is best addressed by randomization methods [22, 15]. In spike train analysis, these methods take the form of surrogate data generation schemes, since one tries to preserve as many properties (that are deemed biologically relevant, e.g. inter-spike intervals) as possible, while destroying the coincidences. A survey of such surrogate data generation methods can be found in [20].

In essence, an assembly detection method then works as follows: a sufficient number of surrogate data sets (say, 1000 or 10,000) are created and mined for frequent item sets, which are identified by their size (number of neurons) and support (number of coincidences). Then the original data set is mined and if patterns of a size and support (but ignoring the exact composition by neurons) can be found that do not show up in any of the surrogate data sets, these patterns can be considered significant results.

4 Soft Pattern Mining

Accepting time binning for now as a simple (though deficient, see below) method for handling temporal imprecision, let us turn to the problem of selective participation. In the framework of frequent item set mining this is a well-known problem for which many approaches exist (see, e.g., [5]). The core idea is this: in standard frequent item set mining a transaction (time bin) supports an item set (neuron set) only if all items in the set are present. By relaxing the support definition, allowing for some items of a given set to be missing from a transaction, we arrive at *fault-tolerant item set mining*. The various algorithms for this task can be roughly categorized into (1) error-based approaches, which allow for a maximum number of missing items, (2) density-based approaches, which allow for a maximum fraction of missing items, and (3) cost-based approaches, which reduce the support contribution of a transaction depending on the number of missing items (and may, in addition, restrict the number of missing items) [5].

However, such approaches suffer from the even larger search space (as more item sets need to be examined) and thus can increase the computational costs considerably. An alternative approach that avoids an exhaustive enumeration relies on distance measures for binary vectors [10] and uses multi-dimensional scaling [11] to a single dimension to group neurons together that exhibit similar spiking activity [6]. The actual assemblies are then discovered by traversing the neurons according to their image location and testing for dependence. The approach of computing distances of time-binned spike trains has been extended to various well-known clustering methods in [7].

All of the mentioned methods work on time binned data. However, the time binning approach has several severe drawbacks. In the first place, the induced concept of synchrony is two-valued, that is, spikes are either synchronous

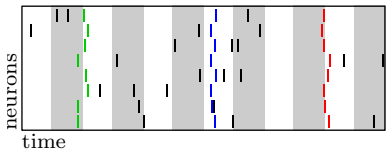

Fig. 2 Eight parallel spike trains with three coincident spiking events (shown in color), two of which are disrupted by time bin boundaries (time bins indicated by gray and white stripes).

(namely if they lie in the same time bin) or not. We have no means to express that the spikes of some coincident event are better aligned than those of another. Secondly, time binning leads to anomalies: two spikes that are (very) close together in time, but happen to be on different sides of a time bin boundary are seen as not synchronous, while two spikes that are almost as far apart as the length of a time bin, but happen to fall into the same time bin, are seen as synchronous. Generally, the location of the time bin boundaries can have a disruptive effect. This is illustrated in Figure 2, where two of the three coincidences of the eight neurons (shown in color) cannot be detected, because they are split by badly placed time bin boundaries.

These problems have been addressed with the influence map approach (see [24, 7]), which bears some resemblance to the definition of a distance measure for continuous spike trains suggested in [26]. The core idea is to surround each spike time with an influence region, which specifies how imprecisely another spike may be placed, which is still to be considered as synchronous. Thus one can define a graded notion of synchrony based on the (relative) overlap of such influence regions. Unfortunately, a direct generalization of binary distance measures to this case (using properly scaled durations instead of time bin counts) seems to lose too much information due to the fact that full synchrony can only be achieved with perfectly aligned spikes [7].

As a solution one may consider specific groups of spikes, one from each neuron, rather than intersecting, over a set of neurons, the union of the influence regions of the spikes of each neuron. This allows to define ϵ-tolerant synchrony, which is 1 as long as the temporal imprecision is less than a user-specified ϵ and becomes graded only beyond that. In addition, extensions to the fault-tolerant case are possible by allowing some spikes to be missing.

5 Future Challenges

The methods reviewed in this paper were devised to detect synchronous activity. However, attention in the neuroscience community shifts increasingly towards spatio-temporal spike patterns as the more general concept, which contains synchronous spiking as a special case. If the time binning approach is accepted, frequent pattern mining offers readily available solutions, for example, in the form of the Spade [33] and cSpade algorithms [32]. However, these approaches require discretized time. Similarly, approaches developed in

the neuroscience community (e.g. [1]) are based on time bins, and thus suffer from the mentioned anomalies. In addition, these methods cannot handle faults, in the sense of individual missing spikes: they only count full occurrences of the potential patterns. It is a challenging, but very fruitful problem to extend these approaches (possibly with influence maps) to continuous time or find alternative methods that can handle both faults and continuous time.

Acknowledgements I am grateful to Denise Berger, Christian Braune, Iván Castro-León, George Gerstein, Sonja Grün, Sebastien Louis, and David Picado-Muiño (listed alphabetically by last name, no precedence expressed), with whom I have cooperated on several of the topics described in this paper.

References

1. Abeles M and Gerstein GL (1988) Detecting Spatiotemporal Firing Patterns among Simultaneously Recorded Single Neurons. *J. Neurophysiology* 60(3):909–924. American Physiological Society, Bethesda, MD, USA
2. Abeles M, Bergman H, Margalit E, and Vaadia E (1993) Spatiotemporal Firing Patterns in the Frontal Cortex of Behaving Monkeys. *J. Neurophysiology* 70(4):1629–38. American Physiological Society, Bethesda, MD, USA
3. Berger D, Borgelt C, Diesmann M, Gerstein G, and Grün S (2009) An Accretion based Data Mining Algorithm for Identification of Sets of Correlated Neurons. *18th Annual Computational Neuroscience Meeting: CNS*2009* 10(Suppl 1):18–23
4. Berger D, Borgelt C, Louis S, Morrison A, Grün S (2010) Efficient Identification of Assembly Neurons within Massively Parallel Spike Trains. *Computational Intelligence and Neuroscience*, Article ID 439648, DOI:10.1155/2010/439648. Hindawi Publishing Corporation, New York, NY, USA
5. Borgelt C, Braune C, Kötter T, and Grün S (2012) New Algorithms for Finding Approximate Frequent Item Sets *Soft Computing* 16(5):903–917. Springer, Berlin, Germany
6. Braune C, Borgelt C, and Grün S (2011) Finding Ensembles of Neurons in Spike Trains by Non-linear Mapping and Statistical Testing. *Advances in Intelligent Data Analysis X* LNCS 7014:55–66. Springer, Berlin / Heidelberg, Germany
7. Braune C (2012) *Analysis of Parallel Spike Trains with Clustering Methods*. Master's Thesis, Otto-von-Guericke-University of Magdeburg, Germany
8. Buzsáki G and Chrobak JJ (1995) Temporal Structure in Spatially Organized Neuronal Ensembles: A Role for Interneuronal Networks. *Current Opinion in Neurobiology* 5(4):504–510. Elsevier, Amsterdam, Netherlands
9. Buzsáki, G.: Large-scale Recording of Neuronal Ensembles. *Nature Neuroscience* 7:446–461. Nature Publishing Group/Macmillian, New York, NY, USA (2004)
10. Choi SS, Cha SH, and Tappert CC (2010) A Survey of Binary Similarity and Distance Measures. *Journal of Systemics, Cybernetics and Informatics* 8(1):43–48. Int. Inst. of Informatics and Systemics, Caracas, Venezuela
11. Cox TF and Cox MAA (2000) *Multidimensional Scaling (2nd ed.)*. Chapman & Hall, London, United Kingdom
12. Eccles JC (1957) *The Physiology of Nerve Cells*. Johns Hopkins University Press, Baltimore, MD, USA
13. Gerstein G, Perkel D, and Subramanian K (1978) Identification of Functionally Related Neural Assemblies. *Brain Research* 140(1):43–62. Elsevier, Amsterdam, Netherlands

14. Gilman S and Newman S (2002) *Essentials of Clinical Neuroanatomy and Neurophysiology (10th ed.).* F.A. Davis Company, Philadelphia, PA, USA
15. Gionis A, Mannila H, Mielikäinen T, and Tsaparas P (2006) Assessing Data Mining Results via Swap Randomization. *Proc. 12th ACM Int. Conf. on Knowledge Discovery and Data Mining (KDD 2006, Philadelphia, PA)*, 167–176. ACM Press, New York, NY, USA
16. Grün A and Rotter S (2010) *Analysis of Parallel Spike Trains.* Springer, Berlin Heidelberg, Germany
17. Hebb DO (1949) *The Organization of Behavior.* J. Wiley & Sons, New York, USA
18. Hopfield JJ (1995) Pattern Recognition Computation using Action Potential Timing for Stimulus Representation. *Nature* 376(6535):33–36. Nature Publishing Group/Macmillian, New York, NY, USA
19. König P, Engel AK, Singer W (1996) Integrator or Coincidence Detector? The Role of the Cortical Neuron Revisited. *Trends in Neuroscience* 19(4):130–137. Cell Press, Maryland Heights, MO, USA
20. Louis S, Borgelt C, and Grün S (2010) Generation and Selection of Surrogate Methods for Correlation Analysis. In [16], 359–382.
21. von der Malsburg C and Bienenstock E (1986) A Neural Network for the Retrieval of Superimposed Connection Patterns. *Europhysics Letters* 3:1243–1249. Institute of Physics Publishing, Bristol, United Kingdom
22. Megiddo N and Srikant R (1998) Discovering Predictive Association Rules. *Proc. 4th Int. Conf. on Knowledge Discovery and Data Mining (KDD 1998; New York, NY)*, 27–78. AAAI Press, Menlo Park, CA, USA
23. Nádasdy Z (1998) *Spatio-temporal Patterns in the Extracellular Recording of Hippocampal Pyramidal Cells: From Single Spikes to Spike Sequences.* PhD Thesis, Rutgers University, NJ, USA
24. Picado-Muiño D, Castro-León I, and Borgelt C (2012) Continuous-time Characterization of Spike Synchrony and Joint Spiking Activity in Parallel Spike Trains. (submitted)
25. Riehle A, Grün S, Diesmann M, and Aertsen A (1997) Spike Synchronization and Rate Modulation Differentially Involved in Motor Cortical Function. *Science* 278:1950–1953. American Association for the Advancement of Science, Washington, DC, USA
26. van Rossum MCW (2001) A Novel Spike Distance. *Neural Computation* 13(4):751–763. MIT Press, Cambridge, MA, USA
27. Ruiz-Villarreal M (2007) Complete Neuron Cell Diagram. `http://commons.wikimedia.org/wiki/File:Complete_neuron_cell_diagram_en.svg`
28. Seidemann E, Meilijson I, Abeles M, Bergman H, and Vaadia H (1996) Simultaneously Recorded Single Units in the Frontal Cortex go Through Sequences of Discrete and Stable States in Monkeys Performing a Delayed Localization Task. *Journal of Neuroscience* 16(2):752–768. Society for Neuroscience, Washington, DC, USA
29. Sherrington CS (1906) *The Integrative Action of the Nervous System.* Yale University Press, New Haven, CT, USA
30. Singer W (1993) Synchronization of Cortical Activity and Its Putative Role in Information Processing and Learning. *Annual Review of Physiology* 55(1):349–374. Annual Reviews, Palo Alto, CA, USA
31. Staude B, Rotter S, and Grün S (2010) CuBIC: Cumulant based Inference of Higher-order Correlations in Massively Parallel Spike Trains. *J. Computational Neuroscience* 29(1–2):327–350. Springer, Berlin, Germany
32. Zaki MJ (2000) Sequences Mining in Categorical Domains: Incorporating Constraints. *Proc. 9th ACM Int. Conf. on Information and Knowledge Management (CIKM 2000, MacLean, VA)*, 422–429. ACM Press, New York, NY, USA
33. Zaki MJ (2001) SPADE: An Efficient Algorithm for Mining Frequent Sequences. *Machine Learning* 42(1/2):31–60. Kluwer, Dordrecht, Netherlands