SPIKE SORTING USING NON PARAMETRIC CLUSTERING VIA CAUCHY SCHWARTZ PDF DIVERGENCE

Sudhir Rao[*], *Justin C. Sanchez*[†], *Seungju Han*[*], *Jose C. Principe*[*]

[*]CNEL, Department of Electrical and Computer Engineering, University of Florida, Gainesville, USA
[†]Department of Pediatrics, Division of Neurology, McKnight Brain Institute, University of Florida, Gainesville, USA

ABSTRACT

We propose a new method of clustering neural spike waveforms for spike sorting. After detecting the spikes using a threshold detector, we use principal component analysis (PCA) to get the first few PCA components of the data. Clustering on these PCA components is achieved by maximizing the Cauchy Schwartz PDF divergence measure which uses the Parzen window method to non parametrically estimate the pdf of the clusters. Comparison with other clustering techniques in spike sorting like k-means and Gaussian mixture elucidates the superiority of our method in terms of classification results and computational complexity.

1. INTRODUCTION

The ability to identify and discriminate the activity of single neurons from neural ensemble recordings is a vital and integral part of basic Neuroscience research. By tracking the modulation of the fundamental constituents of the nervous system, neurophysiologists have begun to formulate the basic constructs of how systems of neurons interact and communicate [1]. Recently, this knowledge of systems neurophysiology has been applied to Brain Machine Interface (BMI) experiments [2] where multielectrode arrays are used to monitor the electrical activity of hundreds of neurons from the motor, premotor, and partietal cortices. In multielectrode BMI experiments, experimenters are faced with the labor intensive task of analyzing each of the extracellular recordings for the signatures of electrical activity related to the neurons surrounding the electrode tip. Separating these different neural sources – a term called "spike sorting", helps the neurophysiologist to study and infer the role played by each individual neuron with respect to the experimental task at hand.

Spike sorting is based upon the property that every neuron has its own characteristic "spike" shape which is dependent on its intrinsic electrochemical dynamics as well as the position of the electrode with respect to the neuron. The key is to separate these spikes from the background noise and use features in each of the shapes to discriminate different neurons. The analysis of this non stationary signal is made more difficult by the fluctuating effects of the electrode-tissue interface which is affected by movement, glial encapsulation, and ionization [3].

To overcome these challenges, many signal processing and machine learning techniques have been successfully applied which are well summarized in [4]. Modern methods use multiple electrodes and sophisticated techniques like Independent Component Analysis (ICA) to address the issue of overlapping and similarity between spikes [4,5]. Nevertheless, in many cases, good single-unit activity can be obtained with a simple hardware threshold detector [6]. After detection, the classification is done using either template matching or clustering of the principal components (PCA) of the waveforms [4,6,7]. PCA of the spike waveforms exploits differences in the variance of the waveshapes to discriminate and cluster neurons.

A common clustering algorithm which is used extensively on the PCA of the waveforms is the ubiquitous k-means[8]. For spike sorting, the k-means algorithm always clusters neurons, but there is no guarantee that it converges to the optimum solution which can produce incorrect sorts. The result depends on the original cluster centers (the random initialization problem) as well as the fact that k-means assumes hyper spherical or hyper ellipsoidal clusters. Lewicki at el. [9] used Bayesian clustering [10] to successfully classify neural signals. The advantage of using Bayesian Framework is that it is possible to quantify the certainty of the classification. Recently, Hillel at el. [11] extended this technique by automating the process of detecting and classification. Other clustering techniques like Gaussian Mixture Model (GMM)[6] and Support Vector Machines (SVM) [12] have also been successfully applied.

The tradeoff of simplicity for accuracy results in techniques suffering from high computational cost and consequently they are not very suitable for online classification of neural signals in low-power devices. Further, model order selection is a difficult task in both GMM and Bayesian clustering. Due to these reasons, k-means is still used extensively for its simplicity and ease of use.

In this paper, we propose the Cauchy Schwartz PDF divergence measure for clustering of neural spike data. We

show that this method not only yields superior results to k-means but also is computationally less expensive with O(N) complexity for classifying online test samples.

2. DATA

2.1. Electrophysiological Recordings

Extracellular cortical neuronal activity was collected using 50µm tungsten microelectrodes from behaving animals. A neural recording system sampling at 24,414.1Hz was used to digitize the extracellular analog voltages with 16 bits of resolution. To emphasize the frequencies contained in action potentials, the raw waveforms were bandpass filtered between 300Hz and 6kHz. Representative microelectrode recordings are shown in Fig. 1. Here, the action potentials from two neurons can be identified (with asterisks) by the differences in amplitude and width of the waveshapes.

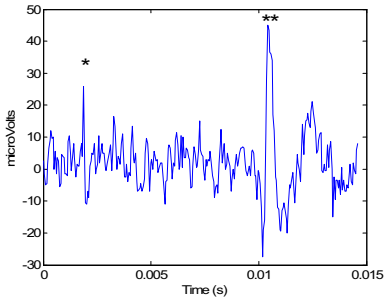

Figure 1. Example of extracellular potentials from two neurons

2.2. Neuronal Spike Detection

The voltage threshold was set by the experimenter using Spike 2 (CED, UK) through visual inspection of the digitized time series. For the example given in Fig. 1, a threshold of 25µV is sufficient to detect each of the two waveshapes. A set of unique waveshapes were constructed from the thresholded waveforms based upon the width which was measured -0.6ms to the left and 1.5ms to the right of the threshold crossing. Using electrophysiological parameters (amplitude and width) of the spike, artifact signals (e.g., electrical noise, movement artifact) were removed. The peak-to-peak amplitude, waveform shape, and interspike interval (ISI) were then evaluated to ensure that the detected spikes had a characteristic and distinct waveform shape when compared with other waveforms in the same channel. Next, the first ten principal components (PC) were computed from all of the detected spike waveforms. Of all the PCs, only the first two eigenvalues were greater than the value one and captured majority of the variance. The first two PCs are plotted in Fig. 2 where two overlapping clusters of points correspond to each of the detected waveshapes. The challenge here is to automatically cluster each of the neurons using automated techniques.

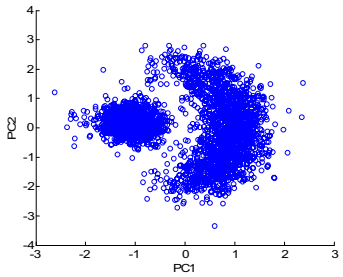

Figure 2. Distribution of PCs from two spike waveforms

3. THEORY

3.1 Renyi's Quadratic Entropy and Information Potential

Given the PCs in Fig. 2, we need a measure of the variability in the distribution of points. Renyi's Quadratic Entropy [13], a continuous and differentiable entropy estimator, for a pdf $p(x)$ is defined as

$$H(X) = -\log\left[\int p^2(x)dx\right] \quad (1)$$

The quadratic entropy can simply be evaluated by plugging in the Parzen window estimator of $p(x)$ using the Gaussian kernel given in eq. (2)

$$p(x) = \frac{1}{N}\sum_{i=1}^{N} G_{\sigma^2}(x - x_i) \quad (2)$$

Using this formulation and the relation that the integral of the product of two Gaussians is another Gaussian with variance equal to the sum of the variances of the two Gaussians, yields a non parametric estimator for Renyi's quadratic entropy

$$H_R(X) = -\log[V(X)]$$

$$V(X) = \sum_{i=1}^{n}\sum_{j=1}^{n} G_{2\sigma^2}(x_i - x_j) \quad (3)$$

where $V(X)$ is called the information potential of the pdf $p(x)$ [14], an analogy borrowed from physics for potential of group of interacting particles. Here, the interacting particles are the PCs from each of the neurons.

3.2. Cauchy – Schwartz PDF divergence measure

Next, the distance between each of the PC distributions needs to be computed using the Cauchy-Schwartz (CS) divergence. Based on the intuition of CS inequality, Principe et al. [14] proposed a divergence measure between two pdfs as expressed in eq. (4).

$$D_{cs}(p,q) = -\log[J_{cs}]$$

$$J_{cs} = \frac{\int p(x)q(x)dx}{\sqrt{\int p^2(x)dx \int q^2(x)dx}} \quad (4)$$

We refer to D_{cs} as the Cauchy Schwartz divergence. It is clear that $J_{cs} \in (0,1]$, such that D_{cs} is always non-negative and symmetric. Maximizing D_{cs} is equivalent to minimizing J_{cs}. Substituting the Parzen window estimator for $p(x)$ and $q(x)$ and using the fact as in eq. (3), we get

$$J_{cs} = \frac{\frac{1}{2}\sum_{i,j=1}^{N,N}(1 - m_i^T m_j)G_{ij,2\sigma^2 I}}{\sqrt{\left(\sum_{i,j=1}^{N,N} m_{i1} m_{j1} G_{ij,2\sigma^2 I}\right)\left(\sum_{i,j=1}^{N,N} m_{i2} m_{j2} G_{ij,2\sigma^2 I}\right)}} \quad (5)$$

where m_i are the fuzzy membership vectors.

Careful observation of (5) shows that J_{cs} is a ratio of between cluster information potential and the within cluster information potentials, and each one can be estimated by (3). Since entropy and information potential are inversely related as shown in eq. (1), minimizing J_{cs} maximizes between cluster entropy and at the same time minimizes the within cluster entropies of the two clusters.

To solve this constrained optimization problem we use Lagrange multipliers as shown below.

$$\min_{m_1, m_2, \ldots, m_N} J_{cs}(m_1, m_2, \ldots, m_N)$$

$$\text{subject to} \quad m_j^T 1 - 1 = 0, \qquad j = 1,2,\cdots,N \quad (6)$$

This technique implements a constrained gradient descent search, with built in variable step-size for each coordinate direction. An elaborate derivation of membership update rule and generalization to more than two clusters can be found in [15]. Jenssen at el. [15] shows superior performance of this algorithm compared to GMM and Fuzzy k-means (FKM) for many non convex clustering problems.

The kernel size is calculated using Silverman's rule of thumb and is given by eq. (7).

$$\sigma_{opt} = \sigma_X \left\{4N^{-1}(2d+1)^{-1}\right\}^{\frac{1}{d+4}} \quad (7)$$

where $\sigma_X^2 = d^{-1}\sum_i \Sigma_{X_{ii}}$ and $\Sigma_{X_{ii}}$ are the diagonal elements of the sample covariance matrix. Further, to avoid local minima we anneal the kernel size from $2\sigma_{opt}$ to $0.5\sigma_{opt}$ over a period of 100 iterations.

Calculating the gradients is $O(N^2)$ complexity. To reduce complexity, we stochastically sample the membership space by randomly selecting M membership vectors where M<<N. Thus the complexity of the algorithm drops to O(NM) per iteration.

3.3. Online Classification of Test Samples

For online spike sorting, the goal is to assign the test point to the cluster whose within cluster entropy has changed the least. This rule automatically ensures the maximization of between cluster entropy and hence the maximization of CS PDF divergence measure for every new test point. The kernel size σ_{opt} can be estimated from the training samples which have already been classified. The kernel size σ_k of each individual cluster was calculated from eq.(7) using the corresponding training samples and then averaged to obtain an estimate of σ_{opt}.

Since information potential and entropy are inversely related as shown in eq. (3), we assign the test point to the cluster which has maximum change in information potential. Assume that the two clusters have N_1 and N_2 points respectively; the change in information potential of the clusters due to a new test point x_t is given by

$$\Delta V_{c_k} = \sum_{j=1}^{N_k} \exp\left\{-\frac{(x_t - x_j)^2}{2\sigma_{opt}^2}\right\} \quad (9)$$

Thus the classification rule can be summed up as

$$\begin{aligned} &\text{If } \Delta V_{c_1} > \Delta V_{c_2} \quad ----> \text{ Classify as Cluster 1} \\ &\text{If } \Delta V_{c_1} < \Delta V_{c_2} \quad ----> \text{ Classify as Cluster 2} \end{aligned} \quad (10)$$

Note specifically this computation takes just O(N) calculations instead of the calculation of CS pdf distance.

4. RESULTS

4.1. Clustering of PCA components

The dataset presented in 2.2 consisted of 2734 points out of which the first 500 points were used in training phase. The algorithm was fully automatic and involved kernel annealing. Only one fourth of the data (M= 0.25N) was used for calculating the membership update thus speeding up the algorithm significantly. Our algorithm took 6-9 seconds (Dell P4, 1.8GHz, 512MB RAM) to converge giving good clustering result for almost every new initialization as shown in Fig. 3. As seen in Fig. 4, k-means clearly fails to separate the two clusters giving poor classification results.

4.2. Online labeling of neural spikes

Using the classification rule presented in section 3.3, the remaining 2234 points were classified as belonging to either of the clusters by just comparing with the training samples as shown in Fig. 5. For comparison we have also plotted the training samples along with the test samples. The algorithm took 2-3 seconds to classify all 2234 test points.

We also used another method of comparison where a new test point is not only compared to the training samples but also to the previous test points which have been already classified. This will linearly increase the computational complexity but yields better classification results. Nevertheless, in our case the two methods gave identical results which may be due to the fact that the training samples defined the clusters sufficiently well.

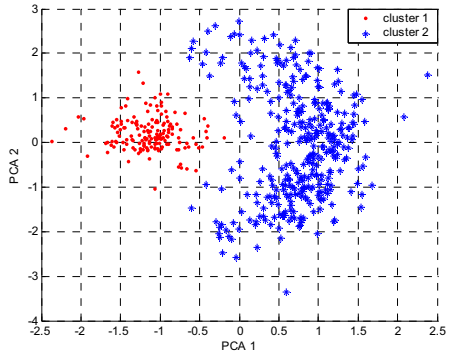

Figure 3. Clustering of training data using CS distance

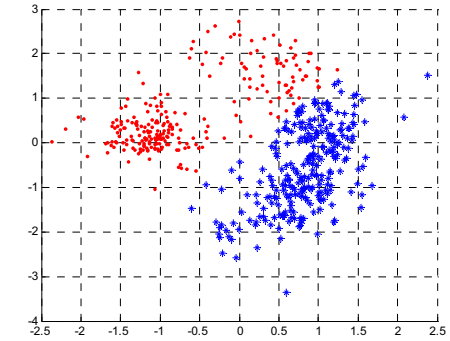

Figure 4. Clustering of training data using k-means

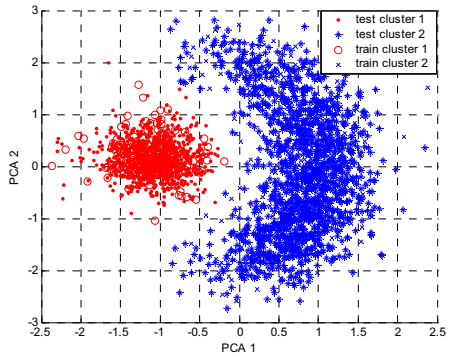

Figure 5. Online Classification of test points

5. CONCLUSIONS

With the advent of multielectrode data acquisition techniques, fast and efficient sorting of neural spike data is of utmost importance for monitoring the activity of ensembles of single neurons. The state-of-the-art in neuronal waveform PCA analysis is faced with a clustering problem due to the electrochemical dynamics in the tissue. We have proposed a technique for clustering that addresses waveform PCA distributions that are not ellipsoid, non Gaussian, and non convex that result from neural sources both near and far from the electrode. Clustering based on Cauchy-Schwartz PDF divergence helps address these issues encountered in multielectrode BMI experiments and classifies new incoming neural spike with O(N) complexity which is suitable for implementation in low-power portable hardware. With no parameters to be selected, this method additionally provides neurophysiologists with an easy and powerful tool for spike sorting. Future research involves extending this technique simultaneously to hundreds of electrodes and extensive comparison with other methods like ICA and Bayesian clustering.

Acknowledgements: This work was partially supported by NSF grants ECS-0300340 and EIA-0135946. The work of Justin C. Sanchez was partially supported from the McKnight Brain Spinal Cord Injury Trust Fund.

11. REFERENCES

[1] Fetz E.E, "Are movement parameters recognizably coded in the activity of single neurons," Behavioral and Brain Sciences, vol. 15, no. 4, pp 679-690, 1992.

[2] J. Wessberg, C.R. Stambaugh, "Real-time prediction of hand trajectory by ensembles of cortical neurons in primates," Nature, vol. 408, no. 6810, pp 361-365, 2000.

[3] J.C. Sanchez, N. Alba, "Structural modifications in chronic microwire electrodes for cortical neuroprosthetics: a case study," submitted to IEEE Trans. on Neural Systems and Rehabilitation Engineering, 2005.

[4] M.S. Lewicki, "A review of methods for spike sorting: the detection and classification of neural action potentials," Network: Computation in Neural Systems vol.9, no. 4, R53-R78, 1998.

[5] E.M. Brown, R.E. Kass and P.P. Mitra, "Multiple neural spike train data analysis: state-of-the-art and future challenges," Nature Neuroscience, vol. 7, no. 5, pp 456-461, May 2004.

[6] B. C. Wheeler, "Automatic discrimination of single units," *Methods for neural ensemble recordings,* M. Nicholelis, Ed.. Boca Raton, Florida: CRC Press, 1999, ch. 4, pp 61-78.

[7] F. Wood, M. Fellows, J.P. Donoghue, M.J. Black, " Automatic Spike Sorting for Neural Decoding," Proceedings of IEEE EMBS, pp 4009-4012, Sept. 2004.

[8] Duda R.O, Hart P.E, Stork D.G., *Pattern Classification*, Wiley Interscience, 2nd Ed., Oct 2000.

[9] M.S. Lewicki, "Bayesian modeling and classification of neural signals," Neural Computation, vol.6, pp 1005-1030, 1994.

[10] Bishop C.M., *Neural Networks for Pattern Recognition*, Oxford, 1995.

[11] A.B. Hillel, A. Spiro and E. Stark, "Spike Sorting: Bayesian Clustering of Non-Stationary Data," NIPS, Dec 2004.

[12] R.J. Vogelstein, K. Murari, P.H. Thakur, G. Cauwenberghs, S. Chakrabartty, C.Diehl, "Spike Sorting with Support Vector Machines," Proceedings of IEEE EMBS, pp 546-549, Sept. 2004.

[13] A. Renyi, "Some Fundamental questions of Information Theory," Selected papers of Alfred Renyi, vol 2, pp 526-552, Akademia Kiado, Budapest, 1976.

[14] J. Principe, D. Xu and J. Fisher, "Information Theoretic Learning," in *Unsupervised Adaptive Filtering*, vol 1, S. Haykin (Ed.), John Wiley and Sons, New York, 2000, ch. 7.

[15] R. Jenssen, "An Information Theoretic Approach to Machine Learning," Ph.D Dissertation, University of Tromso, May 2005.