Unsupervised clustering of evoked potentials by waveform

A. B. Geva H. Pratt

Evoked Potentials Laboratory, Departments of Biomedical & Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel

Abstract—A procedure for clustering evoked potentials (EPs) according to their waveforms is presented. Clustering is performed without a priori selection of basis waveforms, the number of basis waveforms or the number of clusters. The method uses the principal-component-analysis coefficients of EP records as features for unsupervised optimal fuzzy clustering (UOFC) of the records. The validity of the procedure is demonstrated in two instances: visual evoked potentials (VEPs) and cognitive event-related potentials (ERPs) from humans in a memory-scanning task. In the clustering of VEPs, the procedure differentiates between waveforms judged to be clinically normal and abnormal. In the clustering of ERPs, the procedure correctly differentiates between waveforms evoked by the same stimuli which differ in their context to the performance of a memory-scanning task (memorised items against probes). Within this classification, the procedure detects two subgroups to probe-evoked waveforms, which are not obvious from visual inspection of the waveforms. The advantage of the procedure, which conducts clustering by UOFC, is the adaptive and machine-learning nature of its operation.

Keywords—Event-related potentials, Fuzzy clustering, Pattern recognition, Principal component analysis, Visual evoked potentials, Waveform reconstruction

1 Introduction

Evoked potentials (EPs) are typically quantified by latency and amplitude measures of peaks and troughs along the curve, all other features of the record being ignored. Features that are thus ignored include latencies and amplitudes of interposed data points, which comprise the waveform of the recording. Consequently, much of the information in the record is lost. An alternative procedure that reduces EPs to measures that do reflect the entire waveform is principal-component analysis (PCA) (McGILLEM and AUNON, 1987; GONZALES and WINTZ, 1987). In PCA, the basis waveforms from which all members of a set of records can be derived are calculated. These basis waveforms are the eigenvectors of the record set. Each record in the set can then be characterised by the coefficients of the basis waveforms from which it is reconstructed.

When records in a set have common features in their waveforms, each record can be reduced to a small number of coefficients. These coefficients can be used to divide the records into groups. Clustering of such groups is particularly suited to uncovering the structure of a given data set by overlooking dispensable details and removing noise. However, in EP studies, the waveforms are affected by a variety of experimental and subject factors, increasing the variability of coefficients. Therefore, when the coefficients of records within a set of EPs are compared, considerable overlap occurs, and no clear clustering of records is evident. Consequently, these coefficients have only been used for quantifying specific changes in waveforms within a given experimental setting (DONCHIN et al., 1975; WASTELL, 1979; DONCHIN and HEFFLEY, 1979; RUCHKIN et al., 1990).

We propose using fuzzy clustering (ZADEH, 1965; BEZDEK and CASTELAZ, 1971; BEZDEK, 1981; GATH and GEA, 1989a; b) to overcome the problem of EP data variability. In a fuzzy environment, an item can not only belong to one group, but can share common features with a number of groups. Such an item is said to have membership in a number of groups, and not in one group exclusively. Clustering of items in a fuzzy environment is called fuzzy clustering. Fuzzy clustering does not force an item to be a member of only one group, but accommodates membership of each item in a number of groups.

We present a procedure for clustering evoked potentials according to their waveforms, without a priori selection of basis waveforms, the number of basis waveforms or the number of clusters. The method uses the PCA coefficients of EP records as features for fuzzy clustering of the records. In this report, the method is outlined and its validity demonstrated in two instances from humans; visual evoked potentials (VEPs) and cognitive event-related potentials (ERPs) in a memory-scanning task.
2 Methods

The input to the procedure for clustering evoked potentials according to their waveforms is a set of a digitally sampled EP records. Examples include evoked potentials to the same stimuli, recorded from the same electrodes on different subjects, or evoked potentials to different stimulus parameters recorded from the same electrodes on the same subjects, or evoked potentials from the same subject in response to the same stimuli from different electrodes, or any combination of the above. The requirement for meaningful results is that the set includes a large enough number of records that share some features, and the effects on waveform of changes in other features are examined.

The procedure includes two major steps: principal-component analysis (PCA) of the records in the set; and unsupervised optimal fuzzy clustering (UOFC). Two types of preprocessing can precede PCA to negate the effects of irrelevant waveform distortions; first, filtering can be performed to reduce distorting frequencies and enhance the features of interest in the specific waveforms. Filter parameters are based on spectral analysis of the records, sparing the main lobes of the spectrum. Second, Woody filtering (Woody, 1967) can be performed to correct for latency differences between records. Thus, variations among records due to true waveform differences may be distinguished from those resulting from latency differences.

2.1 Principal-component analysis

In PCA, a set of basis waveforms (i.e., principal components) common to all the records are computed and arranged in decreasing order of their contribution to reconstruction of all records in the set. In more formal terms, the basis waveforms are the eigenvectors of the record set covariance matrix, which represents the correlation between all records. The eigenvectors are then arranged in decreasing order of their eigenvalues. These basis waveforms (eigenvectors) constitute an orthogonal basis of the set of records. Therefore each record can be exclusively reconstructed by a linear combination of the basis waveforms and each can be multiplied by an appropriate weight, or reconstruction coefficient (not to be confused with the eigenvalue, see eqn. 6). These reconstruction coefficients are used by our procedure as the features vectors for the UOFC.

2.2 Formulation of the PCA

Our PCA is an implementation of the discrete Karhunen–Loeve transform (McGillem and Aunon, 1987; Gonzales and Wintz, 1987), which is also referred to as hotelling transformation, varimax rotation, singular value decomposition (SVD) or eigenvector PCA. Each sampled EP record can be expressed in the form of row vector \( V_i(n) \), \( n = 1, 2, \ldots, N; i = 1, 2, \ldots, M \), where \( N \) is the number of samples in each record, and \( M \) is the number of EP records:

\[
V_i = \begin{bmatrix}
V_i(1) \\
V_i(2) \\
\vdots \\
V_i(N)
\end{bmatrix}
\]

All EP records can be expressed in the form of \( N \times M \) EP matrix \( V \), as follows:

\[
V = \begin{bmatrix}
V_1(1) & V_2(1) & \cdots & V_M(1) \\
V_1(2) & V_2(2) & \cdots & V_M(2) \\
\vdots & \vdots & \ddots & \vdots \\
V_1(N) & V_2(N) & \cdots & V_M(N)
\end{bmatrix}
\]

The covariance \((N \times N)\) matrix of the EP matrix \( V \) is defined as

\[
C_V = E[(V - M_p) \cdot (V - M_p)^T]
\]

where

\[
M_p(n) = E[V(n)] = \frac{1}{M} \sum_{i=1}^{M} V_i(n) \quad n = 1, 2, \ldots, N
\]

is the grand mean EP vector, and \( E \) is the expected value. (Note that the subtraction of the grand mean \( M_p \) from each record augments differences among records and is thus best suited for their subsequent clustering.)

Let \( G(n) \) and \( x_j, j = 1, 2, \ldots, L; n = 1, 2, \ldots, N \) be the eigenvectors and the corresponding eigenvalues of \( C_V \), where \( L \) is the number of non-zero eigenvalues. The eigenvalues have been arranged in decreasing order so that

\[
x_1 \geq x_2 \geq \cdots \geq x_L.
\]

An \( L \times M \) coefficients matrix is then computed as follows:

\[
Y = G \cdot (V - M_p)
\]

As it can be shown that, in this case, the eigenvectors are orthonormal (Noble, 1969), \( G \cdot G^T = I \) or \( \text{Inverse}(G) = G^T \), \( V \) can be reconstructed by

\[
V = G^T \cdot Y + M_p
\]

If we form an \( F \times N \) matrix \( G_F \) from the first \( F \) eigenvectors that correspond to the largest eigenvalue, and the respective \( F \times M \) matrix \( Y_F \) from the first \( F \) rows of \( Y \), we can approximate \( V \) by

\[
\hat{V} = G_F^T \cdot Y_F + M_p
\]

It can be shown that the mean square error \( R \) between \( V \) and \( \hat{V} \) is given by

\[
R = \frac{1}{N} \sum_{j=F+1}^{N} x_j
\]

The \( M \) columns of the \( Y_F \) coefficients matrix are then used in our procedure as the features vectors for the UOFC.

2.3 Unsupervised optimal fuzzy clustering (UOFC)

Clustering of waveforms was conducted by UOFC (Gath and Geva, 1989a,b). In addition to using fuzzy-clustering algorithms (not forcing an item to be a member of only one group and accommodating membership of each item in a number of groups), UOFC performs clustering of data without \textit{a priori} assumptions about the characteristic features of the clusters, or assumptions about the number of clusters in the data set. The cluster centres and the number of clusters are learned by the system in an unsupervised manner. Clustering begins with the assigning of all records to a single cluster and the calculation of memberships in this cluster. Next, the procedure creates a second cluster to include the records with the lowest memberships in the first cluster. This sequence of adding clusters is repeated until a validity criterion is met. The UOFC validity criteria are based on two parameters:

(a) the sum of memberships within the \( i \)th cluster, \( M_i \)
(b) the standard deviation of members within the \( i \)th cluster \( S_i \).