UDC 612.821:51+519.7+519.8

W.A. CHAOVALITWONGSE, P.M. PARDALOS

ON THE TIME SERIES SUPPORT VECTOR MACHINE USING DYNAMIC TIME WARPING KERNEL FOR BRAIN ACTIVITY CLASSIFICATION[1]

Keywords: *time series, classification, EEG, brain dynamics, optimization, dynamic time warping, epilepsy, support vector machines.*

1. INTRODUCTION

The human brain is among the most complex systems known to man. In neuroscience research, countless number of studies have attempted to comprehend the mechanism of brain functions through detailed analysis of neuronal excitability and synaptic transmission. Many theories of brain functions have been proposed over the last century. Only in the last few years has it become feasible to capture simultaneous responses from large enough numbers of neurons to empirically test those long-standing hypotheses about brain function. However, most neuro-scientific experiments have resulted in massive datasets, in a form of multi-dimensional time series data. These data contain both spatial and temporal properties of brain functions. Making sense of such massive data requires very efficient and sophisticated techniques that are capable of capturing both spatial and temporal properties simultaneously. Current research studies in data mining and classification are mostly focused on the data with only spatial or temporal properties. In addition, very few studies in quantitative neuroscience are not tailored to exploit both spatial and temporal properties of this relentless flood of information.

In this study, epilepsy will be a case point. Epilepsy is the second most common brain disorder after stroke, yet the most devastating one. The most disabling aspect of epilepsy is the uncertainty of recurrent seizures, which can be characterized by a chronic medical condition produced by temporary changes in the electrical function of the brain. Most epilepsy studies employ electroencephalograms (EEGs) as a tool for capturing the electrical changes and evaluating physiological states (normal and abnormal) of the brain. Although EEGs offer excellent spatial and temporal resolution of brain activity, EEG data are so enormous, in a form of long-term multi-dimensional time series, that neuroscientists understand very little about the dynamical transitions to neurological dysfunctions of seizures. A necessary first step to advance epilepsy research is to develop a seizure prediction/warning system. Therefore, the main goal of this study to employ techniques in data mining and optimization to discover seizure-precursor patterns encrypted in the enormous EEGs. In order to validate the reliability of a seizure prediction/warning system, one has to test the hypothesis that the EEGs during the normal period differ from the EEGs during the seizure-precursor. This will, in turn, lead to a classical classification problem. However, the data used in this classification problem has spatio-temporal properties. We herein propose an efficient and effective spatio-temporal data mining/classification method for multi-dimensional time series classification of brain activity.

The organization of the succeeding sections of this paper is as follows. The background of this work including classification techniques in the literature and previous studies in seizure prediction and classification will be discussed in Sec. 2. Subsequently in Sec. 3, basic concepts and standard classification procedure of support vector machines are discussed. The methods employed in this study including the quantification of the brain dynamics, the EEG data acquisition, the support vector machines with dynamic

[1] Research was partially supported by Rutgers Research Council grant-202018, the NSF grants CCF-0546574, DBI-980821, EIA-9872509, CCF 0546574, and NIH grant R01-NS-39687-01A1.

time warping kernel are given in Sec. 4. The design of experiments and the details of the empirical study are described in Sec. 5. The results on the statistical evaluation and the performance characteristics of the proposed classification method are provided in Sec. 6. The concluding remarks and future works are then later discussed in Sec. 7.

2. BACKGROUND

In this section, we give an overview of classification techniques in the literature. We will subsequently focus on optimization-based classification techniques like support vector machines. Later in this section, we give a brief background about epilepsy and the significance of this work to epilepsy research.

2.1. Classification Techniques. Generally, classification techniques are combinatorial in nature as they are involved with discrete decisions. Thus, classification problems can be naturally posted as discrete optimization problems [3, 6, 15, 18, 19, 22, 34]. There have been enormous number of optimization techniques for classification problems developed during the past few decades including classification tree, support vector machines (SVMs), linear discriminant analysis, logic regression, least squares, nearest neighbors, etc. Most optimization methods in classification have been applied to SVMs. A number of linear programming formulations for SVMs have been used to explore the properties of the structure of the optimization problem and solve large-scale problems [5, 35]. The SVM technique proposed in [35] was also demonstrated to be applicable to the generation of complex space partitions similar to those obtained by C4.5 [39] and CART [7]. Current SVM research mainly focuses on extending SVMs to multi-class problems [20, 31, 45]. Nevertheless, there have been very few studies in the literature that address the use of SVMs for time series classification. Moreover, most of the existing studies are only focused on pattern recognition and similarity search [12, 13, 17, 29, 27, 28] or the use of kernel functions for single time series transformation like speech recognition [47, 43, 50, 4, 52].

Multi-dimensional time series classification (MDTSC) has to deal with massive time series data with both spatial and temporal properties (e.g., EEGs). However, to our knowledge, almost all of SVM studies only address either spatial or temporal classification problem. None of current SVM studies attempts to simultaneously consider both problems. In this study, a novel SVM approach for MDTSC is developed based on the improved kernel of finding a separating plane of SVMs in time series sample as well as the idea of projecting the data from the temporal properties. This technique represents a bridge between the parametric techniques that require a priori knowledge of the distributions underlying the data and nonparametric techniques, which presuppose the functional form of the discriminant surfaces separating the different pattern classes.

2.2. Epilepsy Research. Epilepsy will be a case in point in this proposal. Epilepsy is the second most common brain disorder, currently afflicting at least 2 million Americans. The diagnosis and treatment of epilepsy is complicated by the disabling aspect that seizures occur spontaneously and unpredictably. A major epilepsy research lies in the study of how neuronal circuitries of the brain support these electrical changes. Most epilepsy studies use EEGs as a tool for capturing the electrical changes and evaluating physiological states (normal and abnormal) of the brain. Although EEGs have been widely used for the past few decades, neuro-scientists understand very little about the dynamical transitions to neurological dysfunctions of seizures [21]. In some types of epilepsy (e.g., focal or partial epilepsy), there is a localized structural change in neuronal circuitry within the cerebrum which produces organized quasi-rhythmic discharges, which spread from the region of origin (epileptogenic zone) to activate other areas of the cerebral hemisphere. The development of the epileptic state can be considered as changes in network circuitry of neurons in the brain that produce changes in voltage potential, which can be captured by EEG recordings. These changes are reflected by wriggling lines along the time axis in a typical EEG recording. A typical electrode montage for intracranial EEG recordings in our study is shown in Fig. 1. The 10-second EEG profiles during the normal and pre-seizure periods of patient 1 are illustrated in Figs. 2a and 2b. The EEG onset of a typical epileptic seizure is illustrated in Fig. 2c. Fig. 2d shows the post-seizure state of a typical epileptic seizure, respectively.

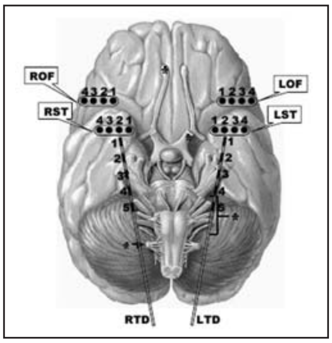


Fig. 1. Inferior transverse views of the brain, illustrating approximate depth and subdural electrode placement for EEG recordings are depicted. Subdural electrode strips are placed over the left orbitofrontal (LOF), right orbitofrontal (ROF), left subtemporal (LST), and right subtemporal (RST) cortex. Depth electrodes are placed in the left temporal depth (LTD) and right temporal depth (RTD) to record hippocampal activity.

2.3. Seizure Prediction and Brain Activity Classification. If seizures could be predicted, it would lead to the development of completely novel diagnostic and therapeutic advances in controlling epileptic seizures. This will tremendously improve the quality of life for those patients who currently suffer from epilepsy. There is growing evidence that human epileptic seizures are preceded by physiological changes that are reflected in the dynamical characteristics of the EEG signals. Our group reported pre-seizure convergence of $STL_{\max}$ values (calculated from intracranial or scalp electrode EEG recordings) which occurred tens of minutes prior to epileptic seizures [24]. Subsequently, Elger and Lehnertz [14, 32] reported reductions in the effective correlation dimension (D_2^{eff}, a measure of the complexity of the EEG signals) that were more prominent in pre-seizure EEG samples than at times more distant from a seizure. They estimated that a detectable change in dynamics could be observed at least 2 minutes before a seizure in most cases [14]. Martinerie and coworkers [36] also reported significant differences between dimension measures obtained in pre-seizure versus normal EEG samples. They found an abrupt decrease in dimension during the pre-seizure transition. This study also employed relatively brief (40 minutes) samples of pre-seizure and normal data. More recently, this group reported changes in brain dynamics obtained from scalp electrode recordings of the EEG. By comparing pre-seizure EEG samples to a reference sample selected from normal data, they found evidence for dynamical changes that anticipated temporal lobe seizures by periods of up to 15 minutes [40]. Recently, Litt and coworkers [33] reported sustained bursts of energy in some EEG channels visually selected by one of the investigators.

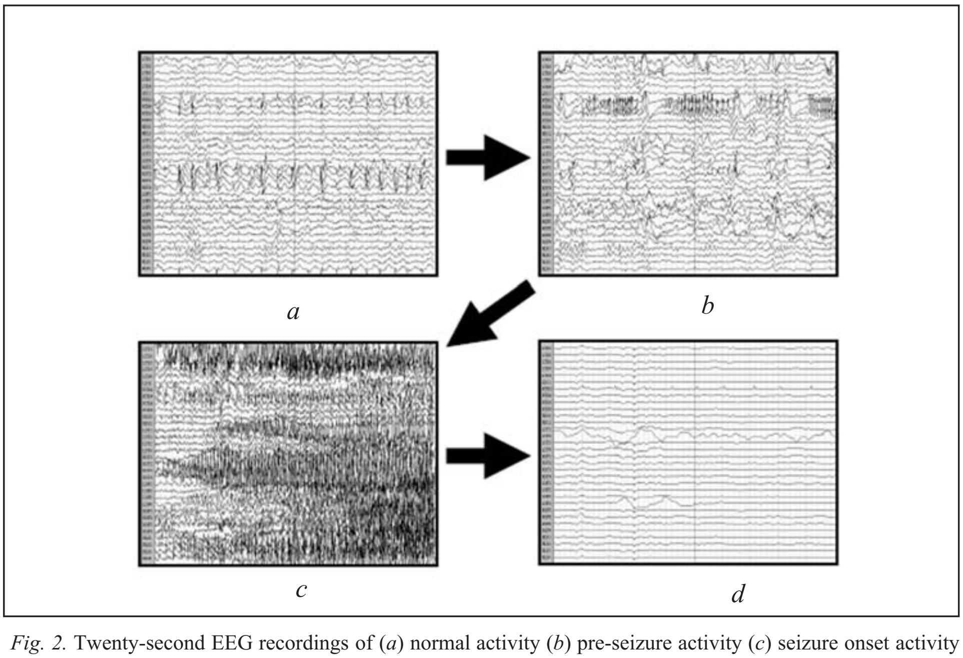


Fig. 2. Twenty-second EEG recordings of (*a*) normal activity (*b*) pre-seizure activity (*c*) seizure onset activity (*d*) post-seizure activity from patient 1 obtained from 32 electrodes. Each horizontal trace represents the voltage recorded from electrode sites listed in the left column (see Fig. 1 for anatomical location of electrodes).

During the past decade, seizure predictability has been demonstrated through the above-mentioned studies including our previous studies in [9,10,25,38]. These studies were motivated by mathematical models used to analyze multidimensional complex systems (e.g., neuronal network in the brain) based on the chaos theory and optimization techniques. The results of those studies demonstrated that a seizure is essentially a reflecting transition of progressive changes of hidden dynamical patterns in EEG. Such transitions have been shown to be detectable through the quantitative analysis of the brain dynamics [9,10,38]. However, in order for one to validate the seizure predictability, one would have to demonstrate, qualitatively and quantitatively, that the normal EEGs differ from the pre-seizure (abnormal) EEGs. The discriminant ability to differentiate and classify a pre-seizure EEG signal is logically a prerequisite and a necessary first step of seizure prediction/warning development. Thus far, to our knowledge, none of current epilepsy studies in the literature is undertaken to test this hypothesis. Our group has attempted to apply data mining techniques using hidden dynamical characteristics to differentiate normal and pre-seizure EEGs [11].

3. SUPPORT VECTOR MACHINES (SVMs)

In this section, we discuss some basic concepts of SVMs. Then, we explain a general classification procedure of SVMs. Later, we address the use of kernel functions, the most widely used trick of SVMs.

3.1. Basic Concepts. SVMs is one of the classification techniques widely used in practice. The essence of support vector machines is to construct separating surfaces that will minimize the upper bound on the out-of-sample error. In the case of one linear surface (plane) separating the elements from two classes, this approach will choose the plane that maximizes the sum of the distances between the plane and the closest elements from each class (i.e., the “gap” between the elements from different classes). The mathematical definition of support vector machines can be described as follows. Let all the data elements be represented as n-dimensional vectors (or points in the n-dimensional space), then these elements can be separated geometrically by constructing the surfaces that serve as the “borders” between different groups of points. One of the common approaches is to use linear surfaces (planes) for this purpose, however, different types of nonlinear (e.g., quadratic) separating surfaces can be considered in certain applications. Note that, in practice, it is not possible to find a surface that would “perfectly” separate the points according to the value of some attribute. In other words, data points with different values of the given attribute may not necessarily lie at the different sides of the surface. However, in general, the number these errors should be small enough. The classification problem of support vector machines can be represented as the problem of finding geometrical parameters of the separating surface(s). As it will be described below, these parameters can be found by solving the optimization problem of minimizing the misclassification error for the elements in the training dataset (so-called “in-sample error”). After determining these parameters, every new data element will be automatically assigned to a certain class, according to its geometrical location in the elements space. The procedure of using the existing dataset for classifying new elements is often called “training the classifier” (and the corresponding dataset is referred to as the “training dataset”). It means that the parameters of separating surfaces are “tuned” (or, “trained”) to fit the attributes of the existing elements to minimize the number of errors in their classification. However, a crucial issue in this procedure is not to “overtrain” the model, so that it would have enough flexibility to classify new elements, which is the primal purpose of constructing the classifier. An example of hyperplanes separating the brain’s pre-seizure, normal, and post-seizure states is illustrated in Fig. 3.

3.2. SVMs Mathematical Formulation. The main idea of applying SVMs to the classification of EEG time series is to embed EEG data (both normal and pre-seizure) into higher dimensional space and try to find a hyperplane to separate the data. The problem can be formally defined as follows. Let all the EEG data samples be represented as n-dimensional vectors (or points in the n-dimensional space). A very common SVM approach is to find a plane which would separate all the vectors (points) in the n-dimensional space defined in A from the vectors in B. If a plane is defined by the standard expression $x^T \omega = \gamma$, where $\omega = (\omega_1, \ldots, \omega_n)^T$ is an n-dimensional vector of real numbers, and γ is a scalar, then this plane will separate all the elements from A and B. Thus, the discrimination rules can be formulated as an optimization problem to determine vectors ω and γ such that the separating hyperplane defines two open half spaces,

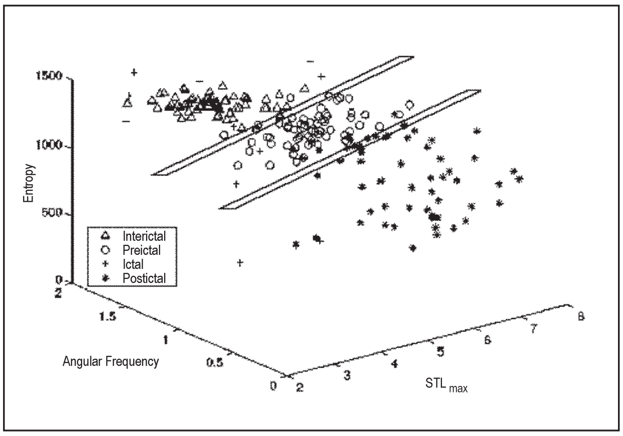


Fig. 3. Example of hyperplanes separating different brain's states.

$$\{x \mid x \in \Re^n, \, x^T \omega < \gamma\}$$

and

$$\{x \mid x \in \Re^n, \, x^T \omega > \gamma\},$$

which contain most data points in A and B respectively. However, in practice it is usually not possible to perfectly separate two sets of elements by a plane. For this reason, one should try to minimize the average measure of misclassifications. The violations of these constraints are modeled by introducing nonnegative variables u and v. The most common mathematical model for SVMs that minimizes the total average measure of misclassification errors is given by:

$$\min_{\omega,\gamma,u,v} \frac{1}{m} \sum_{i=1}^{m} u_i + \frac{1}{k} \sum_{j=1}^{k} v_j \tag{1}$$

$$\text{s.t.} \quad A\omega + u \ge e\gamma + e, \tag{2}$$

$$A\omega - v \le e\gamma - e, \tag{3}$$

$$u \ge 0, \;\; v \ge 0. \tag{4}$$

As one can see, this is a linear programming problem, and the decision variables here are the geometrical parameters of the separating plane ω and γ, as well as the variables representing misclassification error u and v. Although in many cases this type of problems may involve high dimensionality of data, they can be efficiently solved by available LP solvers, for instance Matlab, Xpress-MP, or CPLEX.

3.3. Time Series Kernel Functions. In this section, we will discuss the use of kernel functions in time series classification, one of the most widely used technique in SVM learning. Generally, kernel functions are used to extend the decision functions of SVMs to the nonlinear separation case. The main idea of kernel functions is to map the data from the input space X into a high dimensional feature space $\mathcal{X}$ by a function

$$\Phi : X \to \mathcal{X}$$

and solving the linear learning problem to find a separating hyperplane in $\mathcal{X}$. The actual kernel function Φ does not need to be known, it suffices to have a kernel function k, which calculates the inner product in the feature space

$$k(x, y) = \Phi(x) \cdot \Phi(y).$$

The kernel function can be viewed as a similarity (distance) measure in the input space [46]. The similarity between the samples x and y can be shown as the kernel function $k(x, y)$ as the following:

$$d^2(x, y) = (\Phi(x) - \Phi(y))^2 = k(x,x) - 2k(x, y) + k(y, y).$$

3.3.1. Linear Kernel. The most simple kernel function is the linear kernel, $k(x, y) = x \cdot y$. The decision function takes the formula, $f(x) = wx + b$. In time series prediction, the linear kernel can be interpreted as an statistical autoregressive model of the order k (AR[k]). This can be shown by $x_T = f(x_{T-1}, \ldots, x_{T-k}) = \sum_{t=1}^{k} w_t x_{T-t} + b$. The interpretation of this kernel function is that the time series are considered to be similar if they are generated by the same AR-model.

3.3.2. Radial Basis Function Kernel. Another commonly used kernel function is the radial basis function (RBF) kernel, $k_\gamma(x, y) = \exp(-\gamma \|x - y\|^2)$. The similarity of two samples in the RBF kernel can be interpreted as their euclidian distance. In time series prediction, the RBF kernel, in turn, has a parallel in the phase space representation. This can be explained as follows. Assume the time series is generated by a function f such that $x_T = f(x_{T-1}, \ldots, x_{T-k})$. If one takes the time series $x_1, \ldots, x_k, \ldots, x_N$ and plots it in the $(k+1)$-dimensional phase space. It can be easily observed that the resulting plot is a part of the graph of f, so the function f can be estimated from the time series. Especially, assuming that the function is linear and the time series is generated by $x_T = f(x_{T-1}, \ldots, x_{T-k}) + \eta$, where η is a Gaussian noise. Clearly, the time series model is AR[1]) and it can be shown that most of the time series data points lies in an ellipsoid defined by the mean of the time series and the variance of η. The interpretation of this kernel function is that the time series are considered to be similar in means of the euclidian distance in the phase space.

3.3.3. Fourier Kernel. Fourier transform is among the most common transformation in time series analysis. The Fourier kernel function is advantageous when the information or pattern of the time series does not lie in the individual values at each time point but in the frequency of some events. The inner product of the Fourier expansion of two time series can be directly calculated by the regularized kernel function, $k_F(x, y) = \dfrac{1 - q^2}{2(1 - 2q\cos(x - y)) + q^2}$, where $0 < q < 1$ and $X \subset [0, 2\pi]^n$ [49].

4. METHODS

In this section, we describe the methods used in each step of multi-dimensional EEG classification starting from quantifying the brain dynamics from EEG signals to implementing the dynamic time warping kernel with SVMs to analyze the multidimensional time series of the brain dynamics.

4.1. Quantification of the Brain Dynamics. Quantification of the brain dynamics from EEGs in this study is suitable to the investigation of a nonstationary system such as the brain because it is capable of automatically identifying and appropriately weighing existing transients in the data. This technique is motivated by mathematical models from chaos theory used to characterize multi-dimensional complex systems and reduce the dimensionality of EEGs [1, 26, 37, 42, 48]. To quantify the brain dynamics, we divide EEG signals into sequential 10.24-second epochs (non-overlapping windows) to properly account for possible nonstationarities in the epileptic EEG. For each epoch of each channel of EEG signals, we estimate the measures of chaos to quantify the chaoticity of the attractor. These measures include Short-Term Maximum Lyapunov Exponent and Angular Frequency. A chaotic system like human brain is a system in which orbits that originate from similar initial conditions or nearby points in the phase space diverge exponentially in expansion process. The rate of divergence is an important aspect of the dynamical system and is reflected in the value of Lyapunov exponents and Angular Frequency. In other words, the Lyapunov exponents and Angular Frequency measure the average uncertainty along the local eigenvectors and phase differences of an attractor in the phase space, respectively. Next, we will give a short overview of mathematical models used in the estimation of the Short-Term Maximum Lyapunov Exponent and Angular Frequency from EEG signals.

4.1.1. EEG Time Series Embedding. In the study of the brain dynamics, the initial step in analyzing the dynamical properties of EEG signals is to embed it in a higher dimensional space of dimension *p*, which enables us to capture the behavior in time of the *p* variables that are primarily responsible for the dynamics of the EEG. We can now construct *p*-dimensional vectors $X(t)$, whose components consist of values of the recorded EEG signal $x(t)$ at p points in time separated by a time delay. Construction of the embedding phase space from a data segment $x(t)$ of duration T is made with the method of delays. The vectors X_i in the phase space are constructed as:

$$X_i = (x(t_i), x(t_i + \tau) \ldots x(t_i + (p-1) * \tau)),$$

where τ is the selected time lag between the components of each vector in the phase space, p is the selected dimension of the embedding phase space, and $t_i \in [1, T-(p-1)\tau]$. The vectors X_i in the phase space are illustrated in Fig. 4.

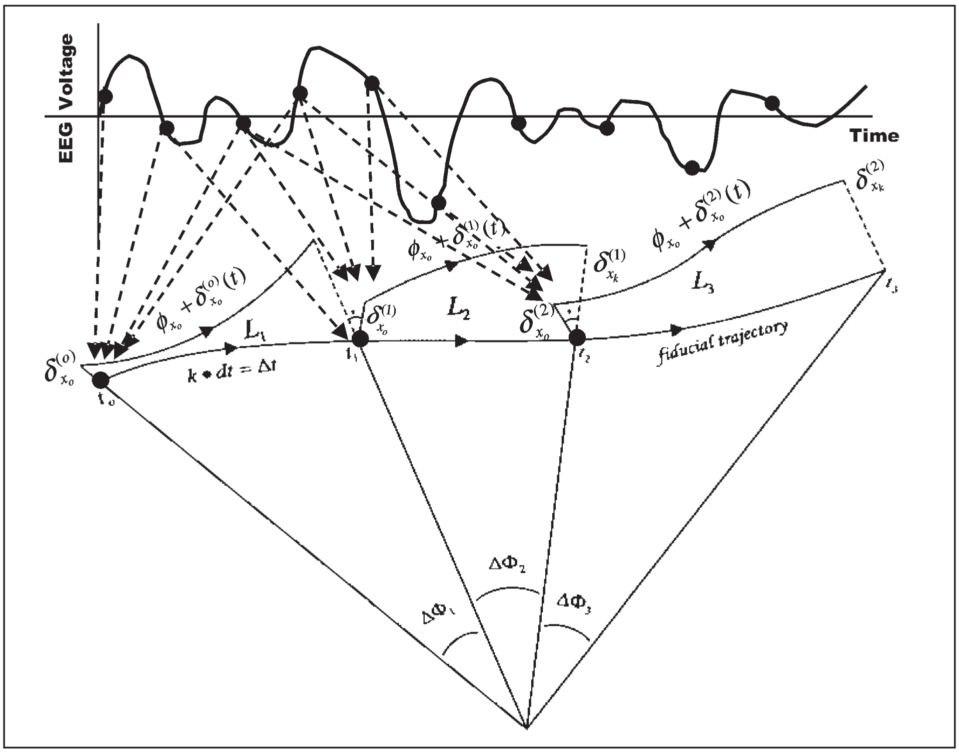


Fig. 4. Diagram illustrating an EEG epoch embedded in phase space for the quantification of brain dynamics: assume $p = 4$. The fiducial trajectory, the first three local Lyapunov exponents ($L1$, $L2$, $L3$), is shown.

4.1.2. Estimation of Short-Term Maximum Lyapunov Exponent ($STL_{\max}$). The method for estimation of $STL_{\max}$ for nonstationary data (e.g., EEG time series) is previously explained in [23, 51]. In this section, we will only give a short description and basic notation of our mathematical models used to estimate $STL_{\max}$. First, let us define the following notation:

— Δt is the evolution time for $\delta X_{i,j}$, that is, the time one allows $\delta X_{i,j}$ to evolve in the phase space. If the evolution time Δt is given in second, then L is in bits per second.

— t_0 is the initial time point of the fiducial trajectory and coincides with the time point of the first data in the data segment of analysis. In the estimation of $STL_{\max}$, for a complete scan of the attractor, t_0 should move within $[0, \Delta t]$.

— N_a is the number of local $STL_{\max}$'s that will be estimated within a duration T data segment. Therefore, if D_t is the sampling period of the time domain data, $T = (N-1)D_t = N_a \Delta t + (p-1)\tau$.

— $X(t_i)$ is the point of the fiducial trajectory $\phi_t(X(t_0))$ with $t = t_i$, $X(t_0) = (x(t_0), \ldots x(t_0 + (p-1) * \tau))$, and $X(t_j)$ is a properly chosen vector adjacent to $X(t_i)$ in the phase space.

— $\delta X_{i,j}(0) = X(t_i) - X(t_j)$ is the displacement vector at t_i, that is, a perturbation of the fiducial orbit at t_i, and $\delta X_{i,j}(\Delta t) = X(t_i + \Delta t) - X(t_j + \Delta t)$ is the evolution of this perturbation after time Δt.

— $t_i = t_0 + (i-1) * \Delta t$ and $t_j = t_0 + (j-1) * \Delta t$, where $i \in [1, N_a]$ and $j \in [1, N]$ with $j \neq i$.

$STL_{\max}$ is defined as the average of local Lyapunov exponents in the state space and can be calculated by the following equation:

$$STL_{\max} = \frac{1}{N_a \Delta t} \sum_{i=1}^{N_a} \log_2 \frac{|\delta X_{i,j}(\Delta t)|}{|\delta X_{i,j}(0)|}.$$

4.1.3. Estimation of Angular Frequency ($\overline{\Omega}$). Similar to the estimation of $STL_{\max}$, the estimation of the Angular Frequency, $\overline{\Omega}$, is motivated by the representation of a state as a vector in the state space. $\overline{\Omega}$ is merely an average uncertainty along the phase differences of an attractor in the phase space. First, let us define the difference in phase between two evolved states $X(t_i)$ and $X(t_i + \Delta t)$ as $\Delta\Phi_i$. Then, denoting with $(\Delta\Phi)$ the average of the local phase differences $\Delta\Phi_i$ between the vectors in the state space, we have

$$\Delta\Phi = \frac{1}{N_a} \sum_{i=1}^{N_a} \Delta\Phi_i,$$

where N_a is the total number of phase differences estimated from the evolution of $X(t_i)$ to $X(t_i + \Delta t)$ in the state space, according to

$$\Delta\Phi_i = \left| \arccos \frac{X(t_i) \cdot X(t_i + \Delta t)}{\|X(t_i)\| \cdot \|X(t_i + \Delta t)\|} \right|.$$

Then, the average angular frequency $\overline{\Omega}$ is defined as

$$\overline{\Omega} = \frac{1}{\Delta t} \cdot \Delta\Phi.$$

If Δt is given in second, then $\overline{\Omega}$ is given in rad/sec. Thus, while $STL_{\max}$ measures the local stability of the state of the system on average, $\overline{\Omega}$ measures how fast a local state of the system changes on average (e.g., dividing $\overline{\Omega}$ by 2π, the rate of the change of the state of the system is expressed in $\sec^{-1} = Hz$).

4.2. Dynamic Time Warping Kernel. Give two time series (or vector sequences) X and Y of equal length $|X| = |Y| = n$, pattern similarity is determined by aligning time series X with time series Y with the distortion of alignment $D_{\text{align}}(X, Y)$. Dynamic time warping (DTW) is used to compute the best possible alignment warp between two time series by selecting the one with the minimum distortion. In other words, The DTW distance is a distance measure (or similarity measure) between two time series by computing the best possible alignment or the minimum mapping (aligning) distance between two time series. In this study, all our EEG data samples are equal in length; however, the DTW can be extended to the case where the lengths of the two time series are not equal. DTW has been widely used in many contexts including data mining [30, 2], gesture recognition [16], robotics [44], speech processing [41, 47, 50], and medicine [8].

The problem of calculating the DTW distance can be solved by a dynamic programming approach. The basic concept can be described as follows. First, construct an alignment in an $n \times n$ matrix so that each vector (data points) in X is matched with a corresponding vector in Y. Typically, the Euclidean distance is used as the local distance between two vectors, $d(x_i, y_j) = (x_i - y_j)^2$, where the ($i$th, jth) element of the matrix is the distance $d(x_i, y_j)$ between the ith point of time series X, and the jth point of time series Y. Then, we construct a warp path $W = w_1, \ldots, w_K$, where K is the length of the warp path and $\max(|X|,|Y|) \le$ $\le K < |X|+|Y|$. The kth element of the warp path represents the matching point of two time series, $w_k = (i, j)$, where (i, j) corresponds to index i from time series X, and index j from time series Y (see Fig. 5). The warp path must start at the beginning of each time series and finish at the end of both time series. In other words, the path starts from the beginning of each time series, $w_1 = (1,1)$, and finishes at the end of both time series, $w_K = (n, n)$. The warp path can actually be calculated in reverse order starting at the end of both time series. There is also a constraint on the warp path that forces indices i and j to be monotonically increasing in the warp path; that is, $w_k = (i, j)$ and $w_{k+1} = (i', j')$ where $i \le i' \le i+1$ and $j \le j' \le j+1$. Note that there can be an exponential number of warping paths that satisfy the above conditions. However, the optimal warp path is the one with a minimum warping (distortion) cost defined by

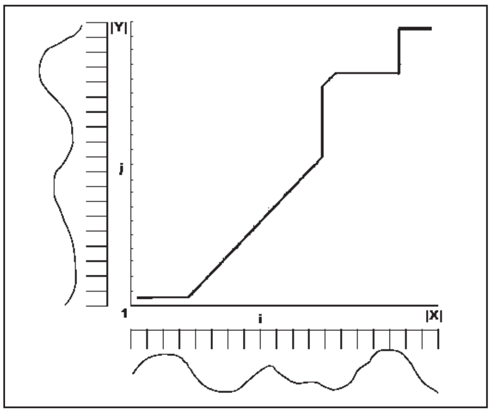


Fig. 5. A warping matrix with the minimum-distance warp path of two time series X and Y.

$$D_{\text{align}}(X,Y) = \min \frac{1}{K} \sum_{k=1}^{K} d(w_{ki}, w_{kj}) .$$

In a dynamic programming approach, the warp path must either be incremented by one unit (adjacent) or stay at the same i or j axes. Therefore, we only need to evaluate the recurrence of the cumulative distance found in the adjacent elements:

$$D(i,j) = d(x_i, y_j) + \min \begin{cases} D(i, j-1), \\ D(i-1, j), \\ D(i-1, j-1). \end{cases}$$

4.3. SVMs with DTW Kernel. The essence of the SVMs framework with DTW kernel in this application can be described as follows. Based on the concept of DTW, one employ a Euclidean distance measure to find the optimal path that minimizes the accumulated distance of the warping path. The SVMs with DTW uses inner product or kernel function to find the optimal path that maximize the accumulated similarity (or minimize the distance) as follows:

$$k_{DTW}(x, y) = \max_{\psi_I, \psi_J} \frac{1}{M_\psi} \sum_{k=1}^{L} m(k) x_{\psi_I(k)} y_{\psi_J(k)}, \tag{5}$$

$$\text{s.t. } 1 \le \psi_I(k) \le \psi_I(k+1) \le L, \tag{6}$$

$$1 \le \psi_J(k) \le \psi_J(k+1) \le L, \tag{7}$$

where $L = |X| = |Y|$, $\psi_I(k)$ and $\psi_J(k)$ are linear warping functions, $m(k)$ is a nonnegative (path) weighting coefficient, and M_ψ is a (path) normalizing factor [47]. The linear discriminant function of SVMs with DTW kernel for time series classification can be then expressed the same way as the original linear SVMs function except using the DTW kernel. It is important to note that, in our case, we have multiple time series; therefore, the similarity of the DTW kernel will be based on a pair-wise manner. In other words, if we let N be the total number of electrodes, then we have to calculate a total of $\frac{N(N-1)}{2}$ kernel functions or similarity indices. These similarity indices can, in turn, be considered as the attributes of the input data. Finally, note that the same algorithms used to solve a standard SVMs can be used to solve the SVMs with DTW kernel as well.

5. EMPIRICAL STUDY

The underlying hypothesis in this empirical study is that the proposed SVMs with DTW kernel is capable of discriminating/classifying different physiological stages (normal and pre-seizure) of the brain. The features of input data are in a form of time series of the the brain dynamics measure (i.e., $STL_{\max}$ and $\overline{\Omega}$). In this section, we discuss in detail each step of our empirical study.

5.1. EEG Data Acquisition. The datasets in this study consisted of continuous long-term (3 to 9 days) multichannel intracranial EEG recordings from bilaterally, surgically implanted macroelectrodes in the hippocampus, temporal and frontal lobe cortexes of 3 epileptic patients with medically intractable temporal lobe epilepsy (outlined in Table 1). The recordings were obtained as part of a pre-surgical clinical evaluation, using a Nicolet BMSI 4000 recording system with amplifiers of an input range of 0.6 mV, sampling rate of 200 Hz and filters with a frequency range of a 0.5–70Hz. Each recording included a total of 28 to 32 intracranial electrodes (8 subdural and 6 hippocampal depth electrodes for each cerebral hemisphere, and a strip of 4 additional electrodes if deemed necessary by the neurologist). Note that we only use the EEG recording from 26 electrodes in this study as those electrodes are most commonly used. The recorded EEG signals were digitized and stored on magnetic media for subsequent off-line analysis. These EEG recordings were viewed by two independent electroencephalographers to determine the number and type of recorded seizures, seizure onset and end times, and seizure onset zones.

Table 1. EEG dataset characteristics

Patient ID	Gender	Age	Seizure Types	Duration of EEG, days	Number of Seizure
1	F	41	CP	9.06	24
2	M	45	CP, SC	3.63	9
3	M	29	CP, SP	6.07	19
Total				18.76	59

CP — Complex Partial; SC — Subclinical

5.2. Data Sampling and Pre-Processing. In this study, the classification will be performed separately for each subject. Per individual, we use the Monte-Carlo sampling technique to randomly select EEG data from 2 groups (normal and pre-seizure states) from the continuous recordings. Each data sample contains a 5-minute epoch of EEG data from 26 electrodes. Note that in this analysis we only consider clinical seizures and un-clustered seizures. From the data set, we consider 22, 7, and 15 seizures in the EEG data from Patients 1, 2, and 3, respectively. Since the data set of each patient is very much different in length and the total number of seizures, for each patient we randomly select three epochs of pre-seizure EEG data per seizure. In other words, 66, 21, and 45 epochs of the EEG data are selected from the group of pre-seizure EEG data in Patients 1, 2 and 3, respectively. Per patient, 200 epochs of EEG data from the normal state are randomly and uniformly sampled. The criteria used in determining normal and pre-seizure states of EEG data is as follows. Normal EEG samples are selected from EEG recordings that is more than 8 hours apart from a seizure. Pre-seizure EEG samples are selected from EEG recordings during the 30-minute interval before. For instance, we analyze 22 seizures from Patient 1’s data; therefore, 266 EEG epochs (200 normal and 66 pre-seizure) are sampled. After EEG data are sampled, we first calculate measures of the brain dynamics (i.e., $STL_{\max}$ and $\overline{\Omega}$) from the EEG data using the methods described in Sec. 4. Each measure is calculated continuously for each non-overlapping 10.24-second segment of EEG data; therefore, each of EEG epoch contains 30 data points of the brain dynamical time series.

5.3. Classification Procedure. The input data consist of 66, 21, and 45 of $N*m$-dimensional time series, where N is the number of electrodes and m is the length of each EEG epoch times two (2 dynamical measures). We then calculate a pair-wise kernel function for every electrode pair of multi-dimensional EEG time series. Therefore, the total number of feature vectors corresponding to Patients 1, 2, and 3 is $\frac{N(N-1)}{2}*2=650$. The method used to calculate the kernel function is described in Sec. 4. Subsequently, we employ SVMs to classify these EEG data. We then use Matlab to solve the constructed SVMs model is to find a plane which would separate all the vectors of normal and pre-seizure EEGs.

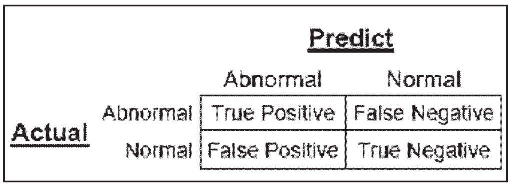
5.4. Training and Testing. In this section, we describe the step of how the SVMs are trained and tested. There are many choices of how to divide the data into training and test sets. In order to reduce the bias of training and test data, we propose to implement a leave-one-out cross validation scheme, extensively used as a method to estimate the generalization error based on "resampling". It is important to note that the classification techniques will be trained and tested individually for each patient. To train the SVMs, it is important to note that, in general, the training of support vectors machines is optimized when the number of pre-seizure and normal samples are comparable. Otherwise, the SVMs will be biased to classify most samples to the physiological state with larger size samples. To adequately implement the SVMs, we train the classifier with the same number of pre-seizure and normal samples by implementing Monte Carlo sampling simulation. First, we shuffle (random order) the pre-seizure and normal samples individually. Since the size of pre-seizure samples is much larger than the size of normal samples, the number of pre-seizure samples will be used to determine the size of the training and testing sets. Then, we divide the first of pre-seizure samples for the training and the other half for the testing. After that, we randomly select training data (with the same size) from normal samples. For individual patient, we run the simulation 100 times.

6. RESULTS

To evaluate the performance characteristics of the proposed classification technique, we calculate the sensitivity and specificity of the proposed classification technique. These results will be discussed in this section.

6.1. Performance Evaluation of Classification Schemes. In general, to evaluate the classifier, we categorize the classification into two classes: positive (pre-seizure) and negative (normal). Then we consider four subsets of classification results: 1. True positives (TP) — denoting correct classifications of positive cases. 2. True negatives (TN) — denoting correct classifications of negative cases. 3. False positives (FP) — denoting incorrect classifications of negative cases into class positive. 4. False negatives (FN) — denoting incorrect classifications of positive cases into class negative.

To better explain the concept of the evaluation of classifiers, let us consider in the case of the detection of pre-seizure EEG data (see Fig. 6). A classification result was considered to be true positive if we classify a pre-seizure EEG sample as a pre-seizure sample. A classification result was considered to be true negative if we classify a normal EEG sample as a normal sample. A classification result was considered to be false positive when we classify a normal EEG sample as a pre-seizure sample. A classification result was considered to be false negative when we classify a pre-seizure EEG sample as a normal sample.



		Predict	
		Abnormal	Normal
Actual	Abnormal	True Positive	False Negative
	Normal	False Positive	True Negative

Fig. 6. The evaluation concept of classification results. Note that we use "pre-seizure" and "abnormal" interchangeably.

Sensitivity and specificity are widely used in the medical domain as classification performance measures Sensitivity measures the fraction of positive cases that are classified as positive. Specificity measures the fraction of negative cases classified as negative. The sensitivity and specificity are defined as follows:

$$\text{Sensitivity } = \frac{TP}{TP+FN}, \qquad \text{Specificity } = \frac{TN}{TN+FP}.$$

In fact, the sensitivity can be considered as a probability of accurately classifying EEG samples in the pre-seizure case. The specificity can be considered as a probability of accurately classifying EEG samples in the normal case. In general, a good classifier is the one with high sensitivity and specificity.

Table 2. Performance characteristics of the support vector machine classifier for individual patient

Patient	SVMs [11]			SVMs-DTW		
	Sensitivity, %	Specificity, %	Overall, %	Sensitivity, %	Specificity, %	Overall, %
1	81.21	87.46	84.34	92.15	93.84	92.99
2	71.18	76.85	74.02	79.45	78.07	78.76
3	74.13	70.60	72.37	80.13	84.66	82.39
Average	75.51	78.30	76.91	83.91	85.52	84.72

6.2. Performance Characteristics of the Proposed Classification Methods. After running 100 simulations, we then report the average classification performance in this section. Figure 7 illustrates the overall classification results of the proposed SVMs-DTW algorithm. Table 2 and Fig. 8 illustrate a performance comparison of the standard SVMs for EEG classification proposed in [11] versus the results from the proposed SVMs with DTW proposed in this paper tested on 3 patients. For all cases, the incorporation of the DTW kernel function, we can achieve substantially better classification results (about 8% better on average). In Patient 1, the proposed algorithm achieve about 92% sensitivity and over 93% specificity on average. This result demonstrates the improvement in classification performance of almost 8% on average. In Patient 2, the proposed algorithm achieve about 80% sensitivity and about 78% specificity on average. This result demonstrates the improvement in classification performance of almost 5% on average. In Patient 3, the proposed algorithm achieve about 84% sensitivity and over 85% specificity on average. This result demonstrates the improvement in classification performance of over 10% on average. Overall, the proposed SVMs-DTW can achieve the sensitivity of correctly classifying pre-seizure of 83.91%, and the specificity of correctly classifying normal EEGs of 85.52%, respectively. This reflects to almost 8% accurate classification improvement. In Fig. 8, we observe that the incorporation of the DTW kernel function can improve the classification performance of SVMs in every case.

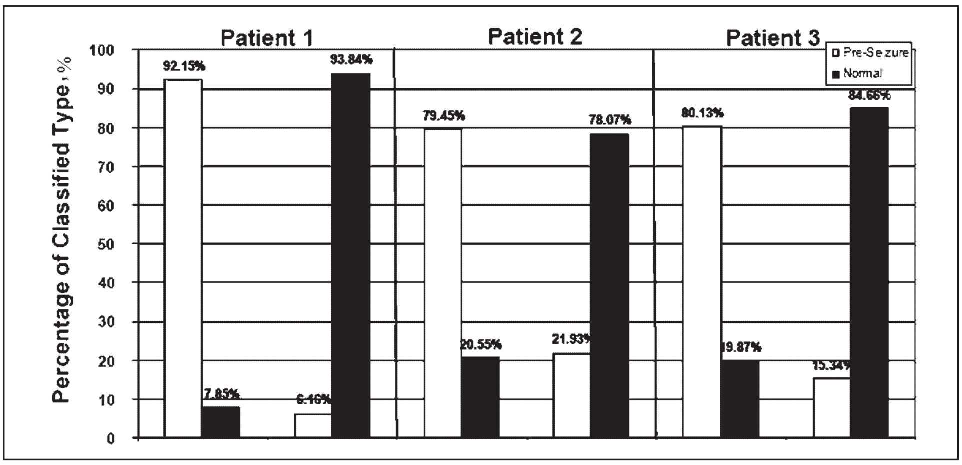


Fig. 7. Average classification results among all 3 patients using leave-one-out cross validation to train and test the algorithm over 100 simulations.

It is very interesting to note that the classification performance of our algorithm for each patient is consistent with the standard SVMs [11]. Specifically, the EEG data from Patient 1 tend to be more classifiable than those of Patients 2 and 3. We speculate that the number of seizures in the EEG data set could play a very important role in terms of

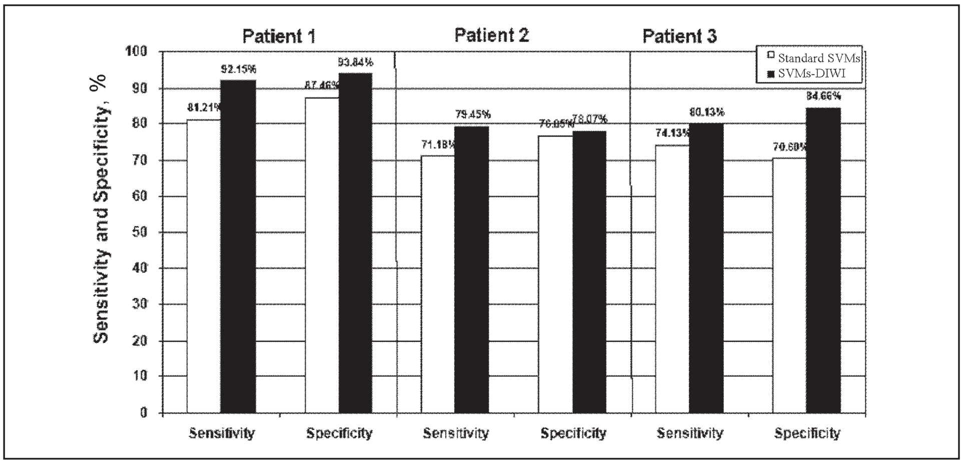


Fig. 8. Performance comparison of the standard SVMs for EEG classification proposed in [11] versus the the proposed SVMs-DTW algorithm among all 3 patients using leave-one-out cross validation to train and test the algorithm over 100 simulations.

providing more training data for abnormal (pre-seizure) EEGs. Because our algorithm yields the worst classification results in Patient 2 among all 3 cases, it is very intuitive to claim that there is so much the classifier can learn from 7 seizure samples as apposed to 22 and 15 samples. Nonetheless, these results confirm our hypothesis that the brain's states are classifiable based on the brain dynamics measures and data mining techniques applied to EEG signals. The framework of classifiers proposed in this study can be extended to development of an abnormal brain activity classifier or an online brain activity monitoring.

7. CONCLUDING REMARKS

This study addresses the open question of the classifiability of the brain's preseizure and normal EEGs. The results of this study is another proof concept of the application of the quantification of the brain dynamics and data mining techniques. This framework was proved successful in providing insights and characterizing different states of brain activities reflected from pathological dynamical interactions of brain network. In addition, these results also confirm our hypothesis that it is possible to differentiate and classify the brain's pre-seizure and normal activities based on optimization, data mining, and dynamical system approaches in multichannel intracranial EEG recordings. Also, the incorporation of DTW kernel function with SVMs is very straightforward and not difficult to implement. The optimization problems in the framework of support vector machines can be solved in reasonable time. All of the programming was done in Matlab environment on a desktop computer Pentium IV 2.4 GHz with 1 GB of RAM. The proposed technique very fast and scalable. The running time for statistical cross-validation technique is less than 5 minutes on average. In the future, more cases (patients and seizures) will be studied to validate the observation across patients as well as the development of multi-class classifier based on support vector machines framework. In addition, the feature selection study will be possible in the future. This study will help us to select electrodes that show prominent changes, which might lead us to the solution to the epileptogenic zone localization problem.

REFERENCES

1. Babloyantz A., Destexhe A. Low dimensional chaos in an instance of epilepsy // Proc. Nat. Acad. Sci. USA. — 1986. — **83**. — P. 3513–3517.
2. Berndt D., Clifford J. Using dynamic time warping to find patterns in time series // Proc. of the AAAI–94 Workshop on Knowledge Discovery in Databases (KDD-94). — 1994.
3. Bertsimas D., Darnell C.R., Soucy R. Portfolio construction through mixed-integer programming at grantham, mayo, van otterloo and company // Interfaces. — 1999. — **29**, N 1. — P. 49–66.
4. Borgwardt M.K., Vishwanathan S.V.N., Kriegel H-P. Class prediction from time series gene expression profiles using dynamical systems kernels // Pacific Symp. on Biocomput. — 2006. — P. 547–558.
5. Bradley P.S., Fayyad U., Mangasarian O.L. Mathematical programming for data mining: Formulations and challenges // INFORMS J. Computing. — 1999. — **11**. — P. 217–238.

6. Bradley P.S., Mangasarian O.L., Street W.N. Clustering via concave minimization. // M.C. Mozer, M.I. Jordan, and T. Petsche, editors. Adv. in Neural Inform. Proces. Systems. — Cambridge: MIT Press, 1997. — P. 368–374.
7. Breiman L., Friedman J., Olsen R., Stone C. Classification and regression trees. — Belmont: Wadsworth Inc, 1993.
8. Caiani E.G., Porta A., Baselli G. et al. Warped-average template technique to track on a cycle-by-cycle basis the cardiac filling phases on left ventricular volume // IEEE Comput. in Cardiology. — 1998. — **25**, N 98. — CH36292.
9. Chaovalitwongse W.A., Pardalos P.M., Iasemidis L.D. et al. Applications of global optimization and dynamical systems to prediction of epileptic seizures // P.M. Pardalos, J.C. Sackellares, L.D. Iasemidis, P.R. Carney, editors. Quantitative Neuroscience. — Dordrecht: Kluwer, — 2003. — P. 1–36.
10. Chaovalitwongse W.A., Pardalos P.M., Prokoyev O.A. Reduction of multi-quadratic 0–1 programming problems to linear mixed 0–1 programming problems // Oper. Res. Letters. — 2004. — **32**, N 6. — P. 517–522.
11. Chaovalitwongse W.A., Pardalos P.M., Prokoyev O.A. Electroencephalogram (EEG) time series classification: Applications in epilepsy // Ann. Oper. Res. — 2006. — **148**, N 1. — P. 227–250.
12. Dasgupta D., Forrest S. Novelty detection in time series data using ideas from immunology // Intern. Conf. on Intell. Systems. — 1999.
13. Diez J.J.R., Gonzalez C.A. Applying boosting to similarity literals for time series classification // Intern. Workshop on Multiple Classifier Systems. — 2000. — P. 210–219.
14. Elger C.E., Lehnertz K. Seizure prediction by non-linear time series analysis of brain electrical activity // Eur. J. Neurosci. — 1998. — **10**. — P. 786–789.
15. Fung G.M., Mangasarian O.L. Proximal support vector machines // 7th ACM SIGKDD Intern. Conf. on Knowledge Discovery and Data Mining. — 2001.
16. Gavrila D.M., Davis L.S. Towards 3-d model-based tracking and recognition of human movement: a multi-view approach // Proc. of the Intern. Workshop on Autom. Face- and Gesture-Recognition. — 1995.
17. Geurts P. Pattern extraction for time series classification // Principles of Data Mining and Knowledge Discovery, 5th Eur. Conf. — 2001. — P. 115–127.
18. Grossman R.L., Kamath C., Kegelmeyer P. et al. Data mining for scientific and engineering applications. — Dordrecht: Kluwer Acad. Publ., 2001. — 628 p.
19. Hand D.J., Mannila H., Smyth P. Principle of data mining. — Concord: Bradford Books, 2001. — 584 p.
20. Hsu C.-W., Lin C.-J. A comparison of methods multi-class support vector machines // IEEE Trans. on Neural Networks. — 2002. — **13**. — P. 415–425.
21. Hurn A.S., Lindsay K.A., Michie C.A. Modelling the lifespan of human *t*-lymphocite subsets // Math. Biosciences. — 1997. — **143**. — P. 91–102.
22. Iannatilli F.J., Rubin P.A. Feature selection for multiclass discrimination via mixed-integer linear programming // IEEE Trans. on Pattern Analysis and Machine Learning. — 2003. — **25**. — P. 779–783.
23. Iasemidis L.D. On the dynamics of the human brain in temporal lobe epilepsy: PhD thesis. — Univ. of Michigan, Ann Arbor. — 1991.
24. Iasemidis L.D., Pardalos P.M., Sackellares J.C., Shiau D.-S. Quadratic binary programming and dynamical system approach to determine the predictability of epileptic seizures // J. Comb. Optimiz. — 2001. — **5**. — P. 9–26.
25. Iasemidis L.D., Shiau D.-S., Chaovalitwongse W.A. et al. Adaptive epileptic seizure prediction system // IEEE Trans. Biomed. Eng. — 2003. — **5**, N 5. — P. 616–627.
26. Iasemidis L.D., Zaveri H.P., Sackellares J.C., Williams W.J. Phase space analysis of eeg in temporal lobe epilepsy // IEEE Eng. in Medicine and Biology Soc., 10th Ann. Intern. Conf. — 1988. — P. 1201–1203.
27. Keogh E., Chakrabarti K., Pazzani M., Mehrotra S. Dimensionality reduction for fast similarity search in large time series databases // Knowledge and Inform. Systems. — 2000. — **3**, N 3. — P. 263–286.
28. Keogh E., Kasetty S. On the need for time series data mining benchmarks: A survey and empirical demonstration // 8th ACM SIGKDD Intern. Conf. on Knowledge Discovery and Data Mining. — 2002. — P. 102–111.
29. Keogh E., Pazzani M. An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback // 4th Int’l Conf. on Knowledge Discovery and Data Mining. — 1998. — P. 239–241.
30. Keogh E., Pazzani M. Scaling up dynamic time warping for datamining applications // Proc. of the 6th ACM SIGKDD Intern. Conf. on Knowledge Discovery and Data Mining. — 2000. — P. 285–289.
31. Krebel U. Pairwise classification and support vector machines // Adv. in Kernel Methods — Support Vector Learning. — Cambridge: MIT Press, 1999. — P. 255–268.
32. Lehnertz K., Elger C.E. Can epileptic seizures be predicted? Evidence from nonlinear time series analysis of brain electrical activity // Phys. Rev. Lett. — 1998. — **80**. — P. 5019–5022.
33. Litt B., Esteller R., Echauz J. et al. Epileptic seizures may begin hours in advance of clinical onset: A report of five patients // Neuron. — 2001. — **30**. — P. 51–64.
34. Mangasarian O.L. Linear and nonlinear separation of pattern by linear programming // Oper. Res. — 1965. — **31**. — P. 445–453.

35. Mangasarian O.L., Street W.N., Wolberg W.H. Breast cancer diagnosis and prognosis via linear programming // Ibid. — 1995. — **43**, N 4. — P. 570–577.
36. Martinerie J., Van Adam C., Van Quyen M.L. Epileptic seizures can be anticipated by non-linear analysis // Nature Medicine. — 1998. — **4.** — P. 1173–1176.
37. Packard N.H., Crutchfield J.P., Farmer J.D. Geometry from time series // Phys. Rev. Lett. — 1980. — **45.** — P. 712–716.
38. Pardalos P.M., Chaovalitwongse W.A., Iasemidis L.D. et al. Seizure warning algorithm based on spatiotemporal dynamics of intracranial EEG // Math. Program. — 2004. — **101**, N 2. — P. 365–385.
39. Quinlan J.R. C4.5: Programs for Machine Learning. — Orlando: Morgan Kaufmann, 1993. — 302 p.
40. Van Quyen M.L., Martinerie J., Baulac M., Varela F. Anticipating epileptic seizures in real time by non-linear analysis of similarity between eeg recordings // Neuro Rep. — 1999. — **10**. — P. 2149–2155.
41. Rabiner L., Juang B. Fundamentals of speech recognition. — Upper Saddle River: Prentice Hall, 1993. — 496 p.
42. Rapp P.E., Zimmerman I.D., Albano A.M. Experimental studies of chaotic neural behavior: cellular activity and electroencephalographic signals // H.G. Othmer, editor. Nonlinear oscillations in biology and chemistry. — New York: Springer-Verlag, 1986. — P. 175–205.
43. Rüping S. SVM kernels for time series analysis // R. Klinkenberg, S. Rüping, A. Fick, N. Henze, C. Herzog, R. Molitor, O. Schröder, editors. LLWA 01 — Tagungsband der GI–Workshop–Woche Lernen –Lehren–Wissen–Adaptivität. — 2001. — P. 43–50.
44. Schmill M., Oates T., Cohen P. Learned models for continuous planning // Proc. of the Seventh Intern. Workshop on Artif. Intell. and Statist. — 1999. — P. 278–282.
45. Scholkopf B., Burges C., Vapnik V. Extracting support data for a given task // Proc. First Intern. Conf. on Knowledge Discovery and Data Mining. — Menlo Park: AAAI Press. — 1995.
46. Schölkopf B. The kernel trick for distances: Techn. Rep. / Microsoft Research. — 2000.
47. Shimodaira H., Noma K., Naka M., Sagayama S. Support vector machine with dynamic time-alignment kernel for speech recognition // Proc. of Eurospeech. — 2001. — P. 1841–1844.
48. Takens F. Detecting strange attractors in turbulence // D.A. Rand, L.S. Young, editors. Dynamical systems and turbulence; Lecture Notes in Mathematics. — Berlin: Springer-Verlag, 1981.
49. Vapnik V.N. The nature of statistical learning. — Berlin: Springer, 1995. — 16 p.
50. Wan V., Carmichael J. Polynomial dynamic time warping kernel support vector machines for dysarthric speech recognition with sparse training data // Proc. of Interspeech. — 2005. — P. 3321–3324.
51. Wolf A., Swift J.B., Swinney H.L., Vastano J.A. Determining Lyapunov exponents from a time series. Physica D. — 1985. — **16.** — P. 285–317.
52. Yang K., Shahabi C. A pca-based kernel for kernel pca on multivariate time series // Proc. of ICDM 2005 Workshop on Temporal Data Mining: Algorithms, Theory and Applications held in conjunction with The Fifth IEEE Intern. Conf. on Data Mining (ICDM'05). — 2005.

Поступила 24.11.2006