EEG signal classification with feature selection based on one-dimension real valued particle swarm optimization

Jun Wang
Department of Communication Engineering
Harbin University of Science and Technology
Harbin, China
wangjunhit@gmail.com

Yan Zhao
Department of Electronics Engineering
Shantou University
Shantou, China
yanzhao3@stu.edu.cn

Abstract—***In this study, a new scheme was presented for the EEG signal classification with feature selection based on one-dimension real valued particle swarm optimization. In the proposed scheme, normal and abnormal EEG signals were decomposed into various frequency bands with one fourth-level wavelet packet decomposition. Approximation entropy value of the wavelet coefficients at all nodes of the decomposition tree were used as a feature set to characterize the predictability of the EEG data within the corresponding frequency bands. Then, the one-dimension real valued particle swarm optimization algorithm was used to find the optimal feature subset by maximizing the classification performance of a support vector machine based EEG signal classifier. Experimental results showed that the proposed method improved the classification performance substantially and got a much less size of optimal feature subset with compared to the other methods.***

Keywords- EEG signals, wavelet packet decomposition, approximation entropy, feature selection, particle swarm optimization

I. INTRODUCTION

The electroencephalogram (EEG), a highly complex signal, is widely used clinically to investigate brain disorders [1].The process of automated diagnosis that can be viewed as the process of doing the classification or decision making of EEG signals can generally be subdivided into a number of disjoint processing modules or stages: preprocessing, feature extraction, feature selection, and classification [2]. In the feature extraction stage, numerous different methods can be used so that several diverse features can be extracted from the same raw data such as auto-correlation function [3], frequency domain analysis methods[4], nonlinear methods[5] and etc.

Feature selection also constitutes a key development phase of pattern recognition [6]. Extensive research into feature selection has been carried out over the past four decades. The feature selection is inherently a combinatorial optimization problem that is a NP-hard problem [7]. There are many searching algorithms used to determine the promising feature subset candidates, such as exhaustive search, branch and bound search (BB), sequential forward selection (SFS), Tabu search (TB), Simulated annealing algorithm (SA), Genetic Algorithm (GA), Binary Particle Swarm Optimization (BPSO) and etc. Exhaustive search algorithm can get the optimal feature subset but its computing complexity increases exponentially with the number of original features increasing [7]. In recent years, stochastic searching algorithms for its good global search ability and relative lower computing complexity have been adopted to do feature selection. Unfortunately, all of existing stochastic searching based feature selection methods are also always falling into local optimum and cannot get optimal feature subset, especially when the number of original features is too great.

In the literature, various studies have been considered related with classifying the EEG signals. While Guler et al obtained 96.79% classification accuracy using recurrent neural networks to detect the epileptic seizure from EEG signals [8]. Subasi obtained 95% and 93.6 classification accuracies using combination wavelet transform and mixture of experts and combination wavelet transform and multilayer perceptron neural network, respectively [9]. 98.68 % accuracy was obtained using combination wavelet transform and ANFIS classifier by Guler et al [10]. Ocak obtained 94.3% and 98% classification accuracy by combining the wavelet transform, approximation entropy, genetic algorithm and LVQ classifier [11].

In this paper, a novel classification method for EEG signals is proposed. In this method, the features were extracted from the EEG signals using the wavelet packet decomposition (WPD) and approximate entropy (ApEn). This feature extraction method can extracted the nonlinear and non-stationary features of EEG signals effectively. And then one dimension real valued particle swarm optimization algorithm was employed to reduce the number of features and find the optimal feature subset by maximizing the classification performance of the SVM classifier. Experimental results demonstrated that the proposed method obtained 100% classification accuracy and only 6 features were selected for classifier training and testing which makes the training and testing of classifier more efficiently.

II. TECHNICAL BACKGROUND

2.1 Wavelet packet analysis

Wavelet packet analysis[10] is a generalized form of the discrete wavelet transform. In the wavelet packet analysis of a signal, first the signal is simultaneously passed through a series of low-pass (LP) and high-pass (HP) filters named as quadrature mirror filters. The cut-off frequency of these filters is one-fourth of the sampling frequency. The bandwidth of the filter outputs are half the bandwidth of the original signal, which allows for the down-sampling of the output signals by two without loosing any information according to the Nyquist theorem. The downsampled signals from the LP and HP filters are referred to as first-level approximation (A) and detail (D) coefficients, respectively. To get the second-level approximation of approximation (AA), detail of approximation (DA), approximation of detail (AD) and detail of detail (DD) coefficients, the same procedure is repeated for the first-level A and D coefficients. At each level of the decomposition, frequency resolution is doubled through filtering while the time resolution is doubled through filtering while the time resolution is halved by downsampling operation.

2.2 Approximation entropy

Approximation entropy (ApEn) derived from the Kolmogorov-Sinai entropy is a measure that quantifies the regularity of predictability of a time series or signal [11]. It is defined as the logarithmic likelihood that runs of patterns of certain length that are close to each other will remain close on next incremental comparisons. The first step in computing the approximation entropy of a time series, $y_i, i=1,2,\ldots,N$ is to construct the state vectors in the embedding space, R^m, using the method of delays,

$$x_i = \{y_i, y_{i+\tau}, y_{i+2\tau}, \ldots, y_{i+(m-1)\tau}\}, 1 \le i \le N-(m-1)\tau \quad (1)$$

where m and τ are the embedding dimension and time delay, respectively. Next, we define for each i,

$$C_i^m(r) = \frac{1}{N-(m-1)\tau}\sum_{j=1}\theta(r-d(x(i),x(j))) \quad (2)$$

where $\theta(x)=1$ for $x>0$, $\theta(x)=0$, otherwise is the standard Heavyside function, r is the vector comparison distance and $d(x(\text{i}),x(\text{j}))$ is a distance measure defined by,

$$d(x(i),x(j)) = \max_{k=1,2,\ldots,m}\left(\left|y(i+(k-1)\tau)-y(j+(k-1)\tau)\right|\right) \quad (3)$$

Then, $\Phi^m(r)$ can be defined as

$$\Phi^m(r) = \frac{1}{N-(m-1)\tau}\sum_{i=1}^{N-(m-1)\tau}\log C_i^m(r) \quad (4)$$

For fixed m, r and τ , ApEn is given by the formula

$$ApEn(m,r,\tau,N) = \Phi^m(r) - \Phi^{m+1}(r) \quad (5)$$

which is basically the logarithmic likelihood that runs of patterns of length m that are close (within r) will remain close on next incremental comparisons.

2.3 The principle of Particle Swarm Optimization

Particle swarm optimization (PSO) is an evolutionary computation technique developed by Kennedy and Eberhart in 1995 [12]. The original intent was to graphically simulate the graceful but unpredictable movements of a flock of birds. Initial simulations were modified to form the original version of PSO. Later, Shi introduced inertia weight into the particle swarm optimizer to produce the standard PSO. PSO is initialized with a population of random solutions, called 'particles'. Each particle is treated as a point in an S-dimensional space. The *i*th particle is represented as $X_i=(x_{i1}, x_{i2}, \ldots, x_{iS})$. The best previous position (*pbest*, the position giving the best fitness value) of any particle is recorded and represented as $P_i= (p_{i1},p_{i2}, \ldots ,p_{iS})$. The index of the best particle among all the particles in the population is represented by the symbol '*gbest*'. The rate of the position change (velocity) for particle *i* is represented as $V_i= (v_{i1}, v_{i2}, \ldots , v_{iS})$. The particles are manipulated according to the following equation:

$$\begin{aligned} v_{id} &= w * v_{id} + c_1 * rand() * (pbest_{id} - x_{id}) \\ &\quad + c_2 * Rand() * (gbest_d - x_{id}) \end{aligned} \quad (6)$$

$$x_{id} = x_{id} + v_{id} \quad (7)$$

where $d = 1,2,\ldots,S$, w is the inertia weight, it is a positive linear function of time changing according to the generation iteration. Suitable selection of the inertia weight provides a balance between global and local exploration, and results in needing much fewer iterations on average to find a sufficiently optimal solution. The acceleration constants c_1 and c_2 in Eq.(6) represent the weighting of the stochastic acceleration terms that pull each particle toward *pbest* and *gbest* positions. Low values allow particles to roam far from target regions before being tugged back, while high values result in abrupt movement toward, or past, target regions. *rand*( ) and *Rand*( ) are two random functions in the range [0, 1].

Particles' velocities on each dimension are limited to a maximum velocity, V_{max}. It determines how large steps through the solution space each particle is allowed to take. If V_{max} is too small, particles may not explore sufficiently beyond locally good regions. They could become trapped in local optima. On the other hand, if V_{max} is too high particles might fly past good solutions.

The first part of Eq.(6) provides the "flying particles" with a degree of memory capability allowing the exploration of new search space areas. The second part is the "cognition" part, which represents the private thinking of the particle itself. The third part is the "social" part, which represents the collaboration among the particles. Eq.(6) is used to calculate the particle's new velocity according to its previous velocity and the distances of its current position from its own best experience (position) and the group's best experience. Then the particle flies toward a new position according to Eq.(7). The performance of each particle is measured according to a pre-defined fitness function.

2.4 The Principle of Support Vector Machine

In the process of feature selection, a performance index or a fitness function that assesses the quality of the selected features in terms of classification error should be adopted to guide the forming of a reduced feature space. So a classifier will be used. Here we adopt the support vector machine (SVM) as the classifier [14].

Let (x_i, y_i), $1\leq i\leq N$, denote a set of training data, where N represents the number of training data. Each datum must conform to the criteria $x_i \in R^d$ and $y_i \in \{-1,1\}$, where d denotes the number of dimensions of input data. SVM attempts to identify a hyper-plane, which functions as a separating plane for classification of data, in a multidimensional space. The parameters *w* and *b* are given by

$$(<w\cdot x_i>+b)=0,\ i=1,2,\ldots,N \tag{8}$$

If a hyper-plane exists that satisfies Eq.(8), then linear separation is obtained. In this case, *w* and *b* can be rewritten as follows. Eq.(8) becomes

$$\min_{1\leq i\leq N} y_i\left(\langle w\cdot x_i\rangle\right)\geq 1, i=1,...,N \tag{9}$$

Let the distance from the data point to the hyper-plane be 1/||*w*||. Among separating hyper-planes, there exists one optimal separating hyper-plane (OSH), and the distance between two support vector points on two sides of this hyper-plane is maximal. Because the distance between two support vector points is $1/||w||^2$, the minimal distance to OSH, $||w||^2$, may be derived from Eq.(9).

The margin of a separating hyper-plane, calculated as 2/||*w*||, determines the hyper-plane's generalization ability. The OSH has the largest margin among separating hyper-planes. $||w||^2$ is minimized with Eq.(9) and Lagrange's polynomial. Let a denote ($a_1,\ldots,a_N$). Combining Lagrange's polynomial (in the order of *N*) with Eq.(9) produces the following equations for maximization.

$$W(a)=\sum_{i=1}^{N}a_i-\frac{1}{2}\sum_{i,j=1}^{N}a_ia_jy_iy_jx_ix_j \tag{10}$$

where $a_i\geq 0$ and under constraint $\sum_{i=1}^{N}y_ia_i=0$.

Quadratic programming method can be adopted to solve the above maximization problem. If a vector $a^0=\left(a_1^0,...,a_N^0\right)$ satisfies the Eq.(10) in maximization, then the OSH expressed in terms of (w_0,b_0) may be expressed as follows:

$$w_0=\sum_{i=1}^{N}a_i^0y_ix_i \tag{11}$$

where the support vector points must comply with $a_i^0\geq 0$ and Eq.(9). When considering expansion in constraint Eq.(11), the determinant function of hyper-plane is expressed as follows:

$$f(x)=sign\left(\sum_{i=1}^{N}a_i^0y_ix_ix+b_0\right)=0 \tag{12}$$

In most cases, the data are not linearly separable, and are consequently mapped to a higher-dimensional feature space. Therefore, if the data cannot be classified clearly in the current dimensional space, then the SVM will map them to a higher dimensional space for classification.

Input data are mapped to a higher dimensional feature space by plotting a nonlinear curve. The OSH is constructed in the feature space. By constructing the feature space $\phi(x)$ can be adopted in constrained Eq.(10) as shown below:

$$W(a)=\sum_{i=1}^{N}a_i-\frac{1}{2}\sum_{i,j=1}^{N}a_ia_jy_iy_j\phi(x_i)\phi(x_j) \tag{13}$$

Given a symmetric and positive kernel function *K*(*x*,*y*), the existence of Mercer's theorem can be deduced. Therefore, $K(x,y)=\phi(x)\phi(y)$. Provided that the kernel function *K* satisfies Mercer's theorem, the derived training algorithm is guaranteed for minimization

$$W(a)=\sum_{i=1}^{N}a_i-\frac{1}{2}\sum_{i,j=1}^{N}a_ia_jy_iy_jK(x_i,x_j) \tag{14}$$

The decision function is expressed as follows:

$$f(x)=sign\left(\sum_{i=1}^{N}a_iy_iK(x_i,x_j)+b\right) \tag{15}$$

III. EEG SIGNAL CLASSIFICATION WITH FEATURE SELECTION BASED ON ONE-DIMENSION REAL VALUED PARTICLE SWARM OPTIMIZATION

3.1 Feature selection based on one-dimension real valued particle swarm optimization

The original PSO technique is designed for the real-valued optimization problems, whereas the feature selection only uses binary values to represent whether one feature is selected or not. So many researchers adopted the binary PSO (BPSO) [6] or proposed the improved binary PSO (IBPSO) [13] to do feature selection. In those works, a binary string is used for representing the feature set in which the presence of a feature is coded as "1" and the absence of a feature as "0". So the binary PSO does searching in high dimension real-valued space in which every original feature is viewed as a dimension, and converts the searched multi-dimension real-valued result into a binary string to test the performance of the searched feature subsets in each iterations. And the computing time may increase substantially as the number of original features increasing. Furthermore, the more dimension, the more complicate searching space. There will be more local minima or maxima in the searching space, and the feature selection algorithms have more chances to trap into the local optima and cannot get the optimal feature subset. In this paper, one-dimension real-valued PSO for feature selection (ODRV_PSO) is proposed, in which the search space will less complicate than that of traditional BPSO and has less chances to trap into local optimum. This approach consists of the following steps.

Step 1: The population of particles x_i is initialized, each particle i having a random position x_i^t and a random velocity v_i^t within the one-dimension real valued space, where iteration t=0.

Step 2: For every particle, x_i^t is coded into a binary string $BinaryString(x_i^t)$ in which every bit indicates a feature present or not, and its fitness is evaluated. Here, the fitness valued is determined by an SVM classifier, which is defined as following form:

$$fitness(x_i^t) = classification \quad accuracy$$
$$+ \alpha \times \frac{number\ of\ "0"\ in\ BinaryString(x_i^t)}{total\ number\ of\ original\ features} \quad (16)$$

where α is a balance factor.

Step 3: For every particles i, the best solution p_i^t until iteration t can be obtained as the Eq. (17).

$$\left\{ p_i^t = x_i^T \mid \max_{T \in \{1,2,\ldots,t\}} fitness(BinaryString(x_i^T)) \right\} \quad (17)$$

Step 4: For all population, the global best solution p_g^t until iteration t can be obtained as the Eq.(18).

$$\left\{ p_g^t = p_i^t \mid \max_i fitness(BinaryString(p_i^t)) \right\} \quad (18)$$

Step 5: Compute the velocity of each particle with Eq.(19).

$$v_i^{t+1} = v_i^t + c_1 r_1 (p_i^t - x_i^t) + c_2 r_2 (p_g^t - x_i^t) \quad (19)$$

where c_1 indicates the cognition learning factor, c_2 indicates the social learning factor, r_1 and r_2 are random number uniformly distributed in U(0,1).

Step 6: Each particle moves to the next position (or solution) according to Eq.(20).

$$x_i^{t+1} = x_i^t + v_i^t \quad (20)$$

Step 7: Stop the algorithm and output *BinaryString* (p_g^t) if termination criterion is satisfied; return to *Step 2* otherwise.

3.2 SVM classification of EEG signals with the proposed feature selection method

In this study, a novel classification method for EEG signals is proposed and can be described as follows:

Step1: The EEG recordings are decomposed into various frequency bands through a fourth-level WPD. And Db2 mother wavelet is used in the decomposition. Table 1 gives the reason why fourth-level WPD is used. From Table 1, we can find that the classification accuracy with fourth-level wavelet packet decomposition is much higher. That is to say, fourth-level wavelet packet decomposition can adequately extract the useful information or features for classification from EEG signals.

Step2: ApEn value of the coefficients at each node of the decomposition structure is computed as a feature representing the regularity or the predictability of the coefficients at that node.

Step3: ODRV_PSO algorithm searches the optimal feature subset.

Step4: Training the SVM classifier with the searched optimal feature subsets and do the prediction with the well-trained SVM classifier.

IV. Experimental results

4.1 feature selection based on one dimension real valued particle swarm optimization

Four datasets from the UCI Machine Learning Repository are used to compare the performance of the BPSO based feature selection algorithm with SVM classifier (Algorithm a) and the One-dimension real valued PSO based feature selection algorithm with SVM classifier (Algorithm b), as shown in Table 2. From Table 2, we can find that the proposed method gets higher classification accuracy and obtains smaller feature subset. These results indicate that the proposed method can get the solution that is much closer to the optimal solution, i.e. better feature subset.

Table 1.The experimental results of the classification based on different level wavelet packet decomposition and approximation entropy

	SVM training time (in second)	SVM testing time (in second)	Classification rate of normal EEG (%)	Classification rate of Epileptic EEG (%)
Two-level WPD	0.016	0.016	97.5	92
Two-level WPD	0.016	0.026	97.7	94
Four-level WPD	**0.015**	**0.021**	**98.3**	**96**
Five-level WPD	0.021	0.020	96.5	94.7

4.2 SVM classification of EEG signals with the proposed feature selection method

Five datasets containing quasi-stationary, artifact-free EEG signals both in normal subjects and epileptic patients were put in the web by Ralph Andrzejak from the Epilepsy center in Bonn, Germany. Each dataset contains 100 single channel EEG segments of 23.6 sec duration. The sampling rate of the data was 173.61Hz. A summary of the data set can be found in the reference Ref.(14).

Fourth-level WPD is applied to both normal and epileptic EEG signals. There are a total of 31 nodes in the wavelet decomposition structure. And ApEn values are computed for each node as the extracted features of EEG signals. Then dataset are partitioned into training dataset and testing set. The training dataset were used for training the SVM classifier and doing feature selection with three stochastic searching algorithms. Three stochastic searching algorithms for feature selection are genetic algorithm (GA), binary particle swarm optimization (BPSO) algorithm and one-dimension real valued particle swarm optimization (ODRV_PSO) algorithm. And testing dataset is used for testing the classification rates. Table 3 shows the classification rates of the SVM classifier with three feature selection algorithms. From Table 3, our proposed method can get 100% classification rate which is higher than other methods and select only 6.2 features which is less than other methods. This experimental result also shows that our proposed method can much better feature subset.

Table 2.Feature selection performance and classification performance comparison of two algorithms using UCI datasets

Dataset	Number of original features	Average number of reduced features with 20 runs		Average classification accuracy with 20 runs (%)	
		Algorithm a	Algorithm b	Algorithm a	Algorithm b
Breast Cancer	9	6.9±0.34	3.1±0.23	95.4±1.24	99.28±0
Segment_ation	19	6.2±0.34	3.0±0.14	92.8±2.05	95.11±0.23
Dermat_ology	34	12.6±0.61	8.0±0.36	80.89±0.19	98.30±0.38
Opt_digits	64	25.0±0.36	16.2±0.26	97.5±0.37	99.50±0.00

Table 3. Classification performance of EEG signals with three different feature selection algorithms

	Number of the selected features	Average classification rates of normal EEG with 20 runs (%)	Average classification rates of epileptic EEG with 20 runs (%)
Without feature selection	31	88.7±0.266	90.1±0.412
Feature selection with GA	11.3±0.331	95.3±0.331	98.1±0.417
Feature selection with PSO	12.6±0.473	98±0.442	96.6±0.256
Feature selection with ODRV_PSO	**6.2±0.311**	**100±0**	**100±0**

V. CONCLUSIONS

In this study, a new method for EEG signal classification is presented. In this method, wavelet packet decomposition and approximation entropy are used for do feature extraction, which will extract the nonlinear information from EEG signals sufficiently. And then an effective feature selection algorithm based on one-dimension real valued Particle Swarm Optimization is proposed and applied to do feature selection of the extracted features of EEG signals, which will reduce the number of features substantially further, and improve the accuracy and efficiency of EEG signal classification. Experimental results showed that the proposed feature selection algorithm could get more optimal feature subset and the proposed classification method could effectively select the useful features and obtain 100% classification rate of EEG signals. All of these indicated that the proposed method was much qualified for doing feature selection in many applications unlimited to EEG signal classification.

REFERENCE

[1] R.Agarwal, J.Gotman, D.Flanagan, B.Rosenblatt, “Automatic EEG analysis during long-term monitoring in the ICU”, Electroencephalography and Clinical Neurophysiology, vol.107, pp.44-58, 1998.

[2] E.D.Ubeyli, I.Guler, “Features extracted by eigenvector methods for detecting variability of EEG signals,” Pattern Recognition Letters, vol.28, pp.592-603, 2007.

[3] L.Chen, H.Hsiao, “Feature selection to diagnose a business crisis by using a real GA-based support vector machine: An empirical study,” Expert System with Applications, vol.35, pp.1145-1155, 2008.

[4] L.Chuang, H.W.Chang, C.J.Tu, C.H.Yang, “Improved binary PSO for feature selection using gene expression data, Computational Biology and Chemistry, vol.32, pp.29-38, 2008.

[5] T.M.Cover, J.M.Van Campenhout, “On the possible orderings in the measurement selection problem,” IEEE Transactions on Systems, Man and Cybernetics, vol.9, pp.657−661, 1997.

[6] A.B.Gardner, G.A.Worrell, E.Marsh, “Human and automated detection of high-frequency oscillations in clinical intracranial EEG recordings,” Clinical Neurophysiology, vol.118, pp.1134-1143, 2007.

[7] N.F.Guler, E.D.Ubeyli, I.Guler, “Recurrent neural networks employing Lyapunov exponents for EEG signals classification,” Expert System with Applications, vol.29, pp.506-514, 2005.

[8] H.Ocak, “Automatic detection of epileptic seizures in EEG using discrete wavelet transform and approximate entropy,” Expert Systems with Applications, vol.36, vol.2027-2036, 2009.

[9] W.Siedlecki, J.Sklansky, “On automatic feature selection,” International Journal of Pattern Recognition and Artificial Intelligence, vol.2, pp.197−220, 1988.

[10] A.Subasi, “EEG signal classification using wavelet feature extraction and a mixture expert model,” Expert System with Applications, vol.32, pp.1084-1093, 2007.

[11] X.Wang, J.Yang, X.Teng, W.Xia, R.Jensen, “Feature selection based on rough set and particle swarm optimization,” Pattern Recognition Letters, vol.28, pp.459-471, 2007.

[12] S.C.Yusta, “Different metaheuristic strategies to solve the feature selection problem,” Pattern Recognition Letters, vol.30, pp.525-534, 2009.

[13] H.Zhang, G.Sun, “Feature selection using tabu search method,” Pattern Recognition, vol.35, pp.701-711, 2002

[14] S.Lin, K.Ying, S.Chen, Z.Lee, “Particle swarm optimization for parameter determination and feature selection of support vector machines,” Expert System with Applications, vol.35, 1817-1824, 2008.