Analysis of Epileptic EEG Signals with Simple Random Sampling J48 Algorithm

Shuaifang Wang, Guohun Zhu, Yan Li, Peng Wen, and Bo Song

***Abstract*—This paper describes the application of a Simple Random SamplingJ48 (SRS-J48) model for classification of electroencephalogram (EEG) signals. Decision making is performed in two stages: feature extraction and classification. Eight statistical features are extracted from a two-level sample set model based on SRS technique and then classified by the J48 decision tree algorithm in Weka. The classification accuracy of the SRS-J48 is 16.35% higher than that of J48 according to the five groups of experiment with only 13% execution time on average. Besides, the proposed SRS-J48 algorithm has competitive or even better results on some of the experimental groups than Siuly's Simple Random Sampling-Least Square-Support Vector Machine (SRS-LS-SVM).**

***Index Terms*—Epilepsy, Simple Random Sampling (SRS), feature extraction, J48.**

I. Introduction

Epilepsy is a prevalent neurological disorder stemming from temporary abnormal discharges of the brain electrical activities and leading to unprovoked seizures. About 1% population in the world is diagnosed as epilepsy [1]. EEGs which record the voltage fluctuations resulting from ionic current flows within the neurons are capable of increasing insights into brain dysfunction and even of yielding information useful for diagnostic purposes [2]. Nowadays, it is widely used in the detection of epilepsy [3], [4] as well as characterization of sleep phenomena [5], encephalopathy [6] or Creutzfeldt-Jakob disease [7] and monitoring the depth of anesthesia [8] and the location of epileptic focus [9]. Automatic epileptic classification systems are the trend in both research and clinical areas because the traditional visual inspection of EEG signals requires highly trained medical professionals. Meanwhile, it is time consuming, error prone and not sufficient enough for reliable detection and prediction. Therefore, how to improve the classification accuracy of an automatic classification system should be studied.

High dimensional feature vectors with relatively few training samples tend to be a big issue in EEG signal classification. To figure out this problem, some countermeasures in both feature extraction and classification stages have been proposed so far. Adbulhamit Subasi decomposed EEG signals into the frequency sub-bands using Discrete Wavelet Transform (DWT) and classified normal and epileptic EEGs with a mixture of expert mode [10]. Güleret al extracted features using wavelet transform (WT) and the adaptive neuro-fuzzy inference system (ANFIS) trained with the back propagation gradient descent method in combination with the least squares method [11]. Toshio *et al*. employed a Gaussian mixture model to conduct EEG pattern classification [12]. Vasicek *et al*. had a test for normality based on sample entropy [13]. Kemal detected epileptic seizure in EEG signals using a hybrid system based on a decision tree classifier and fast Fourier transform (FFT) and obtained 98.72% classification accuracy [14]. Suryannarayana *et al*. introduced a most promising pattern recognition technique called cross-correlation aided SVM based classifier and achieved classification accuracy on normal and epileptic EEGs as high as 95.96% [15].

This study proposes a Simple Random Sampling J48 Algorithm (SRS-J48) to discriminate EEG signals. It extracts eight representative features from the original EEG data by SRS technique and then forwards the obtained features to a J48 classifier to gain the final classification results. To evaluate the efficiency of the proposed algorithm, the original EEG data is also classified by J48 directly. Besides, it is compared with Siuly's SRS-LS-SVM [16] in terms of accuracy as well.

This paper is organized as follows: In Section II, the experimental dataset is briefly introduced. The proposed SRS-J48method is described in Section III. In Section IV, the classification results of both the J48 classifier on original EEG data and the proposed SRS-J48 algorithm on extracted features are presented. Besides, Siuly's SRS-LS-SVM is also applied for the comparison purpose. Finally, the conclusion is drawn in Section V.

II. Experimental Data

The epileptic EEG dataset used in this paper was published by Andrzejak *et al.* [2]. The data was digitized at 173.61 samples per second obtaining from 12-bit A/D convertor. Band-pass filter setting was 0.53-40Hz. The whole dataset consists of five separate classes of EEG signals (denoted as Sets A-E), each containing 100 single-channel EEG signals from that specific class and 4097 data points in each channel. Sets A and B were recorded from five healthy volunteers with eyes opened and eyes closed, respectively. Sets C and D were recorded from the EEGs of epileptic patients during seizure-free intervals from the opposite hemisphere of the brain and within the epileptogenic zone, respectively. Set E contains the seizure activity EEGs.

Manuscript received September 9, 2013; revised November 14, 2013. This work was supported by Centre for Systems Biology.

The authors are with University of Southern Queensland, Toowoomba, 4350, Australia (email: Shuaifang.Wang@usq.edu.au, Guohun.Zhu@usq.edu.au, yan.li@usq.edu.au,peng.wen@usq.edu.au, Bo.Song@usq.edu.au).

III. METHDOLOGY

A. Related Work

The SRS-LS-SVM is a relatively high performance algorithm proposed by Siuly *et al*. in 2011 [16]. It employed SRS technique to reduce the dimensionality of the original data and a least square support vector machine for the classification of the EEG signals. The terms of SRS and LS-SVM are introduced briefly hereafter.

Simple Random Sampling (SRS) is a basic technique for probability sampling. With the SRS technique, there is an equal chance (probability) of selecting each unit from the population being studied when creating sample sets. It reduces the potential human bias in the selection of cases to be included in the sample set population. As a result, the SRS provides us with a sample set that is highly representative for the population being studied, assuming that there are limited missing data [17].

The LS-SVM algorithm was originally proposed by Suykens and Vandewalle in 2002 [18] and corresponds to a modified version of a support vector machine (SVM) [19]. The implementation details can be found in [16].

B. The Proposed Method

The proposed SRS-J48 algorithm is a combination of SRS and J48, which also extracts features by SRS technique as Siuly's SRS-LS-SVM but classifies data by J48 decision tree. The flow chart of the proposed method is shown in Fig. 1.

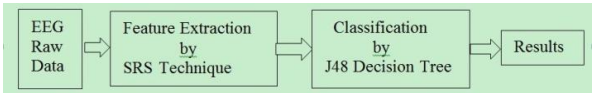


Fig. 1. The flow chart of the proposed SRS-J48 classification system.

C. Feature Extraction

Feature extraction aims at reducing the dimensionality of the original data while remaining as much useful information being included in the original vectors as possible. The implementation detail of the SRS is described in the next section. Fig. 2 depicts the block diagram of the feature extraction by the SRS technique.

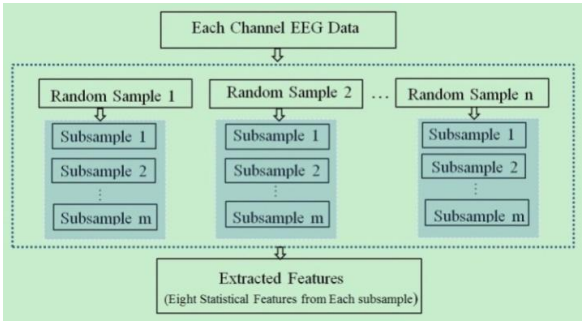


Fig. 2. The block diagram of feature extraction by SRS technique.

The experimental EEG data set consists of five sets and every set contains 100 data files holding one channel EEG data which has 4097 data points each. The use of a large number of time segments leads to high dimensionality of the feature vectors (which is [100 × 4097] for each class). Theoretically, if the number of training data is small compared to the size of the feature vectors, the classifier will most probably give poor results [20]. It is recommended to use at least five to ten times as many training samples per class as the dimensionality [21], [22]. In this paper, the SRS is used twice to gain 50 cases from each channel EEG data. In the first stage, 10 sample sets (n=10) containing 3288 points are selected from 100 channels of EEG data having 4097 points of each set by the SRS. During the second stage, five subsample sets (m=5) making up of 2746 points are chosen from 3288 points of each sample set gained from the previous stage. Finally, we extracted the following eight statistical features to reduce the dimensionality of each subsample set:

1) Minimum
2) Maximum
3) Mean
4) Median
5) First quartile
6) Third quartile
7) Inter-quartile range
8) Standard deviation

The reasons to choose the above eight statistical features as the valuable parameters to represent the high dimensional raw EEG data are as follows: the mean and the standard deviation are appropriate measures for measuring the center and variability of the data sets for a symmetric distribution case. The median and inter-quartile range are usually used to measure the center and the spread of the data for skewed distributions. When it comes to maximum and minimum, they are considered important information about a dataset in most cases.

Therefore, the dimension of 100 × 4097 in each set with one specific label (Sets A-E) has been transferred into a feature vector of size 5000 × 8.

D. Classification

During the classification stage, the extracted features are classified by J48 decision tree algorithm (Weka implementation of C4.5) which was published by Ross Quinlan in 1993 [23]. Decision tree is a classic way to represent information from a machine learning algorithm and offers a fast and powerful way to express structures in data [24]. The J48 algorithm also gives variety of options available which can make a significant difference in the quality of results. In this paper, the default settings are used because they are proven to be adequate in many cases. Weka is an open-source Java application produced by the University of Waikato in New Zealand. This software offers an interface through which many algorithms can be utilized on pre-formatted data sets. Using this interface, several test-domains were experimented to gain insight on the effectiveness of different methods.

There are no testing data provided by the above dataset. For the classification part, we split the dataset randomly with two thirds of the data being used as training data and the remaining for testing purpose. There are 100 channels of data making up of 4097 points for each class (denoted as Sets A-E) in total so that two thirds (100 × 2/3≈66) of them are used as the training data and the remaining (100-66=34) are used as the testing data. When it comes to the proposed SRS extracted data, each original channel data is transferred to 50 subsample sets (n=10 and m=5) and each subsample set has eight statistical features. The class distribution of the

sample set in the training and testing data sets of both original data and the SRS extracted features are summarized in Table I and Table II, respectively. The classifier used is J48 from Weka.

TABLE I: THE DISTRIBUTION OF THE TRAINING AND TESTING DATA SETS FROM ORIGINAL DATA

Data Class	Training Set	Testing Set	Total
A	[66 x 4097]	[34 x4097]	[100 x4097]
B	[66 x 4097]	[34 x4097]	[100 x4097]
C	[66 x 4097]	[34 x4097]	[100 x4097]
D	[66 x 4097]	[34 x4097]	[100 x4097]
E	[66 x 4097]	[34 x4097]	[100 x4097]

TABLE II: THE DISTRIBUTION OF THE TRAINING AND TESTING DATA SETS FROM SRS EXTRACTED FEATURES

Data Class	Training Set	Testing Set	Total
A	[3300 x 8]	[1700 x8]	[5000 x8]
B	[3300 x 8]	[1700 x8]	[5000 x8]
C	[3300 x 8]	[1700 x8]	[5000 x8]
D	[3300 x 8]	[1700 x8]	[5000 x8]
E	[3300 x 8]	[1700 x8]	[5000 x8]

IV. EXPERIMENTAL RESULTS

To evaluate the performance of the SRS-J48algorithm presented in Section III, java programming language is used, while the original data is also been imported to the J48 classifier for comparison reason. The experiments consist of the following five groups: 1) Set D*vs* Set E 2) SetC*vs* Set E 3) Set A*vs* Set C 4) Set B*vs* Set E and 5) Set A*vs* Set E.

Feature extraction stage is implemented by Java programing language and classification is performed using J48 decision tree in Weka version 3.7.10. All the experiments are run on a 3.00GHz Intel(R) Core(TM) 2 Duo CPU processor PC with 4.00G RAM. The operation system is Microsoft Windows 7.

A. Performance Comparison

In this section, performance comparison between SRS-J48 and J48 on the experimental EEG database is presented by Table III in terms of classification accuracy and execution time.

TABLE III: THE CLASSIFICATION ACCURACY AND EXECUTION TIME BY THE SRS-J48 AND J48

Method / Group	SRS-J48		J48	
	Accuracy	Time	Accuracy	Time
D*vs*E	94.09%	0.09s	92.65%	0.7s
C*vs*E	97.29%	0.03s	83.33%	0.6s
A*vs*D	77.85%	0.34s	61.76%	1.19s
B*vs*E	95.59%	0.03s	76.47%	0.84s
A*vs*E	100%	0.02s	85.29%	0.6s
Average	92.96%	0.102s	79.90%	0.786s

The SRS-J48 algorithm based on SRS extracted features results in a 16.35% (which is(92.96%-79.90%)/79.90%) higher accuracy, with only 13% (which is 0.102/0.786) execution time of that of the J48 algorithm based on original data. It is noted that the classification accuracy of Set A*vs*Set E is as higher as 100% due to the nature of the large differences in the data. In contrast, the analogous features result in low classification accuracy of Set A*vs* Set D. Overall, the SRS-J48 algorithm on the SRS extracted features outperforms the J48 algorithm on the same experimental dataset in terms of both efficiency and accuracy, because the high dimensional EEG data are of large size and not that representative. Feature extraction by SRS technique hits the point and turns out to be a good solution.

B. Comparing the Accuracy of the SRS-J48 and SRS-LS-SVM

In this section, theclassification accuracies on the experimental EEG database of the proposed SRS-J48 and Siuly's SRS-LS-SVM [16] are presented in Table IV.

TABLE IV: THE CLASSIFICATION ACCURACY BY THE SRS-J48 AND SRS-LS-SVM

Method / Group	SRS-J48 (Proposed Method)	SRS-LS-SVM (Siuly *et al*. [16])
D*vs*E	94.09%	94%
C*vs*E	97.29%	96.4%
A*vs*D	77.85%	88%
B*vs*E	95.59%	99.5%
A*vs*E	100%	100%

Based on the fore mentioned five groups of experiments, the SRS-J48 has the competitive results as the SRS-LS-SVM on some of the above groups, such as 1) Set D*vs* Set E 2) SetC*vs* Set E 3) Set A*vs* Set E. They adopted the same feature extraction technique but different classification algorithms, which leaded to the differences on the final results. They demonstrated that SRS is a reliable feature extraction technique for epileptic EEG signal detection and different classifiers are also a considerable factor for EEG signal classification system design.

V. CONCLUSION

EEG classification plays an important role in epilepsy detection. The proposed SRS-J48 algorithm in this study uses eight representative statistical features and the classic decision tree classifier to improve the classification performance. It transfers the high dimensional original data into less size preprocessed data by using the SRS technique, which may explain its success. The classification accuracy of the extracted features is 16.35% higher than that of original data with much less execution time (<13%). It throws light on the solution of large and high dimensional data such as EEG. Hence, the SRS-J48 algorithm has potential in the classification EEG signals.

ACKNOWLEDGMENT

Shuaifang Wang thanks her supervisor, Associate Professors Yan Li and Paul Wen, for their continuous inspiration, support, guidance, and individual feedback through the course of my Phd study. She also gratefully acknowledge the University of Southern Queensland (USQ) and Centre for Systems Biology (CSBi) supporting me to attend conferences.

REFERENCES

[1] L. Brian and J. Echauz, "Prediction of epileptic seizures," *The Lancet Neurology*, vol. 1, pp. 22-30, May 2002.

[2] G. Ralph, K. Lehnertz, F. Mormann, C. Rieke, P. David, and C. E. Elger, "Indications of nonlinear deterministic and finite-dimensional

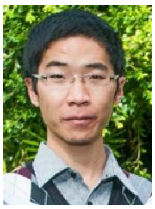
structures in time series of brain electrical activity: Dependence on recording region and brain state," *Physical Review E*, vol. 64, 061907, Nov. 2001.
[3] K. Lehnertz and C. E. Elger, "Can epileptic seizures be predicted? Evidence from nonlinear time series analyses of brain electrical activity," *Phys. Rev.Lett.*, vol. 80, pp. 5019-5023, June 1998.
[4] J. Martinerie, C. Adam, M. L. van Quyen, M. Baulac, B. Renault, and F. J. Varela, "Epileptic seizures can be anticipated by non-linear analysis," *Nature Medicine*, vol. 4, pp. 1173-1176, Oct. 1998.
[5] J. Wackermann, "Beyond mapping: Estimating complexity of multi-channel EEG recordings," *ActaNeurobiol. Exp.*, vol. 56, pp. 197-208, 1996.
[6] C. J. Stam, E. M. H. van der Leij, R.W. M. Keunen, and D. L. J. Tavy, "Nonlinear EEG changes in postanoxic encephalopathy," *Theor. Biosci.*, vol. 118, pp. 209-218, 1999.
[7] C. J. Stam, T. C. A. M. V. Woerkom, and R. W. M. Keunen, "Nonlinear analysis of the electroencephalogram in Creutzfeldt-Jakob disease," *Biol.Cybern.*, vol. 77, pp. 247-256, Oct. 1997.
[8] I. A. Rezek and S. J. Roberts, "Stochastic complexity measures for physiological signal analysis," *IEEE Trans. on Biomedical Engineering*, vol. 45, no. 9, pp. 1186-1191, Sep. 1998.
[9] G. Zhu, Y. Li, P. P. Wen, S. Wang, and M. Xi, "Epileptogenic focus detection in intracranial EEG based on delay permutation entropy," *AIP Conference Proceedings*, vol. 1559, no. 1, pp. 31-36, 2013, doi: http://dx.doi.org/10.1063/1.4824993.
[10] S. Abdulhamit, "EEGsignal classification using wavelet feature extraction and a mixture of expert model," *Expert Systems with Applications*, vol. 32, issue 4, pp. 1084-1093, May 2007.
[11] İ. Güler and E. D. Übeyli, "Adaptive neuro-fuzzy inference system for classification of EEG signals using wavelet coefficients," *Journal of Neuroscience Methods*, vol. 148, issue 2, pp. 113-121, Oct. 2005.
[12] T. Tsuji *et al*., "A recurrent log-linearized Gaussian mixture network," *Neural Networks, IEEE Trans. on Neural Networks*, vol. 14, pp. 304-316, 2003.
[13] O. Vasicek, "A test for normality based on sample entropy," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 38, no. 1, pp. 54-59, 1976.
[14] K. Polat and S. Güneş, "Classification of epileptic form EEG using a hybrid system based on decision tree classifier and fast Fourier transform," *Applied Mathematics and Computation*, vol. 187, issue 2, pp. 1017-1026, April 2007.
[15] S. Chandaka, A. Chatterjee, and S. Munshi, "Cross-correlation aided support vector machine classifier for classification of EEGsignals," *Expert Systems with Applications*, vol. 36, issue 2, part 1, pp. 1329-1336, March 2009.
[16] Y. Li, P. Wen *et al*., "Clustering technique-based least square support vector machine for EEG signal classification," *Computer Methods and Programs in Biomedicine*, vol. 104, no. 3, pp. 358-372, Dec. 2011.
[17] R. R. Frerichs, *Rapid Surveys*, 2008, ch. 3.
[18] J. A. K. Suykens and J. Vandewalle, "Least Squares support vector machine classifiers," *Neural Processing Letters*, vol. 9, issue 3, pp. 293-300, June 1999.
[19] V. Vapnik, *The Nature of Statistical Learning Theory*, New York: Springer-Verlag, 1995.
[20] F. Lotte, M. Congedo, A. Lecuyer, F. Lamarche, and B. Arnaldi, "A review of classification algorithms for EEG-based brain-computer interfaces," *Journal of Neural Engineering Journal*, vol. 4, no. 2, R1-R13, Jan. 2007.
[21] R. J. Sarunas and A. K. Jain, "Small sample size effects in statistical pattern recognition: recommendations for practitioners," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 13, no. 3, pp. 252-264, Mar. 1991.
[22] A. K. Jain and B. Chandrasekaran, "Dimensionality and sample size considerations in pattern recognition practice," *Handbook of Statistics*, vol. 2, pp. 835-855, 1982.
[23] L. Sehgal, N. Mohan, and P. S. Sandhu, "Quality prediction of function based software using decision tree approach," presented at the International Conference on Computer Engineering and Multimedia Technologies, Bangkok, Thailand, September 8-9, 2012
[24] S. Darzin and M. Montag. Decision Tree Analysis using Weka, Machine Learning–Project II. [Online]. Available: http://www.samdrazin.com/classes/een548/project2report.pdf

Shuaifang Wang received the B.S. and M.S. degrees in Computer Science from Henan University and Information Technology from the University of New South Wales, in 2011 and 2013, respectively. She is currently a Phd student of USQ in the Faculty of Health, Engineering and Sciences, University of Southern Queensland Toowoomba. Her research interests include biomedical signal analysis and data mining.

Guohun Zhu is an associate professor of Guilin University of electronic technology. He is currently a Phd student of USQ in the Faculty of Health, Engineering and Sciences, University of Southern Queensland Toowoomba. His research interests include biomedical signal analysis, networking and Computational Intelligence.

Yan Li is currently an associate professor of Computer Sciences and Deputy Associate Dean in Research, Faculty of Health, Engineering and Sciences, University of Southern Queensland. Her research interests are in the areas of Biomedical Engineering, Artificial Intelligent, Blind Signal Separation, Signal/Image Processing, Independent Component Analysis, Computer Communications and Internet Technologies, etc.

Peng Wen is currently an associate professor of Systems & Computer Control, Facultyof Health, Engineering and Sciences, University of Southern Queensland. His research interests are in the areas of Biomedical Engineering (EEG Research), Complex Medical Engineering, Networked System and Intelligent control.



Bo Song received the B.S. and M.S. degrees in Computer Science from Henan University and Information Technology from the University of New South Wales, in 2011 and 2013, respectively. He is currently a Phd student of USQ in the Faculty of Health, Engineering and Sciences, University of Southern Queensland Toowoomba. His research interests include biomedical image processing and data mining.