Unsupervised Classification of Epileptic EEG Signals with Multi Scale K-Means Algorithm

Guohun Zhu[1,2], Yan Li[1,2], Peng (Paul) Wen[1,2], Shuaifang Wang[1,2], and Ning Zhong[3]

[1] Faculty of Health, Engineering and Sciences, University of Southern Queensland, Toowoomba, QLD 4350, Australia
[2] Centre for Systems Biology, University of Southern Queensland, Toowoomba, QLD 4350, Australia
[3] Department of Life Science and Informatics, Maebashi Institute of Technology, Japan
{Guohun.Zhu,Yan.Li,Peng.Wen,Shuaifang.Wang}@usq.edu.au, zhong@maebashi-it.ac.jp

Abstract. Most epileptic EEG classification algorithms are supervised and require large training data sets, which hinders its use in real time applications. This paper proposes an unsupervised multi-scale K-means (MSK-means) algorithm to distinguish epileptic EEG signals from normal EEGs. The random initialization of the K-means algorithm can lead to wrong clusters. Based on the characteristics of EEGs, the MSK-means algorithm initializes the coarse-scale centroid of a cluster with a suitable scale factor. In this paper, the MSK-means algorithm is proved theoretically being superior to the K-means algorithm on efficiency. In addition, three classifiers: the K-means, MSK-means and support vector machine (SVM), are used to discriminate epileptic EEGs from normal EEGs using six features extracted by the sample entropy technique. The experimental results demonstrate that the MSK-means algorithm achieves 7% higher accuracy with 88% less execution time than that of K-means, and 6% higher accuracy with 97% less execution time than that of the SVM.

Keywords: K-means clustering, multi-scale K-means, scale factor.

1 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 are diagnosed as epilepsy [1]. Fortunately, EEG recordings can show the brain electrical activity information and provide valuable insight into disorders of the brain. EEG signals are considered as important data in diagnosing epilepsy and predicting epilepsy seizures. However, the traditional visual inspection by analysts is time consuming, error prone and not sufficient enough for reliable detection and prediction. The randomization nature of epilepsy seizures and their large EEG recording datasets make epileptic EEG classification more difficult. Hence, an automatic epileptic classification system is becoming more and more on demand.

K. Imamura et al. (Eds.): BHI 2013, LNAI 8211, pp. 158–167, 2013.

Most of traditional automatic epileptic classification systems use supervised learning classifiers, such as artificial neural networks (ANN), support vector machines (SVMs) and decision trees. Acharya et al. (Acharya et al., 2012) fed four entropy features to a fuzzy classifier to identify normal, ictal and inter-ictal EEGs. Chua et al. (Chua et al., 2011) employed a Gaussian mixture model and a SVM to identify the epileptic EEGs. Guo et al. (Guo et al., 2011) applied wavelet discrete transform features and an ANN for discriminating ictal EEGs from normal EEGs. Siuly et al. (Siuly et al., 2011) proposed a clustering technique to classify ictal and healthy EEGs. Song and Lio (Song and Liò, 2010) classified ictal, inter-ictal and normal EEGs by features based on sample entropy (SE) and an extreme learning machine algorithm. Zhu et al. (Zhu et al., 2012) implemented visibility graph (VG) based features and a SVM classifier to identify ictal EEGs from healthy EEGs. However, an automatic epileptic classification system normally requires large sets of data to train a classifier, and to improve the accuracy. Meanwhile, all the data are normally required in a specific format and meet certain conditions, such as the number of data segments/epochs should be the same in the training data and testing data. Besides, the target categories for all the data segments in the training set rely on the labels obtained manually by experts. All these limitations impede the current supervised epileptic EEG classification techniques from being used.

K-means clustering is a popular unsupervised learning method which was first presented by MacQueen (Macqueen, 1967). It consists of two simple steps: the first step is to randomly choose k centroids for k clusters. The second step is to separate the input data into k disjoint clusters according to the distance between each data instance and the k chosen centroids. Its simplicity and fast computation clustering make it easy to implement. However, if some data points belonging to the same cluster are incorrectly assigned into other disjoint clusters during the first step, it may lead to wrong classification results. Recently, Vattani (Vattani, 2011) showed that the running time of the K-means algorithm increases exponentially when the data size increases. To solve the cluster initialization issue, Arthur and Vassilvitskii (Arthur and Vassilvitskii, 2007) proposed a K-means++ algorithm and improved the classification accuracy by initializing centroids one by one. Bahmani et al. (Bahmani et al., 2012) reported that the K-means++ did not work well on large sets of data because it relies too much on the central point initialization.

This study proposes a multi-scale K-means (MSK-means) algorithm to discriminate epileptic EEGs from healthy EEGs. It combines several continuous EEG data points as a scale central area to make the centroid choice more robust than that of the K-means algorithm. The calculation of the distance in the second step can also be expanded to multi scales. The proposed method improves its efficiency by decreasing its running iterations.

The paper is organized as follows: In Section 2, the experimental data set is introduced. The traditional K-means algorithm and the proposed MSK-means method are described in Section 3. In Section 4, the comparison results of the K-means, MSK-means and SVM with the same EEG features are presented. Finally, conclusions are drawn in Section 5.

2 Experimental Data

This paper uses the epileptic EEG data set which was described by Andrzejak et al (Andrzejak et al., 2001). 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 database is made up of five EEG data sets (denoted as sets A-E), each containing 100 single-channel EEG signals from five separate classes and 4097 data points. 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.

3 Methodology

The proposed epileptic classification system is shown in Fig.1. The features based on sample entropy extracted from the raw EEG data are directly transferred to a MSK-means classifier for the classification. The K-means clustering algorithm and the SVM classifier in Fig. 1 are for comparison purpose.

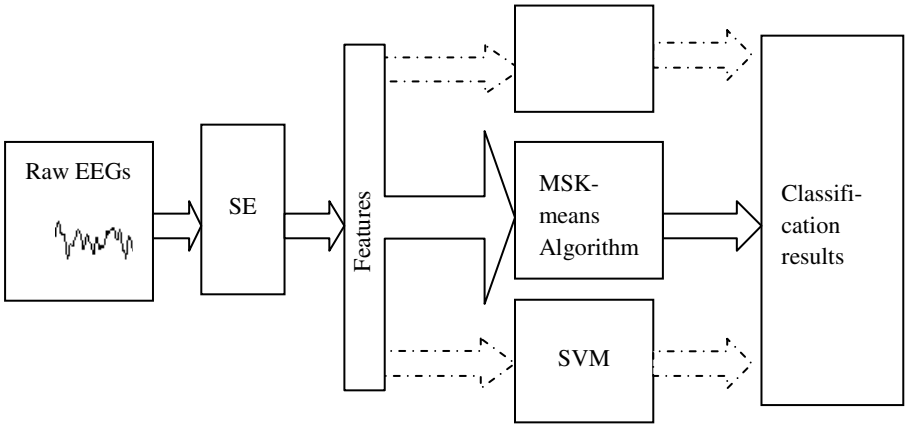

Fig. 1. The structure of the proposed epileptic EEGs classification system

3.1 K-Means Algorithm and K-Means++ Algorithm

Given a set of observations $X = \{x_i \mid i = 1,2,\ldots,n\}$, the K-means clustering technique aims to partition n observations into k sets ($k \leq n$) $C = \{c_j \mid j = 1,2,\ldots,k\}$ based on the Euclidean distance. The Euclidean distance between the i^{th} data point and the j^{th} centroid is defined as follows:

$$d(x_i, c_j) = \sqrt{\sum_{j=1}^{k} (x_i - c_j)^2} \tag{1}$$

The central point of a cluster is recomputed as:

$$C_j = \frac{1}{|C_j|} \sum_{x \in C_j} x \tag{2}$$

The K-means algorithm minimizes the within-cluster sum of squares by Lloyd iteration to make the data to the same cluster more compact and dependent:

$$\varphi = \sum_{j=1}^{k} \sum_{i=1}^{|c_j|} d(x_j, c_i) \tag{3}$$

The main idea of the K-means algorithm is to randomly choose *k* observations as the cluster central points (centroids) and assign all the remaining data to their nearest centroids based on equation (1). Then the new centroid of each cluster is calculated using equation (2). The algorithm converges when the new centroids are as same as the old centroids. The randomness of initialization is error prone if some data points from the same class are assigned to different cluster centroids. The k-mean++ algorithm proposed by Arthur and Vassilvitskii [9] improves the initialization by the following algorithm:

Algorithm1. K-means++ init

```
Input: X, k
n• number of X
C• randomly choose a point from X
While |C|<k {
      Dist [1...n] •the distance between X and C
      U•sum(Dist[1...n])
      j•1
      Do {U=U-Dist[j], j•j+1} while U>0
      C•C union X[j]
     }
end.
```

The K-means++ has an additional computation time for initializing centroids. However, the time complexity of both K-means and K-means++ algorithms are $O(ndk)$ (Arthur and Vassilvitskii, 2007). Where *n* is the number of the given observations; *k* is the number of clusters; and *d* is the time of iterations, respectively.

3.2 Multi Scale K-Means (MSK-Means) Algorithm

The scale of initialization of both K-means and K-means++ is small and limited to the data size, which is not suitable for large sizes of EEG signals. In this paper, a MSK-means algorithm is proposed to improve the performance by optimizing the cluster initialization.

The concept of multi scale analysis of time series was first proposed by Costa (Costa et al., 2002). The multi scale technique transfers one dimensional time serial $X= \{x_i \mid i = 1,2,\cdots,n\}$ into another time serial $Y=\{y_t \mid t = 1,2,\cdots,n/\tau\}$ with a different scale. Here τ is the scale factor. The transformation formula is as follows:

$$y = \frac{1}{\tau} \sum_{i=(j-1)\tau+1}^{js} x_i, \quad 1 \le j \le \frac{n}{\tau} \tag{4}$$

Based on equation (4), the original algorithm is adjusted as:

Algorithm 2. MSK-means init

```
Input: X, k,
Y•construct according to equation (4) and
Med•K median positions of (Y)
C• empty set
i•1
While (i<k)
  C[i] • random a point between Med[i] and Med [i+1]
end.
```

Similar to the K-means++ algorithm, the MSK-means algorithm only improves the initialization part of the K-means algorithm. Lloyd repeat is conducted with the scaled time serious *Y* instead of the original times series *X*. The computational complexity of the MSK-means algorithm is as follows.

Theorem (1). Let us assume that *n* is the number of the data sets, *d* is the time of iterations, *k* is the number of clusters and τ is a parameter, the time complexity of the MSK-means algorithm is $O\left(\max\{ndk/\tau, n\}\right)$.

Proof: In the MSK-means algorithm, the time complexity of equation (4) is *n*. It indicates that the complexity of *k* median value is n/τ. The time complexity of Lioyd repeat is $O(ndk/\tau)$. The time complexity of the MSK-means algorithm is $O\left(\max\{ndk/\tau, n\}\right)$.

According to Theorem (1), the time complexity of the MSK-means algorithm can be linear when τ is large enough, which means it can be of higher efficiency than both the K-means and K-means++ algorithms. The relation of τ and the time complexity of the multi-scale means algorithm is discussed in Section 4.

3.3 Sample Entropy Algorithm

Entropy is often used to measure the complexity of a time series. It has been widely applied in EEG signal processing. Bai et al. (Bai et al., 2007) used approximate entropy (AE) and sample entropy (SE) to analyze epileptic EEG signals and found that SE is more suitable for identifying epileptic seizures than AE. This study also adopts SE features to represent the raw EEG signals. The SE algorithm has three input parameters: (1) *m:* the embedded dimension, (2) *r:* the similarity criterion, (3) *n:* the length of a time series. In this experiment, *m* is assigned as *1, 2, 3* in combination with *r=0.15* and *r=0.2*, respectively. Therefore, six combined SE features from each epoch of EEG signals are extracted in this study. The sample entropy algorithm used in this paper is available from PhysioNet and PhysioToolkit website (http://www.physionet.org/physiotools/sampen/c/).

3.4 Support Vector Machine

To compare the performance of the unsupervised MSK-means algorithm with the supervised classifiers, the support vector machine (SVM) is selected to conduct the binary classification. The SVM has been successfully used in epileptic EEG classification (Siuly et al., 2011, Nicolaou and Georgiou, 2012). It can perform both the linear space discrimination and nonlinear classification by choosing different "kernel" functions which can be linear, polynomial kernel, radical basis function (RBF) and sigmoid. In this paper, the SVM algorithm with RBF kernel is implemented in R package *e1071* (Karatzoglou et al., 2006).

4 Experimental Results

To evaluate the performance of the MSK-means algorithm presented in Section 3, C programming language is used, while the SVM and K-means algorithms are implemented by R package e10171 and stats package, respectively. The experiments include three parts: (1) evaluating the scale factor τ and classification accuracy based on different lengths of a time series; (2) comparing the performances of the K-means, SVM and MS K-means algorithms for classifying seizure EEGs and healthy EEGs with eye closed; (3) comparing the computational speed and the accuracy level of the K-means, SVM and MSK-means for classifying epileptic EEGs and healthy EEGs on five groups. For experiments (1) and (2), every EEG recording is separated into 23, 8, 4, and 2 equal epochs, thus two groups of EEG data can generate 4600, 1600, 800, and 400 non-overlapping signal segments with a new length. Six SE features are extracted from each new epoch. For experiment (3), each EEG recording is divided into four equal epochs, and a total of 4000 new EEG segments are produced. During the SVM classification processing, the extracted features of odd EEG segments are used in the training data set while those of even epochs are used in the testing data set.

4.1 Evaluating the Classification Accuracy of the MSK-Means Algorithm with Different Values of Scale Factor τ and Segment Size

This section is to evaluate the impact of different values of scale factor τ and the number of epochs/segments on the performance of distinguishing normal EEGs from seizure EEGs with the SE features. Firstly, two groups of EEG data, sets A and E, are selected. Each recording is separated into 173, 512, 1024, 2048 and 4096 data points per segment. There is a total of 2300, 800, 400, 200 and 100 epochs in each EEG data set. Lastly, all these data are fed into the MSK-means classifier with the scale factor value as 46, 16, 8, 4 and 1 to conduct the classification, respectively.

Table 1 compares the results of the MSK-means algorithm on set A and set E when the values of scale factor τ and dataset are different.

Table 1. The execution time and accuracy of the MSK-means algorithm with different τ values for set A vs. set E

n \ τ	τ =1		τ =4		τ =8		τ =16		τ =46	
	Accuracy	**d**	Accuracy	**d**	Accuracy	**d**	Accuracy	**d**	Accuracy	**d**
4600	93.9%	9	95.9%	7	97.6%	4	97.9%	4	100%	3
1600	94.7%	8	96.0%	7	95.0%	5	100%	4	97.1%	2
800	94.3%	7	95.0%	5	100%	4	100%	2	94.0%	2
400	95.0%	8	100%	4	100%	3	100%	3	87.5%	4

* **d** is the number of Lioyd iterations.

From Table 1, when the accuracy is 100%, the values of (n, τ) pair are (4600, 46), (1600, 16), (800, 8), (800, 16), (400, 4), (400, 8) and (400, 16). From those values, it is concluded that the performance of the MSK-means algorithm is better when $100 \le n / \tau < 200$.

4.2 Comparing Speed and Accuracy of K-Means, SVM and MSK-Means Algorithms with Different Numbers of Epochs and Scale Factor τ

In this section, we use the K-means, MSK-means and SVM algorithms to discriminate seizure EEGs (set E) from healthy EEGs with eyes closed (set B). The results are demonstrated in Table 2. Where KM indicates the K-means algorithm; MSKM denotes the MSK-means algorithm. In order to obtain good performances, the values of scale factor τ are selected as 46,16,8 and 4, making $n / \tau = 100$.

Table 1 and Table 2 show that the accuracy of the MSK-means algorithm remains as 99% or 100% using a suitable scale factor value τ when the dataset is large, while both K-means and SVM classifiers have a lower accuracy with a longer execution time when the size of the data increases.

Table 2. The **execution** time and accuracy comparisons of the three algorithms with different segment length *n* and scale factor τ for set B and set E

n / A	4600 (τ=46)		1600(τ =16)		800 (τ=8)		400(τ=4)	
	Accuracy	Time (ms)	Accuracy	Time (ms)	Accuracy	Time (ms)	Accuracy	Time (ms)
KM	93.2%	70	94.4%	30	94.8%	15	95.0%	5
MSKM	100%	8	99.0%	6	99.0%	5	99.0%	5
SVM	93.9%	260	94.9%	40	96.5%	20	96.0%	10

4.3 Comparing the Speed and Accuracy of the K-Means, SVM and MSK-Means for Different Pairs of Data Sets

In this section, the performance comparison of the K-means, SVM and MSK-means algorithms for different pairs of data sets are presented. Four same size of data sets containing 1024 epochs and the six SE extracted features from each epoch are used. The results are showed in Table 3.

Table 3. The classification accuracy and execution time of the three algorithms with different pairs of data sets (τ =10, n=1024).

Data groups (Set)	SVM		K-means		MSK-means	
	Accuracy	Time (ms)	Accuracy	Time (ms)	Accuracy	Time (ms)
A vs E	98.0%	20	94.9%	10	100%	10
A vs C	92.3%	25	89.8%	10	95.0%	10
A vs D	93.5%	20	85.6%	10	96.0%	10
A vs B	82.3%	20	52.6%	20	74.0%	10
(A, B) vs E	99.0%	30	95.7%	25	100%	10
(A, B) vs (C, D, E)	98.6%	50	88.6%	30	98.0%	15

From the above table, the classification accuracy for the pair of (A, B) vs. (C, D, E) is 98% using the MSK-means algorithm. However, it can be further improved by changing the values of τ and n. (e.g. it achieves *99.1%* when $\tau = 16$ and n=*2000*.)

The classification accuracies on the epileptic EEG database from different literature are presented in Table 4. Based on Tables 2, 3 and 4, the proposed MSK-means method has better performance in distinguishing the epileptic EEGs from healthy EEGs, especially in identifying the epileptic EEGs from normal EEGs with eye closed. Without clinical history data records, it is impossible for a supervised algorithm to conduct classifications, while the MSK-means algorithm can work well because it is unsupervised.

Table 4. The classification accuracy by the MSK-means and other existing methods

Researchers	Features, (Epochs length) & classifiers	Data sets (Set)	Accuracy
Polat and Güneş (Polat and Güneş, 2007)	PSD, (*n=256*) & Decision tree	A, E	98.72%
Guo et al. (Guo et al., 2011)	DWT, (*n=4097*) & ANN	A, E	99.6%
Siuly et al. (Siuly et al., 2011)	Clustering, (*n=4096*) & SVM	A, E	99.9%
		B, E	96.3%
Zhu et al. (Zhu et al., 2012)	Visibility graph, (*n=4097*) & SVM	A, E	100%
Srinivasan et al.	ApEn, (*n=1024*) & ANN	A, E	100%
Xie and Krishna (Xie and Krishnan, 2013)	Wavelet-based sparse functional linear model, (*n=1024*) & SVM	A, E	100%
		(A,B), (C,D,E)	79.34%
*Orhan et al.(Orhan et al., 2011)	DWT with K-means clustering, (*n=4097*) & ANN	A, E	100%
		(A,B), (C,D,E)	98.8%
This work	**Sample entropy, (*n=1024*) & Multi-scale K-means clustering**	**A, E**	**100%**
		B, E	**99.0%**
		(A, B), (C, D, E)	**99.1%**

*The K-means algorithm was used by Orhan et al. (Orhan et al., 2011) as a feature extraction method instead of a classifier.

5 Conclusion

Unsupervised classification algorithms play an important role in epilepsy detection. The proposed MSK-means algorithm in this study optimizes the initialization stage to improve the classification performance. Both theory and experimental results show that the complexity of the MSK-means algorithm is less than that of the K-means. This study also demonstrates that the MSK-means algorithm improves the classification accuracy by 7% than the K-means when scale factor $\tau=46$, and has 6% higher accuracy with 97% less execution time than the SVM classifier using the half of the data as the training set. Hence, the MSK-means algorithm can be used efficiently for time series analysis and EEG classification.

References

1. Acharya, U.R., Molinari, F., Sree, S.V., Chattopadhyay, S., Ng, K.-H., Suri, J.S.: Automated diagnosis of epileptic EEG using entropies. Biomedical Signal Processing and Control 7, 401–408 (2012)
2. Andrzejak, R.G., Lehnertz, K., Mormann, F., Rieke, C., David, P., Elger, C.E.: 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. 64, 061907 (2001)

3. Arthur, D., Vassilvitskii, S.: k-means++: the advantages of careful seeding. In: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 1027–1035. Society for Industrial and Applied Mathematics, New Orleans (2007)
4. Bahmani, B., Moseley, B., Vattani, A., Kumar, R., Vassilvitskii, S.: Scalable k-means++. Proc. VLDB Endow. 5, 622–633 (2012)
5. Bai, D., Qiu, T., Li, X.: The sample entropy and its application in EEG based epilepsy detection. Journal of Biomedical Engineering 24, 200–205 (2007)
6. Chua, K., Chandran, V., Acharya, U.R., Lim, C.M.: Application of Higher Order Spectra to Identify Epileptic EEG. Journal of Medical Systems 35, 1563–1571 (2011)
7. Costa, M., Goldberger, A.L., Peng, C.K.: Multiscale Entropy Analysis of Complex Physiologic Time Series. Physical Review Letters 89, 068102 (2002)
8. Guo, L., Rivero, D., Dorado, J., Munteanu, C.R., Pazos, A.: Automatic feature extraction using genetic programming: An application to epileptic EEG classification. Expert Systems with Applications 38, 10425–10436 (2011)
9. Karatzoglou, A., Meyer, D., Hornik, K.: Support Vector Machines in R. Journal of Statistical Software 15, 1–28 (2006)
10. Macqueen, J.B.: Some methods of classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297 (1967)
11. Nicolaou, N., Georgiou, J.: Detection of epileptic electroencephalogram based on Permutation Entropy and Support Vector Machines. Expert Systems with Applications 39, 202–209 (2012)
12. Orhan, U., Hekim, M., Ozer, M.: EEG signals classification using the K-means clustering and a multilayer perceptron neural network model. Expert Systems with Applications 38, 13475–13481 (2011)
13. Polat, K., Güneş, S.: Classification of epileptiform EEG using a hybrid system based on decision tree classifier and fast Fourier transform. Applied Mathematics and Computation 187, 1017–1026 (2007)
14. Siuly, L.Y., Wen, P.: Clustering technique-based least square support vector machine for EEG signal classification. Computer Methods and Programs in Biomedicine 104, 358–372 (2011)
15. Song, Y., Liò, P.: A new approach for epileptic seizure detection: sample entropy based feature extraction and extreme learning machine. Journal of Biomedical Science and Engineering 3, 556–567 (2010)
16. Vattani, A.: k-means Requires Exponentially Many Iterations Even in the Plane. Discrete & Computational Geometry 45, 596–616 (2011)
17. Xie, S., Krishnan, S.: Wavelet-based sparse functional linear model with applications to EEGs seizure detection and epilepsy diagnosis. Medical & Biological Engineering & Computing 51, 49–60 (2013)
18. Zhu, G., Li, Y., Wen, P.: Analysing epileptic EEGs with a visibility graph algorithm. In: 2012 5th International Conference on Biomedical Engineering and Informatics (BMEI), pp. 432–436 (2012)