Fast Clustering Algorithm for Large ECG Data Sets Based on CS theory in Combination with PCA and K-NN Methods

Mohammadreza Balouchestani, *Member, IEEE* and Sridhar Krishnan, *Senior Member, IEEE*

Abstract*-- Long-term recording of Electrocardiogram (ECG) signals plays an important role in health care systems for diagnostic and treatment purposes of heart diseases. Clustering and classification of collecting data are essential parts for detecting concealed information of P-QRS-T waves in the long-term ECG recording. Currently used algorithms do have their share of drawbacks: 1) clustering and classification cannot be done in real time; 2) they suffer from huge energy consumption and load of sampling. These drawbacks motivated us in developing novel optimized clustering algorithm which could easily scan large ECG datasets for establishing low power long-term ECG recording. In this paper, we present an advanced K-means clustering algorithm based on Compressed Sensing (CS) theory as a random sampling procedure. Then, two dimensionality reduction methods: Principal Component Analysis (PCA) and Linear Correlation Coefficient (LCC) followed by sorting the data using the K-Nearest Neighbours (K-NN) and Probabilistic Neural Network (PNN) classifiers are applied to the proposed algorithm. We show our algorithm based on PCA features in combination with K-NN classifier shows better performance than other methods. The proposed algorithm out-performs existing algorithms by increasing 11% classification accuracy. In addition, the proposed algorithm illustrates classification accuracy for K-NN and PNN classifiers, and a Receiver Operating Characteristics (ROC) area of 99.98%, 99.83%, and 99.75% respectively.

Keywords -ECG, Sensitivity, Accuracy, Clustering, Classification

I. INTRODUCTION

Through enabling continuous and wireless long-term monitoring of ECG signals, they have the potential to achieve improved personalization and quality of care, increased ability of prevention and early diagnosis, and enhanced patient mobility and safety [1]. Long-term ECG monitoring shows redundancy between adjacent beats due to its quasi-periodic structure which implies a high fraction of common support between consecutive beats [2, 3]. The high fraction of common support demonstrates that long-term ECG monitoring is suitable for distributed CS theory. In this work, we present a novel clustering algorithm based on Compressed Sensing (CS) theory for large ECG datasets. The proposed algorithm is applied to five types of beats as: 1) Normal Left Bundle Branch Block (LBBB); 2) Right Bundle Branch Block (RBBB); 3) Atrial Premature Contraction (APC); 4) Ventricular Premature Contraction (VPC); 5) Paced beats. The proposed algorithm is tested by the MIT-BIH arrhythmia database [4, 5] which includes 48 signals with duration from 14 to 24 hours and total of 668,486 heartbeats including 7707 ventricular ectopic beats, 90,580 non-ectopic, 2973 supra-ventricular-ectopic, 1784 fusion, and 7050 unknown and paces beats. In this study, firstly, two dimensionality reduction approaches are applied to the input ECG dataset. Secondly, two classifier methods are applied independently to output of the proposed algorithm in order to extract the final classification results. The proposed algorithm has the following advantages: 1) Real-time clustering and classification; 2) provide a good starting point in designing low power wireless ECG systems which are commercially viable and could be easily fitted into current healthcare facilities, Electronic Health (EH) and Mobile Health (MH). We show our algorithm shows better performance based on PCA and K-NN methods than other methods. The structure of this paper is organized as follows: In Section II, a quick overview about the available methods in the simulation results is provided. Section III, presents our main contribution for establishing the state-of-the-art clustering algorithm based on CS theory and in Section IV, the simulation results are demonstrated. Conclusion is given in Section V and future works in Section VI.

The authors are with the Department of Electrical and Computer Engineering, Ryerson University, Toronto, Canada, M5B2K3 (e-mails:mbalouch@ee.ryerson.ca, Krishnan@ee.ryerson.ca, Tel: 416975000).

II . OVERVIEW OF EXISTING METHODS

The following methods are applied to the proposed algorithm to provide classification features: 1) K-Means clustering algorithm aims to classify or to group out objects based on attributes or features into number of groups [6, 7]. The grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid; 2) K-NN method is a non-parametric learning algorithm and a statistics-based tool for classification [8, 9]. The similarity between two features is measured by the Euclidean distance between them, where decreasing distance indicates increasing similarity; 3) PNN classifier is an implementation of a statistical algorithm called kernel discriminant analysis [10, 11]. This approach maps any input pattern to a number of classifications and can be forced into a more general function approximation; 4) PCA is a linear dimensionality reduction method that reduces the dimensions of the feature coefficients [12, 13]. In addition, PCA is a reduction dimensionality approach to convert a set of observations in a dataset of interdependent with a large number of variables into a set of values of linearly uncorrelated parameters or Principal Components (PCs); 5) LCC as a dimensionality reduction method aims local normalization in Euclidean length [14, 15].

III. ADVANCED K-MEANS ALGORITHM

CS theory is a mathematical framework in acquiring and recovering sparse signals with the help of an incoherent

projecting basis that provides insight into how a high resolution dataset can be inferred from a relatively small and random number of measurements using simple random linear process [16, 17]. Thus, rather than measuring each sample and then computing a compressed representation, CS suggests that we can measure a compressed representation directly [18]. For example, in our work a group of 10-clusters of one recode could potentially be used to reconstruct a group of 100-clusters by projecting the desired high resolution clusters onto a set of low resolution of random measurements and then recovering a group of 100-clusters through sparse signal reconstruction by solving an optimization problem. In this section, we present our contribution in establishing the state-of-the-art in clustering algorithm based on CS theory. The main purpose of K-Means algorithm based on CS theory is to determine the optimal and minimal number of clusters. The compressed feature vector in $\mathbb{R}^M$ is obtained of original vector in $\mathbb{R}^N$ with $M \ll N$ and can be expressed as [18]:

$$[\mathrm{Y}_i]_{M\times 1} = [\Phi]_{M\times N}[X_i]_{N\times 1} = [\Phi]_{M\times N}[\Psi]_{N\times N}[\mathrm{S}]_{N\times 1}, \quad (1)$$

where Φ, Ψ, and S are random measurements matrix in $\mathbb{R}^M$, orthogonal basis, and coefficients matrix in $\mathbb{R}^N$ respectively. The reconstruction process of feature vector is done by solving a $\ell_1 -$ minimization least-square problem as [19]:

$$\widehat{S} = \arg\min\{\left\|\Phi X_i - \Phi\Psi_{X\,i} S\right\|_{\ell_2}^2 + \beta\left\|S\right\|_{\ell_1}\}. \quad (2)$$

where β is defined as a minimization constant. The process of updating random sensing matrix Φ has a straight-forward solution, as it minimizes to find a rank-one approximation to the matrix of E_K as:

$$E_K = Y - \sum_{j\neq k} \Phi_j X^j, \quad (3)$$

where X^j is the j^{th} row in the coefficient matrix. Therefore, the goal of the proposed algorithm is to find an optimal set of sparsity bases for each data point such that the sum of their reconstruction errors is minimized. The minimization rule to check the reconstruction error is defined as [19]:

$$E^* = \min\ \min \sum_{j=1}^{K}\sum_{c\in C}\left\|X_i - V_j(X_i)\right\|^2, \quad (4)$$

where C and V are cluster membership and reconstruction vector. This rule minimizes all possible zero coefficients of S for all data points of ECG dataset into K cluster. In addition, our algorithm consists of an iterative procedure to find out the minimum number of clusters. The minimization rule in (4) depends on two parameters, we fix value of the first parameter and minimize over the second value and then fixing value of the second while minimizing the first to converge when the minimization rule would be very small. The new value for C is found as:

$$C = \sum_{j=1}^{K}\sum_{i\in\hat{c}(X_i)=j}\left\|X_i - \nu_j(X_i)\right\|^2, \quad (5)$$

After applying (5) to all datasets, we are able to find out minimum value in (4). The proposed algorithm is established on the following steps: 1) For a given Φ, the compressed feature vector is generated after checking the reconstruction error ; 2) For a given value of K, the input dataset partitions into K clusters; 3) Iterate over all data vectors to determine clusters based on nearest subspace according to (5) ; 4) Compute the data vector's contribution to the total residual by re-assignment of the input data vectors to their found clusters; 5) The algorithm is repeated by alternate applications of Steps 2 and 3 until convergence. In each re-assignment step, the new clusters become consistent and in ideal case of non-overlapping clusters, the data vector that belongs to one selected cluster gets more contribution to a given point's residual. Thus, the other coefficients of data vector not belonging to this cluster are minimized in (4). Following is the pseudo-code of the proposed algorithm.

Inputs: $X = \{X_i : X_i \in \mathbb{R}^N\}$, K : the number of clusters, M : the number of random linear measurements and : Random measurements matrix $\Phi^{(j)} \in \mathbb{R}^{M\times N}$ with $j = 1$

Output: Optimal number of clusters

1: Minimize $\{\left\|Y_i - \Phi X\right\|_2^2\}$ subject to $\left\|X\right\|_0 \le \ell$

2: Compute E_K

3: While not satisfied $E_K \le E^*$

4: Repeat until convergence

5:end while

6: $j = j + 1$

7: Gererate compressed vector $Y = \{Y_i : Y_i \in \mathbb{R}^M\}$

8: Initialize cluster $c^{(0)}$ and calculate total residual $E^{(0)}$

9: Select $E_{\min} = E^{(0)}$ with $i = 1$

10:while not converged Euclidean distance d_0

11: Based on re-assignment process, determine the best cluster for each data vector

12: Generate $c^{(i)}$ based on base clusters

13: Calclute total residual $E^{(i)}$ based on nearest subspace of data vectors

14: If $c^{(i)} \equiv c^{(i-1)}$ end If

15: Set $i = i + 1$

16:end while

17: $c^{(i-1)} \to \hat{c}$

While the existing K-Means approach applies mean calculations to evaluate the final clustering, the proposed algorithm generates the update dictionary by (5) in order to provide the optimal and minimum number of clusters.

IV. SIMULATION RESULTS

Our algorithms are applied to long-term ECG dataset in the MIT-BIH database, which contains six two-channel ECG signals sampled at 128 Hz per channel with 12 bit resolution, and one three-channel ECG signal sampled at 128 Hz per

channel with 10-bit resolution. The duration of the 48 signals changes from 14 to 24 hours with total of 668,486 heartbeats. The 300 samples around the R-peak (149 samples before QRS peak and 150 samples after QRS peak and the QRS peak itself) is considered as a signal beat and as a window length which consists approximately one cycle of cardiac activity. The corresponding beats from insufficient samples from the first and last detected QRS complex are neglected. The ECG dataset after dimensionality reduction step, are applied to the proposed algorithm. Then, the output of the proposed algorithm is fed to K-NN and PNN as two important classifiers. Now, we test the proposed algorithm to two different ECG datasets which size 100 by 25 including 100 heartbeats sampled at 25 features. We select the data features to 1st, 2nd, and 3rd principal components which include more the 99.28% of the dataset variance. Figure 1 illustrates two datasets in a 3D plot using the existing K-Means algorithm in combination with PCA and K-NN methods. Figure 2 demonstrates two datasets in a 3D plot with the proposed K-Means. It can be seen from the Fig. 2 that the classification of the heartbeats in the selected datasets is clearly improved.

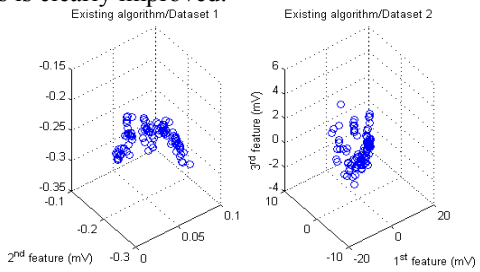

Fig.1: Selected datasets with the existing algorithm

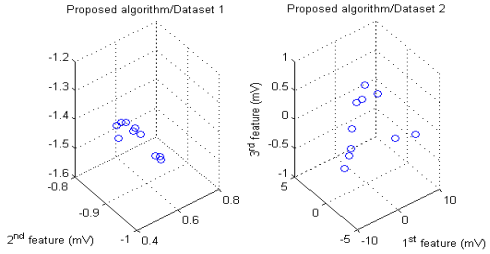

Fig.2: Selected datasets with proposed algorithm

The proposed algorithm confirms that the number of heartbeats can be clustered into a much smaller number of classes that is enough for reconstruction process. Table 1 shows the overall results for sensitivity, specificity, accuracy, and Positive Predictive value (PPV).

Table 1: Comparing between the proposed and existing algorithms

Algorithm	Sensitivity	Specificity	Accuracy	PPV
Proposed and PCA/K-NN	99.92	98.22	99.98	99.02
Existing and PCA/K-NN	87.23	86.22	88.97	86.12
Proposed and LCC/PNN	86.24	87.26	95.06	87.05
Existing and LCC/PNN	74.22	75.55	84.23	76.02

Overall, the proposed algorithm shows over 99.98% of the time within the entire benchmark dataset. Figure 3 compares the excising and proposed algorithms versus of accuracy.

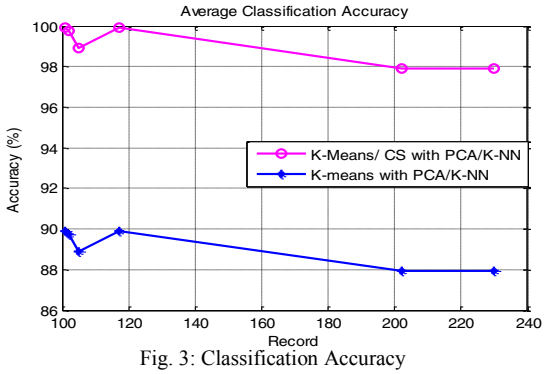

Fig. 3: Classification Accuracy

It can be observed from Fig.3 that the existing clustering algorithm with PCA/K-NN provides less accuracy, and the proposed algorithm with PCA/K-NN provides highest accuracy. This ability allows increasing accuracy to 99.98% which means an increment of 11% for accuracy level comparing to the existing algorithm. ROC area is an important parameter to validate the quality of a classifier approach. ROC is a graphical curve which creates by plotting the true positives rate versus the false positive rate. Classification accuracy is measured by the area under the ROC curve. An area between 98 to 100 % represents perfect accuracy for classification approach. Table 2 compares ROC area between the existing K- Means algorithm and the proposed algorithm for two different types for reduction dimensionality and classification methods.

Table 2: ROC Area

Clustering	Proposed Algorithm		Existing Algorithm	
Reduction/Classifier	PCA/K-NN	PCA/PNN	PCA/K-NN	PCA/PNN
ROC Area (%)	99.98	99.83	95.98	94.22

Figure 3 demonstrates that our algorithm is capable to increase ROC area to 99.98%. It can be noted, the existing K-Means clustering algorithm provides less accuracy for ROC area, whereas the proposed algorithm provides highest accuracy for the same ROC area. This ability can increase the quality of classification in combination of PCA and K-NN methods. It can be seen from Fig. 3, the proposed algorithm exhibits an increment of 10% for ROC area. The ability allows increasing the classification accuracy.

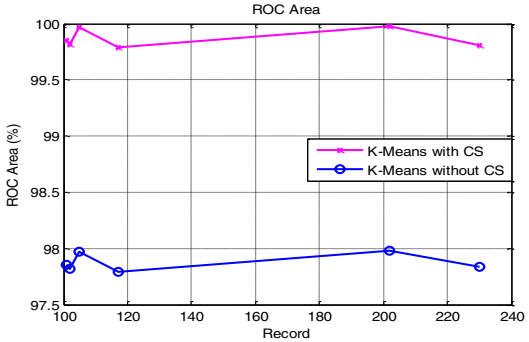

Fig. 3: ROC Area

Also, our algorithm is tested for all possible value of spread parameter σ which uses in the large ECG dataset instead of standard deviation for the PNN classifier. Figure 4 compears spread parameter based on our algorithm in combination with PCA and PNN methodes.

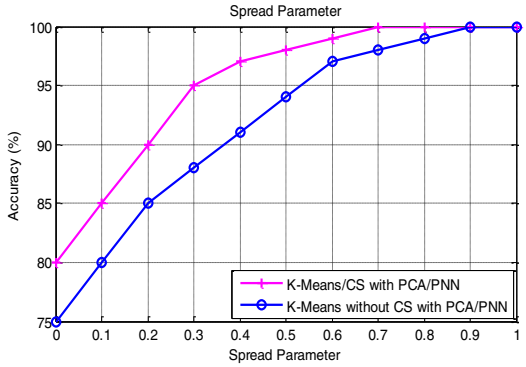

Fig.4: Sperad Parameter

Fig.4 confrims, that the proposed algorithm provides highest average accuracy at $\sigma = 0.7$ and the corresponding accuracy is 99.98% while for the exsisting K-Means algorithm the coressponding accuracy is 97.57%.

V. CONCLUSION

Clustering and classification algorithms play an important role in long-term recording of ECG signals. The fundamental objective of this paper was to establish new K-Means clustering algorithm based on CS theory. We have experimentally confirmed that our algorithm with a collaboration of PCA and K-NN as the dimensionality reduction and classification methods has demonstrated better performances than other methods. The proposed algorithm out-performed existing algorithms by increasing 11% classification accuracy in combination with PCA and K-NN methods in order to increase the classification speed. The proposed algorithm in combination with PCC and K-NN has achieved the following advantages: 1) 99.98% accuracy; 2) 99.98% ROC area; 3) 99.92% sensitivity. In addition, our algorithm in combination with PCA and PNN methods has demonstrated the following advantages: 1) high average accuracy for PNN classifier at $\sigma = .7$; 2) 99.22% PPV; 3) 98.22% sensitivity. The benefit of using our algorithm can be divided into two areas. One area is to improve the wired ECG systems to wireless ECG. The second area of benefits emphasizes within increasing the efficiently of treatment for heat diseases.

VI. FUTURE WORKS

We would like to extend the capabilities of the proposed algorithm to other types of classification and dimensionality reduction methods.

REFERENCES

[1] M. Balouchestani, K. Raahemifar and S. Krishnan, "High - resolution QRS detection algorithm for wireless ECG systems based on compressed sensing theory," in *Circuits and Systems (MWSCAS), IEEE 56th International Midwest Symposium* , 2013, pp. 1326-1329.

[2] S. K. Ambat, S. Chatterjee and K. V. S. Hari, "Fusion of Algorithms for Compressed Sensing," *Signal Processing, IEEE Transactions, 2013,* vol. 61, pp. 3699-3704.

[3] F. A. Ram and S. H. Khayat, "ECG signal compression using compressed sensing with nonuniform binary matrices," in *Artificial Intelligence and Signal Processing (AISP), 16th CSI International Symposium,* 2012, pp. 305-309.

[4] M. Balouchestani, K. Raahemifar and S. Krishnan, "A high reliability detection algorithm for wireless ECG systems based on compressed sensing theory," in *Engineering in Medicine and Biology Society (EMBC), 35th Annual International Conference of the IEEE/EMBC,* 2013, pp. 4722-4725.

[5] A. M. R. Dixon, E. G. Allstot, D. Gangopadhyay and D. J. Allstot, "Compressed Sensing System Considerations for ECG and EMG Wireless Biosensors," *Biomedical Circuits and Systems, IEEE Transactions, 2012,* vol. 6, pp. 156-166.

[6] J. Selvakumar, A. Lakshmi and T. Arivoli, "Brain tumor segmentation and its area calculation in brain MR images using K-mean clustering and fuzzy C-mean algorithm," in *Advances in Engineering, Science and Management (ICAESM), International Conference,* 2012, pp. 186-190.

[7] S. Lin, "Comparison of kohonen feature map against K-mean clustering algorithm with application to reversible image compression," in *Circuits and Systems Conference Proceedings, China, International Conference* , 1991, vol.2,pp. 808-811.

[8] F. Darko, S. Denis and Z. Mario, "Human movement detection based on acceleration measurements and k-NN classification," in *EUROCON, International Conference on "Computer as a Tool",* 2007, pp. 589-594.

[9] P. H. Tang and M. H. Tseng, "Medical data mining using BGA and RGA for weighting of features in fuzzy k-NN classification," in *Machine Learning and Cybernetics, International Conference,* 2009, pp. 3070-3075.

[10] J. L. Roux and C. Gueguen, "A fixed point computation of partial correlation coefficients in linear prediction," in *Acoustics, Speech, and Signal Processing, IEEE International Conference ,ICASSP '77.* 1977, pp. 742-743.

[11] S. Mishra, C. N. Bhende and K. B. Panigrahi, "Detection and Classification of Power Quality Disturbances Using S-Transform and Probabilistic Neural Network," *Power Delivery, IEEE Transactions, 2008,* vol. 23, pp. 280-287.

[12] S. Schmeelk and J. Schmeelk, "Image authenticity implementing principal component analysis (PCA)," in *Emerging Technologies for a Smarter World (CEWIT),10th International Conference,* 2013, pp. 1-4.

[13] V. Kumar, J. Sachdeva, I. Gupta, N. Khandelwal and C. K. Ahuja, "Classification of brain tumors using PCA-ANN," in *Information and Communication Technologies (WICT),* 2011, pp. 1079-1083.

[14] A. K. Sinha and K. K. Chowdoju, "Power system fault detection classification based on PCA and PNN," in *Emerging Trends in Electrical and Computer Technology (ICETECT), 2011 International Conference* , pp. 111-115.

[15] Y. Zhou and K. E. Barner, "Locality Constrained Dictionary Learning for Nonlinear Dimensionality Reduction," IEE *Signal Processing Letters, 2013,* vol. 20, pp. 335-338.

[16] Q. Zhou, "Study on ECG data lossless compression algorithm based on K-means cluster," in *Future Computer and Communication, FCC '09. International Conference on*, pp. 91-93.

[17] K. T. Fan, Y. C. Tzeng, Y. F. Lin, Y. J. Su and K. Chen, "Tree identification using a distributed K-mean clustering algorithm," in *Geoscience and Remote Sensing Symposium (IGARSS), IEEE International,* 2010, pp. 3446-3449.

[18] M. Balouchestani, K. Raahemifar and S. Krishnan, "New sampling approach for wireless ECG systems with compressed sensing theory," in *Medical Measurements and Applications Proceedings (MeMeA), IEEE International Symposium,* 2013, pp. 213-218.

[19] A. Ruta and F. Porikli, "Compressive Clustering of Hihh Dimensional Data," in *Machine Learning and Application (LCMLA),* 2011, pp. 380-385.