Intra-patient Arrhythmia Heartbeat Modeling by Gibbs Sampling

Ethery Ramírez-Robles, Miguel Angel Jara-Maldonado, and Gibran Etcheverry(✉)

Universidad de las Américas Puebla, Sta. Catarina Mártir, Cholula, Puebla, Mexico
gibran.etcheverry@udlap.mx

Abstract. Heartbeat modeling allows to detect anomalies that reflect the functioning of the heart. Certain approaches face this problem by using Gaussian Mixture Models (GMMs) and other statistical classifiers by extracting the fiducial points provided by the MIT-BIH database. In this work, MIT-BIH database heartbeats are modeled into different heartbeat types from a single subject by using the Gibbs Sampling (GS) algorithm. Firstly, a data pre-processing step is performed; this step involves several tasks such as filtering the raw signals from the MIT-BIH database and reducing the heartbeat types to five. Secondly, the GS is applied to the resulting signals of one subject. Thirdly, the Euclidean distance between each heartbeat type is calculated, and lastly, the Bhattacharyya distance is used to classify heartbeats. The results obtained by the GS algorithm were also compared to results obtained by applying the Expectation Maximization (EM) algorithm to the same data-set. Results allow to conclude that GS is a proper solution for separating each heartbeat type; by providing a significant difference between each heartbeat type which can be used for classification.

Keywords: Arrhythmia · Electrocardiogram · Gibbs Sampling algorithm · Expectation Maximization · QRS complex · R programming

1 Introduction

Electrocardiograms (ECG) are measurements of the electricity with which the heart operates. The QRS complex (which is a deflection on the ECG that states ventricular contraction and myocardial depolarization) can be used to analyze the ECGs. According to [1], cardiac disorders can be diagnosed by analyzing the perturbations in the normal electrical patterns. An arrhythmia is "any disturbance in the rate, regularity, site of origin, or conduction of the cardiac electrical impulse" [1]. An arrhythmia can be a single abnormal beat, or a series of different beats that cause rhythm disturbances during the whole lifetime of the patient.

The classification of arrhythmias detected in ECG signals has been investigated in different works. There exist several approaches such as linear discriminant classifiers in [2] and Gaussian Mixture Models (GMMs) in [3–5], among

J. A. Carrasco-Ochoa et al. (Eds.): MCPR 2019, LNCS 11524, pp. 195–205, 2019.
https://doi.org/10.1007/978-3-030-21077-9_18

others. Although their results are promising, accuracy and false positive rates are not yet unerring. This work subscribes to the electrocardiogram ECG raw signal treatment for arrhythmia classification and to the Markov Chain Monte Carlo (MCMC) filtering for ECG nonlinear dynamical modeling; see [6,7]. These approaches are considered given the difficulty encountered when modeling and classifying heart diseases because an ECG signal varies for each person, and "different patients have separate ECG morphologies for the same disease" [8]. Hence, here we consider the intra-patient analysis as a first step, given that the inter-patient protocol considers different patients with the same disease [9].

2 Heartbeat Dataset Description

For this work, the MIT-BIH Arrhythmia Database was used [10]. According to its creators, it was the first open access database that provided standard test material for arrhythmia detection, and it has been used since 1980. This database has a total of 48 records of over 30 min long (including records 201 and 202 which belong to the same subject). There are 25 men subjects, and 22 women subjects; and it includes a wide variety of waveforms, including normal beats, complex ventricular, junctional and supraventricular arrhythmias and conduction abnormalities. All heartbeats from each subject are presented as a collection of amplitudes, along with a file that allows to determine the key positions for the R waves of each heartbeat type. According to [2], the number of possible heartbeat types was reduced to the following five types: N, S, V, F, Q. This types are adopted in this work because they are a recommended standard by the Association for the Advancement of Medical Instrumentation (AAMI) [8]. The mapping procedure to obtain the N, S, V, F, Q nomenclature is shown in Table 1, which was obtained from [2].

3 Methodology

Two different methods were tested in this work; namely the GS algorithm, which is used in this work to generate samples from an ECG; and the EM algorithm suited for cases in which the data-set is not complete. According to [11], the GS can be thought of as a stochastic analog of the EM approach, used to obtain likelihood functions when missing data are present. The difference is that in the GS, random sampling replaces the expectation and maximization steps. For this reason, both methods are compared in this work in order to asses whether a stochastic solution performs better than its iterative analogue.

Before using the MIT-BIH heartbeat dataset, each heartbeat type had to be converted into one of the AAMI classes presented in Table 1. Once that this was achieved, the GS algorithm was used to obtain characteristics of the posterior distribution for each heartbeat type. Then, the obtained characteristics were used to calculate the Euclidean distance from each heartbeat type, and finally, the Bhattacharyya distance was used to classify the signal. This process is discussed in detail in the following sub-sections.

Table 1. MIT-BIH arrhythmia database heartbeat types convertion into AAMI heartbeat classes.

AAMI class	Description	MIT-BIH heart types
N	Non S, V, F, Q class heartbeats	Normal Beat (NOR), Left Bundle Branch Block (LBBB), Right Bundle Branch Block Beat (RBBB), Atrial Escape beat (AE), Nodal/Junctional Escape beat (NE)
S	Supraventricular ectopic beat	Atrial Premature beat (AP), aberrated Atrial Premature beat (aAP), Nodal/Junctional Premature beat (NP), Supraventricula Premature beat (SP)
V	Ventricular ectopic beat	Premature Ventricular Contraction (PVC), Ventricular Escape beat (VE)
F	Fusion beat	Fusion of Ventricular and Normal beat (fVN)
Q	Unknown beat	Paced beat (P), Fusion of Paced and Normal beat (fPN), Unclassified beat (U)

3.1 Pre-processing Step

All heartbeat types were mapped into one of the five AAMI heartbeat classes mentioned in Sect. 2. Each signal was pre-processed by a band-pass filter to reduce the influence of muscle noise, interference, and baseline wander. The chosen values for the filter ranged from 5 Hz to 15 Hz, as suggested by Pan and Tompkins, due to the fact that this is approximately the desirable band-pass to maximize the QRS energy, achieving a 99.3% detection of the QRS complex [12]. Hence, we separated each heartbeat by using the R peak location provided by the MIT-BIH dataset. We followed the Ghorbani et al. statement about separating heartbeats by using samples 225 ms before the R peak, and 400 ms after the R peak; yielding 0.65 s for each heartbeat [4]. Therefore, we divided each beat from 81 samples before the R peak (250 ms interval) to 82 samples before the R peak of the next QRS complex; this is shown in Fig. 1.

3.2 Gibbs Sampling (GS)

GS is an algorithm used to approximate a sequence of observations from a continuously distributed parameter vector Θ [11]. This algorithm was used under the assumption that heartbeats can be modeled by observing to which heartbeat type Probability Density Function (PDF) they approximate better. In order to

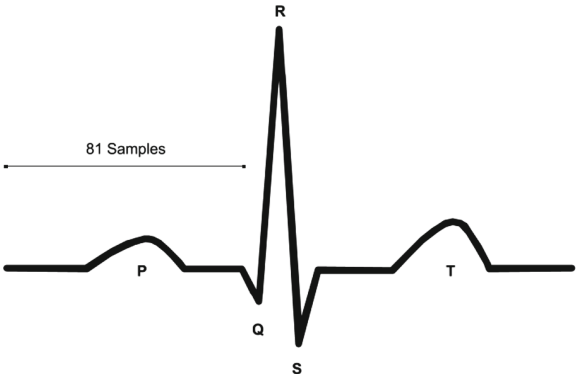

Fig. 1. The P wave (atrial depolarization and contraction) plus the QRS complex and the T wave (repolarization) [13].

achieve this, a Markov Chain (MC) is created to repeatedly sample the parameter sub-vectors $\Theta_1, ..., \Theta_M$, by using the following process. First, the starting value $\Theta^{(0)}$ of the parameter vector Θ is arbitrarily initialized (i.e. all $\Theta_i^{(0)}$ are randomly initialized). Then, the sub-vector $\Theta_0^{(1)}$ is sampled from the full conditional of Θ_0 with the rest of the Θ_i sub-vector values randomized in the previous step. This is done by using Eq. 1 [11]. This process is repeated until each $\Theta_1, ..., \Theta_M$ of the actual sub-vector (i.e. Θ^j) has been updated, yielding a new $\Theta^{(t)}$; where t stands for the current step (thus $t-1$ is the last calculated step).

$$\begin{aligned} \Theta_1^t &\sim P(\Theta_1 \mid \Theta_2^{t-1}, \Theta_3^{t-1}) \\ \Theta_2^t &\sim P(\Theta_2 \mid \Theta_1^t, \Theta_3^{t-1}) \\ \Theta_3^t &\sim P(\Theta_3 \mid \Theta_1^t, \Theta_2^t) \end{aligned} \tag{1}$$

The subsequent $\Theta^{(2)}$s are calculated using $\Theta^{(1)}$ instead of the arbitrary $\Theta^{(0)}$, and so on until the sequence $\Theta^{(0)}, ..., \Theta^{(N)}$ is obtained, which is a MC whose stationary distribution is the posterior distribution of Θ. Once converged to the stationary distribution, the MC samples the posterior distribution and can be used to obtain different characteristics of it [14]. For this work, those characteristics were used to calculate each heartbeat PDF and this is explained in Sect. 3.3.

In order to apply the GS algorithm to the MIT-BIH signals, the Windows implementation of the Bayesian analysis using GS (*WinBUGS* software)[1] was employed. It consists of a program capable of automatically tuning the most suitable Markov Chain Monte Carlo (MCMC) algorithm for a particular model. A normal distribution was used to explore each heartbeat likelihood type an the

[1] https://www.mrc-bsu.cam.ac.uk/software/bugs/the-bugs-project-winbugs/.

mean μ and precision τ variables were used to specify the mean and variance of it. The mean was updated by multiplying the y_{t-1} value by a ϕ normal distribution with zero mean and a variance of 0.0001. Afterwards, τ was updated from a gamma distribution with 0.1 mean and 0.0001 variance. Also, a scale parameter σ was occupied for the gamma distribution and was calculated by using Eq. 2. The WinBUGS software was called from a script developed in the R programming language, by using the *R2WinBUGS* library; and three MCMC chains were used with 2,600 iterations each, and 100 *burn-in* iterations (discarded iterations). No *thinning* was used (a strategy for reducing auto-correlation in the outputs [15]).

$$\sigma = \frac{1}{\sqrt{\tau}} \tag{2}$$

As an example of the posterior distribution characteristics obtained, Table 2 shows the posterior distribution characteristics obtained by applying the GS algorithm to the N-AAMI class heartbeats of subject 208. Three parameters are recovered; ϕ, σ, and deviance, from which the mean and standard deviation are calculated. This values are later used to calculate the proximity between each heartbeat type and to classify a heartbeat in one of the five AAMI classes.

Table 2. Subject 208 posterior distribution characteristics matrix.

Parameter	Mean	Standard deviation
ϕ	0.985	2.941×10^{-4}
σ	0.028	3.524×10^{-5}
Deviance	-1.388×10^{6}	0

3.3 Euclidean Distance

The posterior distribution characteristics obtained with the process explained earlier were used to determine the Euclidean distance between every type of heartbeat. In other words, first the GS was applied to the whole set of heartbeats of each class separately. Then, those characteristics were compared, by using the *Euclidean distance* as shown in Eq. 3 [16].

$$d(u,v) = \| u - v \| = \sqrt{(u_1 - v_1)^2 + (u_2 - v_2)^2 + ... + (u_n - v_n)^2} \tag{3}$$

where u and v are the two vectors to be compared. In this case, each vector would contain the posterior distribution characteristics of the different heartbeat types recorded in the signal. A matrix was generated, containing the distances of each heartbeat type. This allowed to better understand the separability of the data.

3.4 Bhattacharyya Distance

According to [17], the *Bhattacharyya distance* is used as a class separability measure. For this work, the Bhattacharyya distance between the p and q classes (which is applied to the case of two uni-variate normal distributions) was calculated by using Eq. 4.

$$D_{BC}(p,q) = \frac{1}{4} ln \left(\frac{1}{4} \left(\frac{\sigma_p^2}{\sigma_q^2} + \frac{\sigma_q^2}{\sigma_p^2} + 2 \right) \right) + \frac{1}{4} \left(\frac{(\mu_p - \mu_q)^2}{\sigma_p^2 + \sigma_q^2} \right) \tag{4}$$

where σ_p^2 and μ_p are respectively the variance and mean of the *p-th* distribution, and p, q are two different distributions.

In this work, the Bhattacharyya distance was used to classify heartbeats and was calculated between the ϕ values previously obtained (i.e. mean and standard deviation values obtained from the posterior distribution characteristics of each heartbeat). In concrete, the ϕ values obtained from applying GS to the whole set of an AAMI class of heartbeats, against the ϕ values obtained from the heartbeat to be classified. In other words, the **patient heartbeat** to be classified is compared against each of the AAMI class values obtained previously, to determine to which class it belongs to.

3.5 Expectation Maximization (EM)

As a means to compare the performance of the GS algorithm against another method, the EM algorithm was implemented and used for the classification of heartbeat arrhythmia. According to [18], the EM algorithm is occupied in those cases where the data set presents incompleteness. In this case, the algorithm was used to generate a model that allowed to separate the heartbeats into different PDFs in order to classify them. The PDF of the incomplete data is given by Eq. 5.

$$p_x(x,\theta) = \int_{Y(x)} p_y(y,\theta) dy \tag{5}$$

where $p_y(y,\theta)$ is the corresponding PDF and y contains the complete data samples, but cannot be directly observed, and θ is an unknown parameter vector. The Maximum Likelihood Estimate (MLE) of θ is given by Eq. 6.

$$\hat{\theta}_{ML} : \sum_k \frac{\delta ln(p_y(y_k,\theta))}{\delta\theta} = 0 \tag{6}$$

Since the y's are not available, the EM algorithm maximizes the expectation of the log-likelihood function, conditioned on the observed samples and the current iteration estimate of θ [12]. The three steps of the algorithm are enunciated next.

– Initialization: the GMM parameters are determined by the k-means clustering algorithm.

– Expectation (E)-step: the initial parameters are used to determine the probability that an observation at the $(t+1)th$ step of the iteration belongs to a component. This is achieved by using Eq. 7.

$$Q(\theta, \theta(t)) = E\left[\sum_k ln(p_y(y_k : \theta \mid X, \theta(t)))\right] \tag{7}$$

– Maximization (M)-step: the component parameters are re-estimated by maximizing $Q(\theta, \theta(t))$ through the use of Eq. 8.

$$Q(\theta + 1) : \frac{\delta Q(\theta, \theta(t))}{\delta \theta} = 0 \tag{8}$$

In order to use the EM algorithm, the pre-processed signals were separated into each heartbeat class (i.e., they were concatenated into a different vector for each heartbeat type). Then, the *cepstrum* vector for each heartbeat was obtained and its mean value was calculated to use it as an *expert* to better separate each heartbeat type. In other words, for each heartbeat contained in a class vector, the cepstrum was obtained, giving a vector of cepstrums of the same type, then the mean was obtained, and that scalar value was considered as an expert. During the classification process, only those amplitudes found in the R point were used, given they offer a better separation. For each heartbeat type, its corresponding R point amplitude vector was multiplied by their corresponding expert, and these vectors were then used by the k-means algorithm to initialize the GMMs. Finally, to classify a new heartbeat, the Bhattacharyya distance between the heartbeat to be classified and each heartbeat type vector was calculated, and the lower distance obtained was then used to decide to which class the heartbeat belonged to.

4 Results and Discussion

From the 48 patients included in the MIT-BIH Arrhythmia Database, 19 patients were selected for this work. The subjects that had more than 100 heartbeat records on two or more heartbeat types were chosen, so that the algorithm had enough data to classify. Those subjects that did not comply with this condition, as well as those with pacemakers were discarded. The chosen records were 106, 116, 119, 200, 201, 203, 207, 208, 209, 210, 213, 214, 215, 221, 222, 223, 228, 232 and 233. Most of the selected subjects had the N, S and V type heartbeats, while only two patients presented the F type too. Almost all analyzed patients presented an Euclidean distance between the mean and the standard deviation of the posterior distribution characteristics greater than 1,000,000 (i.e., the mean and standard deviation of μ, σ and deviance). The shortest distance obtained was 93,000 belonging to the N - V distance of subject 208.

As mentioned before, the Bhattacharyya distances were used to perform the beat by beat classification. Only the mean and standard deviation of the ϕ posterior distribution characteristics parameter were occupied. The accuracy

Table 3. Classification accuracy using GS and the Bhattacharyya distance.

Subject	Accuracy				
	N	V	S	F	Total
116	99.56%	98.16%	-	-	99.50%
119	58.91%	83.55%	-	-	64.41%
200	49.85%	79.17%	73.33%	-	59.44%
201	86.66%	81.81%	21.21%	-	80.78%
203	68.72%	54.27%	-	-	66.56%
207	71.01%	78.73%	88.78%	-	74.08%
208	75.16%	87.29%	-	93.54%	80.01%
209	80.57%	-	65.27%	-	78.62%
210	91.20%	80.51%	-	-	90.40%
213	81.82%	17.27%	-	66.57%	75.70%
214	61.58%	49.21%	-	-	60.18%
215	99.71%	81.70%	-	-	98.8%
221	90.69%	96%	-	-	91.55%
222	65.46%	-	40.06%	-	63.37%
223	74.57%	72.09%	15.06%	-	72.44%
228	98.16%	77.34%	-	-	94.48%
232	64.07%	-	35.09%	-	41.57%
233	72.15%	51.92%	-	-	66.66%

results obtained from using the Bhattacharyya distances with GS are presented in Table 3, where the accuracy was calculated for each heartbeat class, and the total accuracy is also presented. Recall that all the subjects have a different number of heartbeat records for each heartbeat type, being the N type the most frequent in most of the cases. Therefore, the classification performance is mostly influenced by the results obtained for the N heartbeat types. The best result was obtained from subject 116, which had an accuracy percentage of 98.8%. Furthermore, subjects 201 and 223 presented the lowest accuracy in the S heartbeat type classification, where more than 75% of the heartbeats were misclassified. This may be caused by a confusion between the S and V heartbeat types, whose proximity is one of the closest (the S - V Euclidean distance for subject 201 is 110,691.8; whereas the greatest Euclidean obtained is greater than 3,000,000 and corresponds to subject 215). Similarly, subject 213 had a low accuracy in the V class, probably because the V class varies in its morphology, and may resemble to the N class. Finally, in the case of the subject 201, 97 S class heartbeats were classified as V class heartbeats, from a total number of 165; while in the case of the subject 223, the S beats were principally misclassified as N beats.

Table 4. Classification accuracy using the EM algorithm.

Subject	Accuracy				
	N	V	S	F	Total
116	33.17%	87.88%	-	-	47.21%
119	80.10%	1.83%	-	-	76.56%
200	97.64%	83.89%	0%	-	92.15%
201	63.60%	99.49%	62.42%	-	67.06%
203	75.40%	96.62%	-	-	78.57%
207	91.70%	99.26%	0%	-	89.70%
208	48.99%	90.52%	-	94.36%	68.68%
209	7.43%	-	96.34%	-	18.77%
210	93.47%	92.82%	-	-	93.42%
213	19.46%	0%	-	96.96%	26.83%
214	51.39%	60.93%	-	-	52.48%
215	47.23.%	100%	-	-	49.80%
221	60.95%	99.49%	-	-	67.24%
222	97.18%	-	17.22%	-	90.45%
223	75.25%	87.94%	0%	-	75.45%
228	78.25%	98.61%	-	-	81.85%
232	92.71%	-	12.87%	-	30.73%
233	100%	65.66%	-	-	90.68%

In the case of the results obtained by the EM algorithm, some of the accuracy results are shown in Table 4. From this table it can be observed that for subjects 200, 207 and 223, the S type heartbeats were completely misclassified. This is produced by a bad initialization given in the k-means algorithm step. The S heartbeats in subjects 207 and 223, were clustered in the V type cluster (and in the case of subject 223, all the S beats were sent to the V type cluster, while the S type cluster included some heartbeats of type N and V). Furthermore, the performance drop presented in the N type heartbeats for subject 208 (for the EM algorithm), may be caused by the fact that the F type is a fusion heartbeat that occurs when electrical impulses from different sources act upon the same region of the heart at the same time (i.e., the F type is a fusion of the ventricular and the normal heartbeat types, which may cause confusion, generating N type heartbeats to be classified as F type).

5 Conclusions

In the present work we have used the Gibbs Sampling (GS) algorithm to model heartbeats from individuals in the MIT-BIH Arrhythmia Database according to

the AAMI classes, and compared its results with the ones obtained by using the Expectation Maximization (EM) algorithm. The posterior distribution characteristics obtained from the GS algorithm were used for each class separability and classification. A possible improvement to the results obtained by the GS algorithm could be the application of an expert to the heartbeat signals to classify.

Acknowledgements. Authors would like to acknowledge the Mexican National Council on Science and Technology (CONACyT) and the Universidad de las Américas Puebla (UDLAP) for their support through the doctoral scholarship program.

References

1. Thaler, M.S.: The Only EKG Book You'll Ever Need. Board Review Series. Lippincott Williams & Wilkins, Philadelphia (2007)
2. de Chazal, P., O'Dwyer, M., Reilly, R.B.: Automatic classification of heartbeats using ECG morphology and heartbeat interval features. IEEE Trans. Biomed. Eng. **51**(7), 1196–1206 (2004)
3. Povinelli, R.J., Johnson, M.T., Lindgren, A.C., Ye, J.: Time series classification using Gaussian mixture models of reconstructed phase spaces. IEEE Trans. Knowl. Data Eng. **16**(6), 779–783 (2004)
4. Ghorbani Afkhami, R., Azarnia, G., Ali Tinati, M.: Cardiac arrhythmia classification using statistical and mixture modeling features of ECG signals. Pattern Recognit. Lett. **70**, 45–51 (2016)
5. Martis, R.J., Chakraborty, C., Ray, A.K.: A two-stage mechanism for registration and classification of ECG using gaussian mixture model. Pattern Recognit. **42**(11), 2979–2988 (2009)
6. Escalona-Moran, M.A., Soriano, M.C., Fisher, I., Mirasso, C.R.: Electrocardiogram classification using reservoir computing with logistic regression. IEEE J. Biomed. Health Inform. **19**(3), 892–898 (2015)
7. Edla, S., et al.: Sequential Markov chain monte carlo filter with simultaneous model selection for electrocardiogram signal modeling. In: 34th Annual International Conference of the IEEE EMBS, August 2012
8. Kaplan Berkaya, S., et al.: A survey on ECG analysis. Biomed. Signal Process. Control **43**, 216–235 (2018)
9. Da, E.J., Luz, S., et al.: ECG-based heartbeat classification for arrhythmia detection: a survey. Comput. Methods Programs Biomed. **127**, 144–164 (2015)
10. Moody, G.B., Mark, R.G.: The impact of the MIT-BIH arrhythmia database. IEEE Eng. Med. Biol. Mag.: Q. Mag. Eng. Med. Biol. Soc. **20**, 45–50 (2001)
11. Walsh, B.: Markov chain monte carlo and gibbs sampling. Lecture Notes Online (2002)
12. Pan, J., Tompkins, W.J.: A real-time QRS detection algorithm. IEEE Trans. Biomed. Eng. **BME–32**(3), 230–236 (1985)
13. Association for the Advancement of Medical Instrumentation and American National Standards Institute. Testing and Reporting Performance Results of Cardiac Rhythm and ST-segment Measurement Algorithms. ANSI/AAMI. The Association (1999)
14. Lambert, B.: A Students Guide to Bayesian Statistics. SAGE Publications, Thousand Oaks (2018)

15. Ruppert, D.: Statistics and Data Analysis for Financial Engineering. Springer Texts in Statistics, 1st edn. Springer, Berlin (2010). https://doi.org/10.1007/978-1-4419-7787-8
16. Anton, H.: Elementary Linear Algebra. Wiley, Hoboken (2010)
17. Kashyap, R.: Combining dimension reduction, distance measures and covariance, November 2016
18. Theodoridis, S., Konstantinos, K.: Pattern Recognition, 4th edn. Academic Press Inc., Orlando (2008)