IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON DATA-ENABLED INTELLIGENCE FOR DIGITAL HEALTH

Received January 26, 2019, accepted February 17, 2019, date of publication March 4, 2019, date of current version April 5, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2902619

Continuous-Valued Annotations Aggregation for Heart Rate Detection

YUTING XIE[1], JIANQING LI[1,2], TINGTING ZHU[3], AND CHENGYU LIU[1]

[1]The State Key Laboratory of Bioelectronics, Jiangsu Key Laboratory of Remote Measurement and Control, School of Instrument Science and Engineering, Southeast University, Nanjing 210096, China
[2]School of Biomedical Engineering and Informatics, Nanjing Medical University, Nanjing 210029, China
[3]Department of Engineering Science, University of Oxford, Oxford OX1 2JD, U.K.

Corresponding authors: Jianqing Li (ljq@seu.edu.cn) and Chengyu Liu (chengyu@seu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61571113 and Grant 81871444, and in part by the Natural Science Foundation of Jiangsu Province under Grant BE2017735.

ABSTRACT In the medical field, experts usually annotate the bio-signals manually, and this is regarded as the gold standard. The manual annotating mode is time-consuming so widely replaced by an automated annotating algorithm. To address the low precision and low robustness of algorithm, we used a probabilistic model to synthesize the heart rate (HR) annotations from multiple annotators for electrocardiograph (ECG) signals and inferred the underlying true annotations and the precision of each annotator when the ground truth was not available. We further introduced signal quality indices in the model to improve our estimation. The 100 noisy ECG recordings in 2014 PhysioNet/computing in cardiology challenge database were divided into two parts, and various annotations for HR were generated by six available annotators. By employing the expectation maximization algorithm, we obtained the estimated true annotations for 80 recordings, and this result had an improvement not only over the best single annotator (17.46%) used in this paper but also to the mean and median strategies (the highest of 23.12% and 42.23%). Furthermore, the estimated precision of the single annotator from the proposed model served as the weight of the test data. In independent test, the weighted average of multiple annotations was superior to the single annotator and the mean strategy on 20 recordings, and its root mean square error (14.22 bpm) was close to that (13.96 bpm) of the proposed model on 80 recordings, demonstrating the robustness of the proposed continuous-valued annotation aggregation model.

INDEX TERMS Annotation aggregation, electrocardiograph, heart rate, ground truth, probabilistic model.

I. INTRODUCTION

Estimating ground truth of samples using the labels obtained from multiple annotators has been receiving increasing attention [1]. In order to enhance the data processing ability of models in machine learning, numerous labeled data are necessary for training and evaluating models. With the advent of crowdsourcing platform, it is convenient and efficient to release labeling tasks online, to collect multiple annotations of given data. However, the performances of annotators (manual or automatic) on crowdsourcing platform vary greatly due to their different levels of experience and accuracy.

Even though the crowdsourcing approach improves the efficiency of labeling tasks, it is impossible to be applied in some domains where expert annotations are required due to the complexity and individual differences of bio-signals [1]. However, the subjectivity in final annotations relies heavily on the levels of annotators and aggregators. In addition, there is no clear correct annotations for almost all subjective opinion tasks, such as the task of sentiment classification or lesion severity judgment from medical images [2]. For crowdsourcing platform or expert annotating, it is necessary to collect opinions from more than one annotator and build a model to fuse the multi-annotator opinions since the label from one annotator is unrepresentative and probably incorrect. In addition, these two methods of annotating are time consuming and unable to perform in real-time intelligent systems. On the contrary, the results of annotating algorithms that can operate automatically are not as accurate as manual annotations.

Cardiovascular diseases (CVDs) are the leading cause of death globally [3]. Wearable electrocardiograph (ECG) monitoring is considered as an essential tool for CVDs

The associate editor coordinating the review of this manuscript and approving it for publication was Yongqiang Cheng.

diagnosis. Heart rate (HR) is an important indicator of health [4]. Abnormal HR implies an increased risk of heart attack, providing insight into how hard the heart has to work both while resting and during exercise [5]. Thus, real-time and accurately HR monitoring has an important clinical importance.

Compared with the wristband-type HR monitor, HR calculated from ECG recordings is more accurate [5]. However, recent progress in mobile ECG device and portable battery-operated systems is challenging the accuracy of embedded QRS detectors due to variety of noises, resulting in the inaccurate dynamic HR estimation in wearable environments. In this study, we model the annotating process using a probabilistic approach to estimate the ground truth of HR from multiple QRS annotators. The model can handle the issue of multiple continuous-valued annotations. By fusing numerous annotations, the model can increase the annotating accuracy, on both high-quality and low-quality ECG signals.

II. RELATED WORK

In the case of the unknown true labels, majority voting and a wide variety of expectation & maximization (EM) methods have been used to aggregate multiple labels [6]. Majority voting is the strategy to aggregate the labels on which the majority of them agree as an estimate of the actual gold standard [7]. This strategy is only applicable for crisp labels such as binary labels or ordinal labels. EM algorithm is widely employed to compute the maximum-likelihood solution in presence of missing/hidden data [7], which includes an expectation step (E step) and a maximization step (M step). Latent variables (i.e., missing data) are computed in the E step given the current parameters and next the maximum likelihood (or maximum a posterior) estimation of parameters is updated with the current latent variables in the M step. These two steps iterate alternately until the convergence of latent variables or parameters [8], [9]. More details about the EM algorithm can refer to the reference [8].

Studies about estimating the gold standards of multiple labels mostly adopt the probabilistic model and EM algorithm to work in an unsupervised manner. The latent variables of EM algorithm commonly refer to the true labels and the parameters of the model usually include the precision of each annotator. Dawid and Skene [9] utilized maximum likelihood estimation and EM algorithm to estimate the observer error, resolving the diagnostic differences of multiple doctors caused by the inconsistent expressions of patients. Hosseini *et al*. [10] applied the model proposed by Dawid *et al*. to infer the relevance of documents during information retrieval and simplified the multivariate labels into binary labels. Whitehill *et al*. [11] established a probabilistic model for binary classification, assessing the ability of each annotator, the labeling difficulty of each picture and the true labels of pictures simultaneously.

All studies mentioned above focus on aggregating crisp labels concerned with classification problems. However, annotations of some instances are continuous values, such as the HR values from ECG signals. Raykar *et al*. [7] aimed to build a classifier to realize the computer aided diagnosis. Their method learned the binary classifier and the ground truth simultaneously. This method is bench-marking in this area and it was also extended to cope with multivariate, ordinal and continuous-valued labels. Kara *et al*. [12] proposed a new evaluation mechanism that divided annotators into eight basic types. Moreover, they built four Bayesian models to deal with different types of annotators, and adopted the maximum a posteriori estimation to solve the unknown parameters. Zhu *et al*. [1] presented a model for continuous labels and introduced feature vectors related to ECG signal quality. The model outperformed the median and mean strategies on the 2006 PhysioNet/Computing in Cardiology (CinC) Challenge database for QT interval measurement. Zhu *et al*. [13] then updated their method using Bayesian model and took the biases of annotators into the model for further optimization. Welinder and Perona [14] provided an online algorithm that solved the problem of annotation fusion of multi-type labels, including the binary, multivariate and continuous-valued labels.

Previous studies assumed that the accuracy of annotator is consistent with all the samples, which is not always true. When the annotating work is implemented manually, the error rates may fluctuate due to distractions or other reasons. When the data is annotated by algorithms, the accuracy may decrease on account of signal noises. Therefore, Zhang and Obradovic [15] used the Gaussian Mixture Model to simulate the distribution of the data and assumed that annotators had different labeling precisions for different Gaussian components. Later, they added a procedure to eliminate low-quality annotators at each Gaussian component according to evaluation score in [16] and [17]. Yan *et al*. [2] thought that the reliability of annotator depended on its own competence and the type and quality of input instances. Therefore, they assumed that multiple labels were subject to the Gaussian distribution and its variance was a function of feature vectors of instances.

Contributions of the current study can be summarized as follows. Firstly, we introduced a probabilistic model to compute the true medical annotations automatically so that the labels provided by smart devices would be more reliable. Secondly, we applied the concept of selective ensemble to search the optimal result in experiment. Finally, we verified the efficiency of the proposed model for HR estimation and HR annotator assessment.

III. METHODOLOGY

A. PROBABILISTIC MODEL

Continuous-valued annotation aggregation model (CVAAM) proposed by Raykar *et al.* [7] was used in this study to infer the true labels. We supposed that there are N samples labelled by R annotators. y_i^j represents the annotation provided by the jth annotator for the ith sample and the unknown ground truth is denoted by z_i. Assume that y_i^j follows a Gaussian

distribution with the mean of z_i and the variance of $1/\tau^j$:

$$P\left(y_i^j|z_i,\tau^j\right)=N\left(y_i^j|z_i,1/\tau^j\right) \tag{1}$$

Since the true annotations are closely associated with the features of signal samples, we incorporated the linear regression model and the feature vector of the ith sample $\boldsymbol{x}_i$ to reckon the actual target annotation z_i.

$$z_i=\boldsymbol{w}^T\boldsymbol{x}_i+e \tag{2}$$

In (2), $\boldsymbol{w}$ is the regression coefficient and e is a zero-mean Gaussian noise with inverse-variance a. The $\boldsymbol{x}_i$ and $\boldsymbol{w}$ have the same length, and $\boldsymbol{x}_i=[x_{i1}x_{i2},\ldots,x_{id},1]^T$. Note that $\boldsymbol{x}_i$ contains d features although it is a $d+1$-dimensional vector. Consequently, the probability density function of z_i is:

$$P\left(z_i\,|\,\boldsymbol{w},\boldsymbol{x}_i,a\right)=N\left(z_i\,|\,\boldsymbol{w}^T\boldsymbol{x}_i,1/a\right) \tag{3}$$

Combining (1) and (3), we have $P\left(y_i^j\,|\,\boldsymbol{w},\boldsymbol{x}_i,a,\tau^j\right)=\int P\left(y_i^j\,|\,z_i,\tau^j\right)P\left(z_i\,|\,\boldsymbol{w},\boldsymbol{x}_i,a\right)dz_i=N\left(y_i^j\,|\,\boldsymbol{w}^T\boldsymbol{x}_i,1/\tau^j+1/a\right)$. Then the posterior distribution of y_i^j can be simplified by replacing the $1/\tau^j+1/a$ with $1/\lambda^j$:

$$P\left(y_i^j\,|\,\boldsymbol{w},\boldsymbol{x}_i,\lambda^j\right)=N\left(y_i^j\,|\,\boldsymbol{w}^T\boldsymbol{x}_i,1/\lambda^j\right) \tag{4}$$

where λ^j represents the precision of the jth annotator.

B. DEVELOPED MODEL

We developed CVAAM by specifying the types of the inputting multiple annotations $\boldsymbol{y_i^j}$ ($\boldsymbol{i}=\boldsymbol{1},\ldots,N,\boldsymbol{j}=\boldsymbol{1},\ldots,\boldsymbol{R}$) and defining the feature vectors $\boldsymbol{x_i}$ ($\boldsymbol{i}=\boldsymbol{1},\ldots,\boldsymbol{N}$) of ECG signals to enhance the accuracy of HR annotations.

1) HEART RATE ANNOTATIONS

HR annotations were generated by annotating algorithms in two steps consisting of QRS detection and HR estimation. Firstly, we selected six QRS detectors to detect the QRS complexes of ECG signals. The six detectors were separately named as Hamilton-median algorithm [18], sixth-power algorithm [19], U3 transform algorithm (U3 algorithm) [20], difference operation algorithm (DOM algorithm) [21], 'jqrs' algorithm [22]–[24], optimized knowledge based algorithm (OKB algorithm) [25] in this paper. Though they were built on different basic theories, they all had high efficiencies and could be executed in nearly real-time processing on the mobile devices [26].

1) Hamilton-median algorithm is an improved version of Pan&Tompkins (P&T) method [27] that is one of the most widely used QRS detectors. Based on P&T method, Hamilton optimized the parameter selection and threshold estimation. When evaluated on the MIT-BIH arrhythmia database, Hamilton-median algorithm had an error rate of 0.46% while the P&T method of 0.7%.
2) Sixth-power algorithm mainly handle the sixth power of ECG signal, which enhances the strength of QRS complex more than that of the noises. This algorithm has simple rules for determining threshold.
3) U3 algorithm was performed by transforming the filtered ECG signals into a curve-length signal through U3 transform that is a non-linear transform in time-domain. Then a custom-built set of heuristics was applied on the curve-length signal for searching the R peaks.
4) DOM algorithm was an efficient method, in which the positions of positive extremes and negative extremes within each segment of differential signals should be firstly determined. The differential signal was calculated from the filtered ECG signal. Then R peaks were detected by matching the positions of those extreme points with the original ECG data.
5) 'jqrs' algorithm was also built on P&T method. However, the original band-pass filter was replaced by a QRS matched filter (Mexican hat) and an additional heuristic was added to ensure no detection during flat lines.
6) OKB algorithm included two moving averages that were calibrated by a knowledge base using only two parameters. In contrast to other high-accuracy methods, OKB algorithm shows great superiority on time efficiency [26].

Then we segmented the ECG signal into 10-s segments with a 90% overlap and the HR annotation y_i^j in the ith signal segment was calculated by the following equation:

$$y_i^j=\frac{60}{mRR_i^j}bpm \tag{5}$$

where mRR_i^j means the median of RR intervals that were detected by the jth QRS detector for the ith segment. It's apparent that the 10-s segment was regarded as a signal sample to be annotated in our work. HR obtained in this way represented the beating state of the heart for the past 10 s and was updated every second (90% overlap).

2) FEATURE VECTOR

Estimating the true HR by data fusion requires multiple labels generated by various annotators. In this paper, the distinction among diverse HR annotations for one signal sample mainly lied in the detecting results of QRS detectors.

Generally, various noises such as baseline drift and motion artifacts in poor-quality ECG signals make it hard to detect the QRS complexes and to reckon cardiac parameters. Hence, three signal quality indices of ECG proposed by Clifford *et al*. [28] and Li *et al*. [29] were used in this study.

- *kSQI*: the fourth moment (kurtosis) of the distribution.
- *basSQI*: the relative power in the baseline and it can be formulated as $1-\frac{\int_0^{1\text{Hz}}P(f)df}{\int_0^{40\text{Hz}}P(f)df}$.
- *fSQI*: the percentage of the signal which is not a flat line.

These indices provide additional information about the signal quality, which can assist the probabilistic model to obtain

the optimal results. Similar to HR calculation, the indices were also assessed from the 10-s ECG segments with 9 s of overlap. The signal feature with all indices of the ith 10-s segment can be represented as:

$$\boldsymbol{x}_i = [kSQI_i, basSQI_i, fSQI_i, 1]$$

Note that a scalar value of one was added in the feature vector to account for biases.

3) MODEL SOLUTION

We adopted the maximum likelihood estimation (MLE) to estimate the model parameters. It was assumed that the samples to be annotated are independently and each annotator works independently. Given the input data $\mathbf{D} = \left[\boldsymbol{x}_i, y_i^1, y_i^2, \ldots, y_i^R\right]_{i=1}^N$, the likelihood function of parameters $\boldsymbol{\theta} = (\boldsymbol{w}, \boldsymbol{\lambda})$ ($\boldsymbol{\lambda} = [\lambda^1, \lambda^2, \ldots, \lambda^R]$) can be formulated as:

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{N} P\left(y_i^1, y_i^2, \ldots, y_i^R \,|\, \boldsymbol{x}_i, \boldsymbol{\theta}\right) \\ &= \prod_{i=1}^{N} \prod_{j=1}^{R} N\left(y_i^j \,|\, \boldsymbol{w}^T \boldsymbol{x}_i, 1/\lambda^j\right) \end{aligned} \tag{6}$$

Then we estimated the parameters $\boldsymbol{\theta}$ by maximizing the log-likelihood and obtained the solution as below.

$$\boldsymbol{w} = \left(\sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^T\right)^{-1} \sum_{i=1}^{N} \left(\frac{\sum_{j=1}^{R} \lambda^j y_i^j}{\sum_{j=1}^{R} \lambda^j}\right) \boldsymbol{x}_i \tag{7}$$

$$\frac{1}{\lambda^j} = \frac{1}{N} \sum_{i=1}^{N} \left(y_i^j - \boldsymbol{w}^T \boldsymbol{x}_i\right)^2 \tag{8}$$

Then we simplified (7) as:

$$\boldsymbol{w} = \left(\sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^T\right)^{-1} \sum_{i=1}^{N} \hat{z}_i \boldsymbol{x}_i \tag{9}$$

where $\hat{z}_i$ is the estimation of the ground truth z_i and is formulated as

$$\hat{z}_i = \frac{\sum_{j=1}^{R} \lambda^j y_i^j}{\sum_{j=1}^{R} \lambda^j} \tag{10}$$

With the assumption that $\hat{z}_i$ is the weighted average of the annotations collected from all participating annotators, the weight of each annotator is its estimated precision λ^j. Thus, the annotations from high-precision annotators contribute more to the estimated actual labels, which is reasonable and logical.

Since (9) includes missing variables $\hat{z}_i$, we adopted the EM algorithm to work out the maximum likelihood problem. In the EM algorithm, the inferred true annotations $\hat{z}_i (i = 1, \ldots, N)$ and the unknown parameters $\boldsymbol{\theta}$ can be computed as:

1) Initialization: initializing the parameters $\boldsymbol{\theta} = (\boldsymbol{w}, \boldsymbol{\lambda})$
2) E step: estimating the $\hat{z}_i$ by (10) given the $\boldsymbol{\theta}$ from the last M step, $i = 1, \ldots, N$
3) M step: computing the MLE of model parameters $\boldsymbol{\theta}$ by Eqs. (8) and (9) with the $\hat{z}_i$ obtained in E step, $i = 1, \ldots, N$
4) Repeating the E step and M step until $\boldsymbol{\lambda}$ converges

The converging criterion for $\boldsymbol{\lambda}$ meant that its difference between two consecutive iterations was less than 0.0001.

Besides the MLE and EM algorithm, the concept of selective ensemble proposed by Tang and Zhou [30] was incorporated to the model solution as summarized in Algorithm 1 below and the complete implementation process was also presented in Fig. 1.

Algorithm 1 The Schematic Illustration of the Experiment and Evaluation

for $\boldsymbol{b}_s \in \boldsymbol{B}, s = 1, \ldots, 42$
  EM algorithm
  Input: annotations y_i^j provided by the annotators in $\boldsymbol{b}_s, j = 1, \ldots, R$ (R equals to the size of $\boldsymbol{b}_s$) and $i = 1, \ldots, N$
  feature vectors $\boldsymbol{x}_i$ composed of quality indices, $i = 1, \ldots, N$
  Initialize $\boldsymbol{\lambda}$
  E step: estimate the $\hat{z}_i$ using (9), $i = 1, \ldots, N$
  M step: estimate $\boldsymbol{w}$ using (8)
    update λ^j using (7), $j = 1, \ldots, R$
  Repeat E step and M step until convergence of $\boldsymbol{\lambda}$ or $\hat{\boldsymbol{z}} = [\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_N]$, or 100 runs are reached
  Calculate Root Mean Square Error (RMSE) between the referenced ground truth and the estimated annotations
  Output: $\hat{\boldsymbol{z}}$, $\boldsymbol{\lambda}$ and RMSE
end

The annotator selection was conducted by extracting 3 to 6 unrepeated annotators as an annotator subset each time and the labels provided by each annotator were then fed into CVAAM. As there are a total of six available QRS detectors in our experiment, the numbers of possible HR annotator subsets composed of 3, 4, 5, and 6 annotators are 20, 15, 6 and 1, respectively. We denoted the set of all annotator subsets as $\boldsymbol{B}$, and $\boldsymbol{B} = \{\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_{42}\}$. In Algorithm 1, the EM algorithm ran 42 times iteratively and output the results of $\hat{z}$, λ and RMSE each time.

C. EVALUATION METHOD

Root mean square error (RMSE) between the true z_i and estimated annotations $\hat{z}_i$ was used as evaluation index, which was calculated as:

$$\mathrm{RMSE} = \sqrt[2]{\frac{1}{N} \sum_{i=1}^{N} \left(\hat{z}_i - z_i\right)^2} \tag{11}$$

In this paper, the true annotations z_i were obtained by calculating the HR of 10-s ECG segment with the beat annotations available in database.

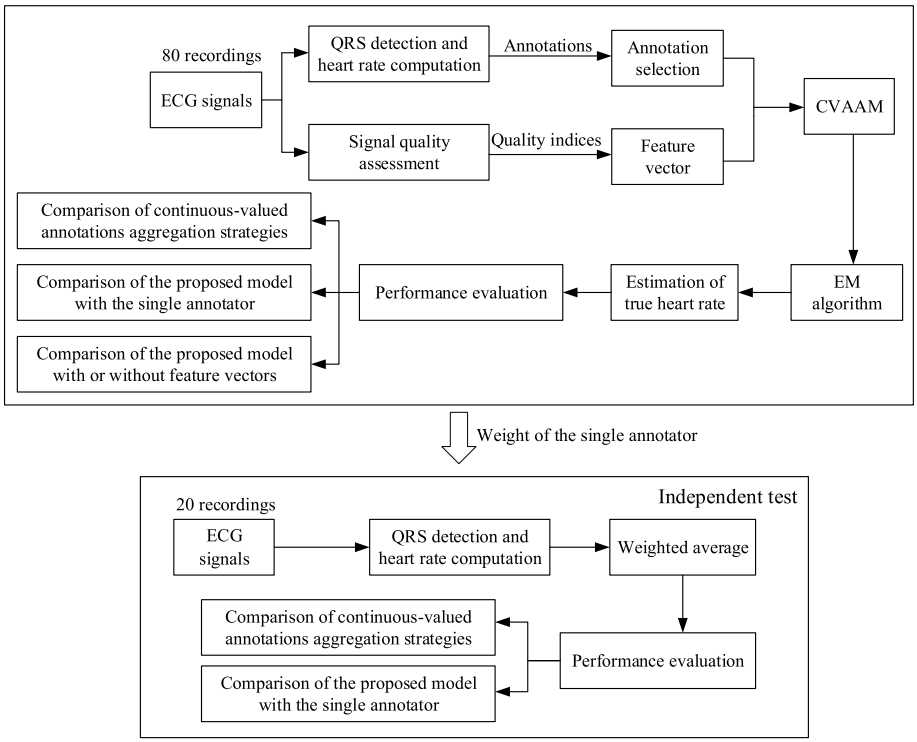

FIGURE 1. The schematic illustration of the experiment and evaluation.

The proposed model was evaluated in three aspects, as indicated in Fig. 1:

Firstly, we compared the proposed method with other annotation fusion strategies (mean and median strategies), and analyzed their performances using combination of different annotator algorithms.

Secondly, we compared the optimal result from the proposed model with that from single annotator algorithm, so that we could intuitively confirm whether the proposed model can enhance the automatic annotating results.

Thirdly, in order to observe the effect of employed ECG signal features on the performance of CVAAM, we also designed a simplified version of CVAAM without using the feature vectors. The solution to the simplified model was similar to the original, and it should be noted that the values of $\hat{\boldsymbol{z}}$ and $\boldsymbol{\lambda}$ of this simplified model were updated according to Eqs. (10) and (12) in the E step and M step of EM algorithm respectively until convergence. However, the EM algorithm was sensitive to initial values in this simple model. Thus the EM algorithm was repeatedly performed 100 times with uniformly distributed random initial values between 0 and 1 for each annotator combination and we took the average RMSE across 100 runs as the error of the current annotator group in the CVAAM without features.

$$\frac{1}{\lambda^j} = \frac{1}{N}\sum_{i=1}^{N}(y_i^j - \hat{z}_i)^2 \tag{12}$$

Finally, the precision of the individual annotator learnt from the CVAAM served as weight to compute the estimated ground truth of test data according to (10). The result was compared to that of single annotator algorithm and other strategies to observe the robustness of CVAAM for unseen data.

IV. EXPERIMENTS AND RESULTS

A. DATABASE

In this study, we utilized ECG recordings from the augmented training set in the 2014 PhysioNet/CinC Challenge, which contained 100 ECG recordings with 10-min or occasionally shorter length [31], [32]. The beats of these ECG recordings were annotated by experts manually and the calculated HR annotations could be regarded as the referenced ground truth z_i to evaluate algorithms. Liu *et al.* [26] tested ten widely used QRS detection algorithms on this augmented dataset and found that no algorithm had the detection accuracies higher than 80% since the ECG recordings were much noisy. Thus, it provided us a chance to utilize the CVAAM method to estimate the true HRs, and thus to test the effectiveness of this model for computing the true values of continuous labels and appraising the annotating precision of each annotator. We divided the dataset into two parts and the first part contained about 45,150 segments from 80 ECG recordings for training. The automated HR annotations generated by annotator algorithms was fed into CVAAM to estimate the true annotations and the precision of each annotator algorithm. The multiple annotations of the other part including 7736 segments from 20 recordings served as the test set.

TABLE 1. The RMSE for different aggregation methods.

No. of subset	Annotator subset	RMSE (bpm)			No. of subset	Annotator subset	RMSE(bpm)		
		Proposed	Mean	Median			Proposed	Mean	Median
1	[M1,M2,M6]	13.96	14.93	14.60	22	[M2,M3,M5,M6]	17.55	22.07	25.98
2	[M1,M2,M4,M6]	14.38	16.51	16.38	23	[M1,M2,M3,M5]	17.66	21.23	22.42
3	[M2,M4,M6]	14.82	18.42	23.34	24	[M2,M3,M4,M5,M6]	17.76	21.73	27.40
4	[M1,M2,M4]	15.20	16.87	16.00	25	[M2,M3,M4,M5]	17.87	22.21	24.21
5	[M1,M2,M5,M6]	15.33	17.43	17.76	26	[M2,M3,M5]	18.21	23.31	28.78
6	[M2,M5,M6]	15.40	18.64	23.35	27	[M1,M4,M6]	20.06	20.76	23.89
7	[M1,M2,M3,M6]	15.49	18.62	19.07	28	[M1,M4,M5,M6]	20.64	21.22	22.17
8	[M1,M2,M4,M5,M6]	15.50	17.81	19.15	29	[M1,M5,M6]	21.12	21.59	25.95
9	[M1,M2,M3,M4,M6]	15.88	19.11	23.33	30	[M1,M3,M4,M6]	21.67	22.95	25.57
10	[M2,M4,M5,M6]	15.91	19.21	21.14	31	[M1,M4,M5]	21.79	22.19	23.45
11	[M2,M3,M6]	15.94	20.73	27.58	32	[M1,M3,M6]	21.92	23.34	29.76
12	[M1,M2,M4,M5]	16.00	18.15	17.65	33	[M1,M3,M4,M5,M6]	22.20	23.28	29.18
13	[M2,M4,M5]	16.07	19.21	20.52	34	[M1,M3,M4]	22.73	23.99	27.57
14	[M1,M2,M3,M4]	16.23	19.36	19.13	35	[M1,M3,M5,M6]	22.88	24.12	27.64
15	[M1,M2,M3]	16.53	19.49	20.18	36	[M1,M3,M4,M5]	23.19	24.30	25.95
16	[M1,M2,M5]	16.64	18.87	19.94	37	[M4,M5,M6]	24.53	24.54	27.83
17	[M2,M3,M4,M6]	16.69	21.15	24.75	38	[M1,M3,M5]	24.93	26.27	29.56
18	[M2,M3,M4]	16.69	21.36	25.23	39	[M3,M4,M5,M6]	26.02	26.34	29.10
19	[M1,M2,M3,M5,M6]	16.86	20.20	25.75	40	[M3,M4,M6]	26.98	27.28	29.23
20	[M1,M2,M3,M4,M5,M6]	16.94	20.07	22.20	41	[M3,M4,M5]	27.74	28.04	32.73
21	[M1,M2,M3,M4,M5]	17.27	20.50	23.03	42	[M3,M5,M6]	27.78	28.19	32.62

All subsets of annotators and their corresponding RMSEs are listed respectively when estimating HR with the proposed, the mean and median strategies. The symbol M1-M6 in the second and seventh column separately indicates the Hamilton-median annotator, sixth-power annotator, U3 annotator, DOM annotator, 'jqrs' annotator, and OKB annotator. These HR annotators were named according to the QRS detection methods.

B. RESULTS

In this stage, the Algorithm 1 was used to estimate the true HR of 80 ECG recordings. The traditional EM algorithm is sensitive to initial values. However, changing the initial values of $\boldsymbol{\lambda}$ had no effect on the output when we kept the input fixed in Algorithm 1. But the starting values were related to the number of iterations for convergence. Furthermore, the time spent by Algorithm 1 was affected by the number of signal samples, annotations and feature components [1]. When inputting annotations of 80 ECG recordings (a total of 45,150 segments) from all six annotators, it took approximately 0.078s to converge for EM algorithm using MATLAB on a 3.2 GHz Intel Core processor.

1) PERFORMANCE OF DIFFERENT STRATEGIES

We obtained a series of inferred annotations and the estimated precisions of available annotators through the Algorithm 1, and the RMSEs between the estimated and referenced annotations were listed in Table 1. In addition, we also computed and displayed the RMSEs of mean and median strategies for 42 annotator subsets in Table 1, which were common methods in the problem of continuous-valued annotations aggregation.

According to Table 1, the proposed, mean and median strategies all obtained the minimal RMSE when we aggregated the information produced by annotator subset [M1, M2, M6] (i.e., an annotator combination consisting of Hamilton-median annotator, six-power annotator and OKB annotator). The minimum RMSE from the proposed model (13.96 bpm) was superior to the mean strategy (14.93 bpm) and median strategy (14.60 bpm). The maximum RMSEs were all from the annotator combination of subset [M3, M5, M6] (i.e., an annotator combination consisting of U3 annotator, 'jqrs' annotator and OKB annotator), with 27.78 bpm for the proposed model, 28.19 bpm for the mean strategy and 32.62 bpm for the median strategy. Compared with the two comparable mean and median strategies, the proposed model reported the lowest RMSE values at any annotator subset. The highest improvement of the CVAMM over the mean and median

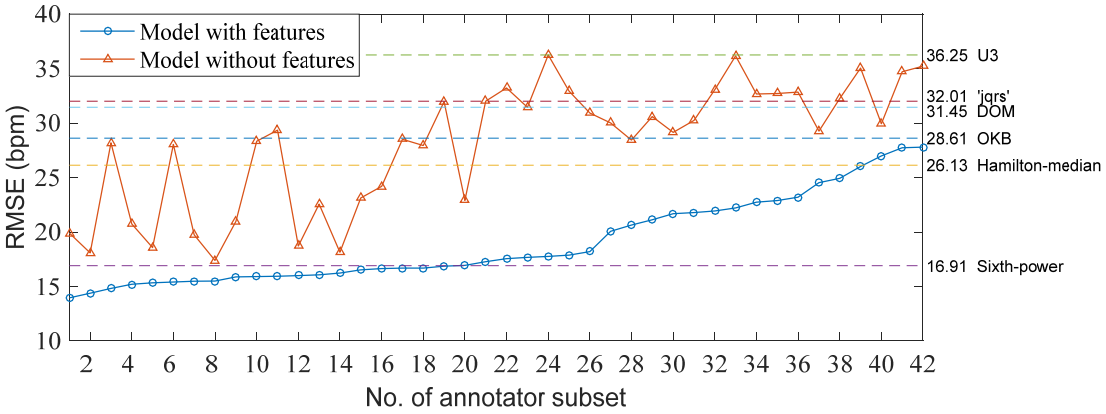

FIGURE 2. The RMSE of the proposed model with and without signal quality feature vectors for different annotator subsets. The dashed straight lines indicate the RMSE results from the six individual annotators.

TABLE 2. The RMSE for different annotators on test set.

Annotator	Hamilton-median annotator	Six-power annotator	OKB annotator	Mean strategy	Weighted average
RMSE (bpm)	39.34	15.14	27.48	15.59	14.12

annotation fusing methods were 23.12% and 42.23% respectively on annotator subset [M2, M3, M6].

2) COMPARISON WITH THE SINGLE ANNOTATOR

Fig. 2 shows the comparable results from the single annotator. The lowest RMSE of the individual annotator was 16.91 bpm from the sixth-power annotator and the maximum was 36.25 bpm from the U3 annotator. The lowest RMSE (13.96 bpm) from CVAAM using annotator subset of [M1, M2, M6] had an improved error rate of 17.46% and 61.50% when compared with the optimal and worst result of an individual annotator. Moreover, the top 19 best annotator subsets had RMSE values smaller than that from the best single annotator as shown in Fig. 2. In addition, 83.33% of 42 annotator subsets surpassed their own members which worked alone.

In Fig. 2, Sixth power annotator (RMSE $=$ 16.91 bpm) obviously performed better than 23 annotator subsets because other individual annotators were relatively inaccurate and their RMSEs ranging from 26.13 bpm to 36.25 bpm were much larger than that of Sixth power annotator. Among these 23 annotator subsets, there were 16 subsets that contained no Sixth power annotator and thus their inferior performance to the Sixth power annotator were reasonable on certain degree. In addition, the other 7 annotator subsets all contained two worst individual annotators, which may lead to some advantages of Sixth power annotator being neutralized.

3) EFFECTS OF FEATURES ON RMSE

Fig. 2 also shows the RMSEs of the proposed model without using signal quality feature vectors. In this scenario, the annotator subset of [M1, M2, M4, M5, M6] achieved a minimum mean RMSE of 17.40 bpm and the annotator subset of [M2, M3, M4, M5, M6] obtained a maximum mean RMSE of 36.25 bpm, both were worse than the best RMSE of 13.96 bpm from the proposed model with added features. From Fig. 2, the performance of model with signal quality features clearly outperformed that of model without features, which fully illustrated the importance of incorporating signal quality features.

4) INDEPENDENT TEST FOR ROBUSTNESS OF CVAAM

In order to observe the generalization ability of CVAAM, we applied it to the other 20 ECG recordings for independent test. In the test, we directly used the weight of the single annotator obtained by CVAAM in the previous stage to calculate the weighted average of multiple annotations according to (10) as the estimated ground truth. The results in the preceding paragraph revealed that synthesized annotations generated by CVAAM for 80 recordings achieved the lowest error when the sourced annotations were provided by the annotator subset [M1, M2, M6]. Thus we only aggregated the annotations from this subset in the independent test, and the weights of the Hamilton-median annotator (M1 annotator), six-power annotator (M2 annotator) and OKB annotator (M6 annotator) in (10) were respectively 3.3516, 5.8393 and 2.3035. Besides, we also utilized the Hamilton-median annotator, six-power annotator, OKB annotator and the mean strategy to estimate the true HR for test data. TABLE 2 shows the RMSE between the reference annotation and the annotations estimated by different methods.

In the Table 1, RMSE of the method of weighted average (14.12 bpm) was obviously smaller than that of the mean strategy (15.59 bpm) and all individual annotators (39.34 bpm, 15.14 bpm, 27.48 bpm). Moreover, the result of the weighted average method on test data was slightly larger than that of the CVAAM on 80 recordings, demonstrating the great robustness of the CVAAM for unseen data.

V. CONCLUSION

In this paper, we introduced CVAAM to estimate the unknown ground truth by fusing multiple continuous-valued labels in the absence of prior knowledge about used annotators. This model works in an unsupervised way and requires no training set. We applied CVAAM to HR measurement and obtained the optimal result with the RMSE of 13.96 bpm on 80 ECG recordings, which made an improvement of 17.46% over the optimal individual annotator and was also better than the commonly used mean and median multiple label fusing strategies. Furthermore, the model including signal quality features apparently outperformed the model without features and confirmed that feature vectors were crucial for improving the overall performance of the model. In independent test, the weight of the participating single annotator provided by CVAAM was used to calculate the true HR annotation by the method of weighted average on 20 ECG recordings, and the result (RMSE = 14.12 bpm) was slightly inferior to the result of CVAAM on 80 recordings but better than that of other methods on test data, which demonstrated the robustness of CVAAM. However, we assumed the precision of each annotator was consistent with all samples, which varied from the real condition in life and needed to be improved in the future work.

ACKNOWLEDGMENT

The authors thank the support from the Southeast-Lenovo Wearable Heart-Sleep-Emotion Intelligent Monitoring Lab.

REFERENCES

[1] T. Zhu, A. E. Johnson, J. Behar, and G. D. Clifford, "Crowd-sourced annotation of ecg signals using contextual information," *Ann. Biomed. Eng.*, vol. 42, no. 4, pp. 871–884, Apr. 2014.

[2] Y. Yan *et al.*, "Modeling annotator expertise: Learning when everybody knows a bit of something," in *Proc. 30th Int. Conf. Artif. Intell. Statist. (AISTATS)*, Sardinia, Italy, vol. 9, Mar. 2010, pp. 932–939.

[3] World Health Organization. *Cardiovascular Diseases (CVDs)*. Accessed: May 4, 2018. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs317/en/

[4] *How's Your Heart Rate and Why it Matters*. Accessed: Apr. 28, 2018. [Online]. Available: https://www.health.harvard.edu/heart-health/hows-your-heart-rate-and-why-it-matters

[5] S. W. Porges and E. A. Byrne, "Research methods for measurement of heart rate and respiration," *Biol. Psychol.*, vol. 34, nos. 2–3, pp. 93–130, Nov. 1992.

[6] H. Kajino, Y. Tsuboi, I. Sato, and H. Kashima, "Learning from crowds and experts," in *Proc. HCOMP*, Toronto, ON, Canada, 2012, pp. 107–113.

[7] V. C. Raykar *et al.*, "Learning from crowds," *J. Mach. Learn. Res.*, vol. 11, pp. 1297–1322, Apr. 2010.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc., B, Methodol.*, vol. 39, no. 1, pp. 1–38, 1977.

[9] A. P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the EM algorithm," *J. Roy. Stat. Soc. C, Appl. Statist.*, vol. 28, no. 1, pp. 20–28, 1979.

[10] M. Hosseini, I. J. Cox, N. Milić-Frayling, G. Kazai, and V. Vinay, "On aggregating labels from multiple crowd workers to infer relevance of documents," in *Proc. ECIR*, Barcelona, Spain, 2012, pp. 182–194.

[11] J. Whitehill, T. Wu, J. Bergsma, J. R. Movellan, and P. Ruvolo, "Whose vote should count more: Optimal integration of labels from labelers of unknown expertise," in *Proc. NIPS*, Vancouver, BC, Canada, vol. 22, 2009, pp. 2035–2043.

[12] Y. E. Kara, G. Genc, O. Aran, and L. Akarun, "Modeling annotator behaviors for crowd labeling," *Neurocomputing*, vol. 160, pp. 141–156, Jul. 2015.

[13] T. Zhu, N. Dunkley, J. Behar, D. A. Clifton, and G. D. Clifford, "Fusing continuous-valued medical labels using a Bayesian model," *Ann. Biomed. Eng.*, vol. 43, no. 12, pp. 2892–2902, Dec. 2015.

[14] P. Welinder and P. Perona, "Online crowdsourcing: Rating annotators and obtaining cost-effective labels," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.-Workshops*, San Francisco, CA, USA, Jun. 2010, pp. 25–32.

[15] P. Zhang and Z. Obradovic, "Learning from inconsistent and unreliable annotators by a gaussian mixture model and Bayesian information criterion," in *Proc. ECML-PKDD*, Athens, Greece, 2011, pp. 553–568.

[16] P. Zhang and Z. Obradovic, "Integration of multiple annotators by aggregating experts and filtering novices," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Philadelphia, PA, USA, Oct. 2012, pp. 1–6.

[17] P. Zhang, W. Cao, and Z. Obradovic, "Learning by aggregating experts and filtering novices: A solution to crowdsourcing problems in bioinformatics," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, Sep. 2013, vol. 14, no. 12, p. S5.

[18] P. S. Hamilton and W. J. Tompkins, "Quantitative investigation of QRS detection rules using the MIT/BIH arrhythmia database," *IEEE Trans. Biomed. Eng.*, vol. BME-33, no. 12, pp. 1157–1165, Dec. 1986.

[19] A. K. Dohare, V. Kumar, and R. Kumar, "An efficient new method for the detection of QRS in electrocardiogram," *Comput. Electr. Eng.*, vol. 40, no. 5, pp. 1717–1730, Jul. 2014.

[20] M. Paoletti and C. Marchesi, "Discovering dangerous patterns in long-term ambulatory ECG recordings using a fast QRS detection algorithm and explorative data analysis," *Comput. Methods Programs Biomed.*, vol. 82, no. 1, pp. 20–30, Apr. 2006.

[21] Y.-C. Yeh and W.-J. Wang, "QRS complexes detection for ECG signal: The difference operation method," *Comput. Methods Programs Biomed.*, vol. 91, no. 3, pp. 245–254, Sep. 2008.

[22] J. Behar, J. Oster, and G. D. Clifford, "Non-invasive FECG extraction from a set of abdominal sensors," in *Proc. Comput. Cardiol. (CinC)*, Zaragoza, Spain, vol. 40, Sep. 2013, pp. 297–300.

[23] A. E. Johnson, J. Behar, F. Andreotti, G. D. Clifford, and J. Oster, "Multimodal heart beat detection using signal quality indices," *Physiol. Meas.*, vol. 36, no. 8, pp. 1665–1677, Aug. 2015.

[24] J. Behar, J. Oster, and G. D. Clifford, "Combining and benchmarking methods of foetal ECG extraction without maternal or scalp electrode data," *Physiol. Meas.*, vol. 35, no. 8, pp. 1569–1589, Aug. 2014.

[25] M. Elgendi, "Fast QRS detection with an optimized knowledge-based method: Evaluation on 11 standard ECG databases," *PLoS ONE*, vol. 8, no. 9, Sep. 2013, Art. no. e73557.

[26] F. Liu *et al.*, "Performance analysis of ten common QRS detectors on different ECG application cases," *J. Healthcare Eng.*, vol. 2018, May 2018, Art. no. 9050812.

[27] J. Pan and W. J. Tompkins, "A real-time QRS detection algorithm," *IEEE Trans. Biomed. Eng.*, vol. BME-32, no. 3, pp. 230–236, Mar. 1985.

[28] G. D. Clifford, J. Behar, Q. Li, and I. Rezek, "Signal quality indices and data fusion for determining clinical acceptability of electrocardiograms," *Physiol. Meas.*, vol. 33, no. 9, pp. 1419–1433, Sep. 2012.

[29] Q. Li, R. G. Mark, and G. D. Clifford, "Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a Kalman Filter," *Physiol. Meas.*, vol. 29, no. 1, pp. 15–32, Jan. 2008.

[30] W. Tang and Z. H. Zhou, "Bagging-based selective clusterer ensemble," *J. Softw.*, vol. 16, no. 4, pp. 496–502, Apr. 2005.

[31] D. S. Benitez, P. A. Gaydecki, A. Zaidi, and A. P. Fitzpatrick, "A new QRS detection algorithm based on the Hilbert transform," in *Proc. Comput. Cardiol. (CinC)*, Sep. 2000, pp. 379–382.

[32] N. M. Arzeno, Z.-D. Deng, and C.-S. Poon, "Analysis of first-derivative based QRS detection algorithms," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 2, pp. 478–484, Feb. 2008.

Authors' photographs and biographies not available at the time of publication.