Performance evaluation of various learners on Human Identification based Moving on ECG signal

Pooja Ahuja[1] Abhishek Shrivastava[2]
[1,2]Department of Computer Science and Engineering
[1,2]DIMAT Raipur, India

Abstract— There is strong evidence that heart's electrical activity embeds highly distinct characteristics, suitable for applications such as the identification of human subjects. In other words, they contain satisfactory discriminative information to let the identification of individuals from a large population. Therefore, this paper presents a robust identification system using 20 healthy subjects from Physikalisch-Technische Bundesanstalt (PTB) database, 25 subjects from MIT-BIH arrhythmia database and 15subjects from The MIT-BIH Normal Sinus Rhythm database. This paper presents a new method which extracts essential amplitude, duration and gradient parameterized features on processed ECG signal essential for human identification. Finally, bagged tree classifier (ensemble classifier) is utilized to evaluate the accuracy of our method. With this system, we obtained a high identification rate (97.5%)

Key words: PTB, MIT-BIH, ECG signal

I. INTRODUCTION

These days strong efforts have been made for the development of next generation of biometric characteristics that are essentially robust to various attacks. Security biometrics is a secure substitute to conventional methods of identity proof of individuals, such as authentication systems based on user name and password.

Recently, it has been found that the electrocardiogram (ECG), bioelectrical activity is unique to each individual. ECG captures cardiac features from persons that are distinctive in nature. It describes the electrical activity of heart over time. It is been recorded with electrodes attached at surface of body. ECG analysis is not only a very useful diagnostic tool for clinical proposes, but also is recently studied as a potential biometric .The advantage of ECG over other biometrics is that it is impossible to mimic and forge, as they are internal biometrics internal biometric and far more reliable.

The ECG signal acquired from different persons is varied, generally reflected in the change in amplitude, time interval and morphology of the heartbeats. The most important features with the ECG includes the information lying in the P, Q, R, S, and T waves corresponding to the atrium and ventricular depolarization and repolarization. A sample ECG signal and the labeled wave fiducials [11] are shown in Figure 1.

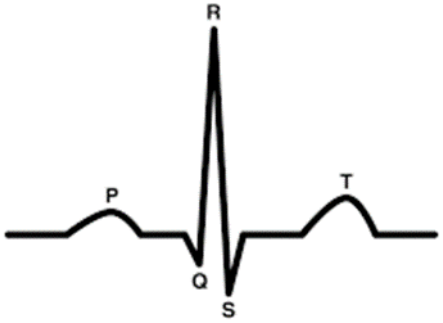

Fig. 1: a sample ECG signal and the labeled P, Q, R, S and T wave fiducials of single heartbeat

In this paper, a robust identification system based on selection of the best threshold values of ECG variables, as well as the best association of features for classification purposes, which is based on maximization of information content is proposed. The amplitude and duration of the waveform components, as well as QRS-loop rotation on the frontal plane and T-loop [10] orientation on the horizontal plane were obtained for each tracing. These, together with the independent classification of each case, were recorded on a floppy disk for computer processing. Each case was automatically allocated at random to one of two independent groups, labeled "training" and "test." In the next phase, an ensemble classifier is used to identify human from their ECG features. One's experiments were carried out using three Physionet datasets [1-3] and the evaluation was drawn on the basis of measuring quantities, such as subject identification (SI), heartbeat recognition (HR), and false acceptance\false rejection rate (FAR\ FRR).

II. ENSEMBLE CLASSIFIERS

Classification Ensemble combines a set of trained weak learner models and data on which these learners were trained. It can predict ensemble response for new data by aggregating predictions from its weak learners. It also stores data used for training and can compute resubstitution predictions. It can resume training if desired. B¨uhlmann and Yu (2003) pointed out that the history of ensemble methods starts as early as 1977 with Tukeys Twicing, an ensemble of two linear regression models. The main idea of ensemble methodology is to combine a set of models, each of which solves the same original task, in order to obtain a better fused global model, with more precise and reliable estimates or decisions than can be obtained from using a particular model. In this chapter we provide an overview of ensemble methods in classification tasks. We present all important types of ensemble methods including boosting and bagging [8]

A. Boosting:

Boosting is a general method for improving the performance of any learning algorithm. The method works by repeatedly running a weak learner (such as classification rules or decision trees), on various distributed training data. The classifiers produced by the weak learners are then combined into a single composite strong classifier in order to attain a higher precision

Schapire introduced the first boosting algorithm in 1990. In 1995 Freund and Schapire introduced the AdaBoost algorithm. The pseudo-code of the AdaBoost algorithm is described in Figure 2 .The algorithm assumes that the training set consists of m instances, labeled as -1 or +1.

Input: I (a weak inducer), T (the number of iterations), S (training set)
Output: $C_t, \alpha_t; t = 1, \ldots, T$
1: $t \leftarrow 1$
2: $D_1(i) \leftarrow 1/m; i = 1, \ldots, m$
3: **repeat**
4: Build Classifier C_t using I and distribution D_t
5: $\varepsilon_t \leftarrow \sum_{i:C_t(x_i) \neq y_i} D_t(i)$
6: **if** $\varepsilon_t > 0.5$ **then**
7: $T \leftarrow t - 1$
8: exit Loop.
9: **end if**
10: $\alpha_t \leftarrow \frac{1}{2} \ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$
11: $D_{t+1}(i) = D_t(i) \cdot e^{-\alpha_t y_t C_t(x_i)}$
12: Normalize D_{t+1} to be a proper distribution.
13: $t++$
14: **until** $t > T$

Fig.2. AdaBoost Algorithm

Boosting seems to improve performances for two main reasons:

1) It generates a ultimate classifier whose error on the training set is small by combining many hypotheses whose error may be large.
2) It produces a combined classifier whose variance is considerably lower than those produced by the weak learner

B. Bagging:

The most familiar technique that processes samples concurrently is bagging (bootstrap aggregating). The method aims to improve the accuracy by designing an improved composite classifier, I*, by amalgamating the various outputs of learned classifiers into a single prediction. Figure 3 presents the pseudo-code of the bagging algorithm (Breiman,1996). Each classifier is trained on a sample of instances taken with replacement from the training set. Generally each sample size is equal to the size of the original training set.

Input: I (an inducer), T (the number of iterations), S (the training set), N (the subsample size).
Output: $C_t; t = 1, \ldots, T$
1: $t \leftarrow 1$
2: **repeat**
3: $S_t \leftarrow$ Sample N instances from S with replacment.
4: Build classifier C_t using I on S_t
5: $t++$
6: **until** $t > T$

Fig. 3: Bagging Algorithm.

Bagging, like boosting, is a method for improving the precision of a classifier by producing different classifiers and combining multiple models. Though, the bagging method is rather hard to analyze and it is not easy to understand by intuition what are the reasons and factors for the improved decisions.

III. MATERIALS AND METHODS

A. Databases:

The ECG records were chosen primarily from existing ECG databases, including the MIT-BIH arrhythmia database [1], the Normal Sinus Rhythm Database [2] and the PTB database [3]. The existing databases are an excellent source of varied and well characterized data, and represent a wide variety of QRS and ST-T morphologies to which have been added reference annotations marking the waveform boundary locations.

For our experimental setup, 25 subjects were selected to form a subset of the MIT-BIH arrhythmia database. This selection was performed in a way that the subset consisted of ECGs which show mostly premature ventricular and atrial contractions. Next from the Normal Sinus Rhythm Database 15 subjects were selected on the basis of length of recording .And a subset of 20 healthy subjects was formed from the PTB database for our experiments. The criteria for the selection of the records were, to demonstrate healthy ECG waveforms and to have at least two recordings for every subject.

B. Preprocessing:

The raw ECG is often somewhat noisy and contains distortions of various origins. Great attention has been paid to the design of filters for the intention of removing baseline wander and powerline interference; both types of disturbance imply the design of a narrowband filter. The filtering techniques are primarily used for preprocessing of the signal and have as such been implemented in a wide variety of systems for ECG analysis. It should be remembered that filtering of the ECG is relative and should be performed only when the desired information remains undistorted. This important insight may be exemplified by filtering for the subtraction of powerline interference.

For noise reduction and baseline line removal, Bandpass filter and Derivative filer [9] was utilized .The processed ECG signal is shown in Figure 4.

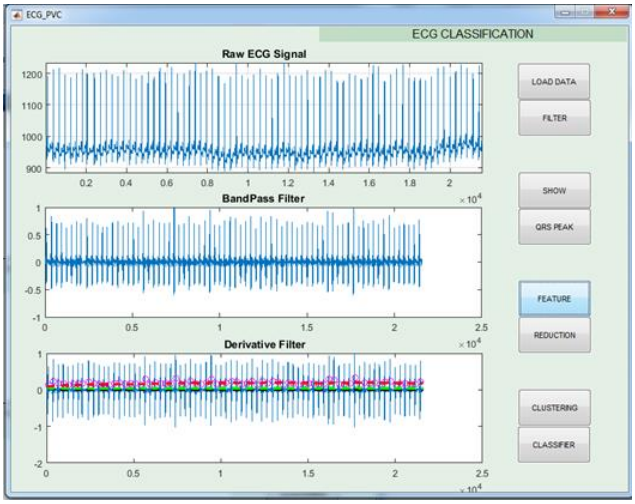

Fig. 4: a raw ECG signal passed through Bandpass filter and Derivative filter.

C. Features Extraction:

In the presented literature, the most frequently encountered types of features for human identification are morphological characteristics of single heartbeats. It has been suggested [2], [3], [4], [5], [6], [7] that amplitude and normalized time distances between successive fiducial points constitute unique patterns for different individuals. However, in these applications, it is implied that fiducial points can be success-fully detected.

The amplitude and duration of the waveform components, as well as QRS-loop rotation on the frontal plane and T-loop orientation on the horizontal plane were obtained for each tracing. These, together with the independent classification of each case, were recorded on a floppy disk for computer processing. Each case was automatically allocated at random to one of two independent groups, labeled "training" and "test."

Normal Values of Amplitude and Duration of ECG Parameters:

Amplitude-:

- p wave -: 0.25mV
- R wave -: 1.6 mV
- Q wave -:25 percent of R wave
- T wave -:0.1 to 0.5 mV

Duration -:

- P-R Interval -:0.12 to 0.20 sec
- Q-T Interval -:0.35 to0.44 sec
- S-T Interval -:0.05 to0.15 sec
- P wave Interval -:0.11 sec

Gradient -:

- Frontal P axis (°) -: 50-63
- Frontal QRS axis (°) -: 57-69
- Frontal T axis (°) -: 40-54
- QRS-T angle (°) -: 34-51

D. Identification Process:

Identification is done in the way that a record of a person's unique characteristic is captured and kept in a database. Later on, when identification verification is required, a new record is captured and compared with the previous record in the database. If the data in the new record matches that in the database record, the person's identity is confirmed.

Identification represents the last step of the proposed system. For this step, every input feature vector is compared to the ones stored in the gallery set in order to find the best match. To achieve this classification is done; in our work we have used ensemble classifier.

As expected, Ensemble classifier outperformed other single model classifier in every test, regardless of feature selection configurations. It is interesting to note that the margin by which Ensemble increased accuracy over other classifiers for very small feature vectors decreased to almost 0, whereas the accuracy of ensemble was significantly higher than others for larger feature vectors. The success of ensemble on small feature vectors can be attributed to its assumption that all features are mutually independent. For large feature vectors, there is a high likelihood that two words in the vector are dependent. This likelihood decreases for small feature vectors since any two words are less likely to be within each other's neighborhood in the document, thus their dependency is decreased

IV. RESULT AND DISCUSSION

In this paper, an ensemble classifier is used to predict identification rate. Ensemble method not only provides the best identification result in the test set, but also minimizes the recognition error in the training set as compared to single model classifiers. Figure 5 demonstrates ensemble classifier identification rates.

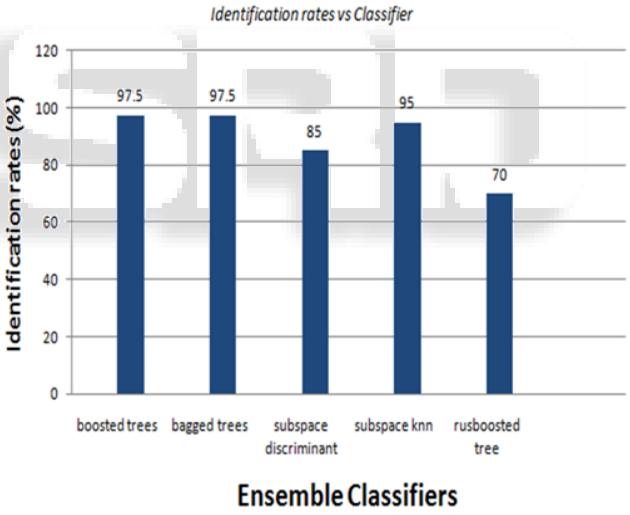

Fig. 5: ensemble classifiers identification rate

The performance of our identification system is measured on the parameters of false accept rate (F AR) and false reject rate (FRR) reported by the system.When the values attributed to a given ECG variable are gradually changed in successive classification experiments and the corresponding true-positive and false-positive ratios (TPR and FPR) are plotted against each other, the result is the so-called received operating characteristic (ROC) curve.

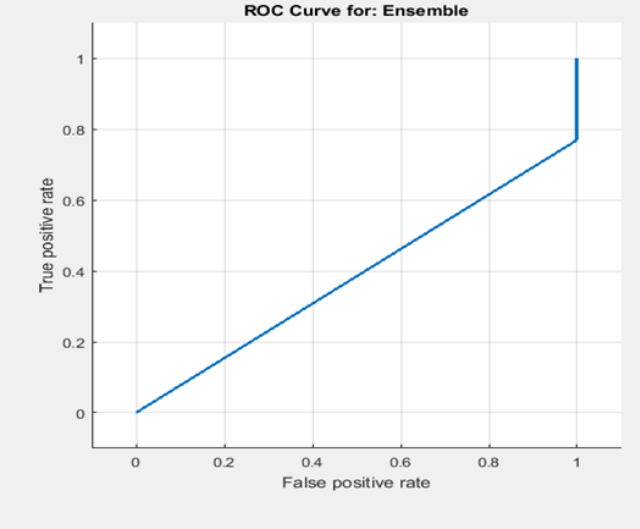

Fig. 6: ROC curve for the identification model .

The thresholds used are shown in the graph. We can see that with higher thresholds the system rejects too many legitimate users and with lower thresholds too many imposters are accepted.

These plots, developed as a method of observer performance analysis in detection experiments of electromagnetic signals transmitted through noisy channels, The Receiver Operating Characteristic (ROC) is plotted between genuine acceptance rate, GAR (i.e., 1 − F RR) and F AR to measure the system performance. The accuracy (Acc) of system is also determined using the factors FAR and FRR as, Acc(%) = 100 – (F AR+F RR /2).ROC curve is plotted for ensemble classifier as shown in figure 6.

V. Conclusion

ECG is a potential identification mechanism as a result of its competitive classification performances which promotes user acceptability. Automatic liveness detection of ECG signals has become the strength which lifts up ECG as a biometric modality that is suitable not only for normal individuals but also convenient to people with disabilities. We presented a robust personal identification approach on ECG signal. A procedure for selection of the best threshold values of ECG variables, as well as the best association of features for classification purposes, which is based on maximization of information content, is being used in our work. The ensemble classification method was used as a measure for the identification mechanism. ECG data for this investigation was obtained from Physiobank website (three databases) .Using this approach, 97.5% identification rate was achieved for 60 subjects. The results demonstrated the validity of our proposed method and the feasibility of using the ECG as a biometric measure for human identification.

The ECG is an interior feature and easy to combine with other exterior biometric features. Multimodal biometrics is a trend for future biometric identification systems. Also, future works can also concentrate in the design of a self-encryption identification system, based on the autocorrelation of ECG signals.

References

[1] G. B. Moody and R. G. Mark, The impact of the MIT-BIH arrhythmia database, IEEE Engineering in Medicine and Biology Magazine (2001) 45-50.
[2] Goldsmith RL, Bigger JT, Steinman RC, et al. Comparison of 24-hour parasympathetic activity in endurance-trained and untrained young men. J Am Coll Cardiol 1992; 20:552-558..
[3] Bousseljot R, Kreiseler D, Schnabel, A. Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet. Biomedizinische Technik, Band 40, Ergänzungsband 1 (1995) S 317.
[4] Maglaveras N. ECG pattern recognition and classificationnon linear transformations and neural network ks: a review. Int. J. Med. Inf., 52: 191-208. NIST report to Congress (2004).
[5] Haykin S . Adaptive filter theory. 4th Ed., New Jersey: Prentice- Hall, pp. 313-322. 2001..
[6] S. A. Israel, J. M. Irvine, A. Cheng, M. D. Wiederhold, and B.K. Wiederhold, ”ECG to identify individuals”, Pattern Recognition 38 (1): 133-142, 2005.
[7] Worck W. J. Irvine J. M. Israel S. A., Scruggs W. T., “Fusing face and ecg for person identification,” IEEE App. Imag. Paternt. Recogn. Workshop,, p. 226, 2003.
[8] Bauer, E. and Kohavi, R., “An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants”. Machine Learning, 35: 1-38, 1999.
[9] Indu Udai, Lekshmi P R, Sherin K Mathews and Tinu Maria Daie, “ECG Signal Processing Using DSK TMS320C6713”, IOSRJEN pp 24-43,2012.
[10] Cassiano Abreu-Lima, Duarte M. Correia, Jorge Almeida, Manuel Antunes-Lopes and Mario Cerqueira-Gomes,” A New ECG Classification System for Myocardial Infarction Based on Receiver Operating Characteristic Curve Analysis and Information Theory”,ahajournal 2016
[11] F. Agrafioti, J. Gao, D. Hatzinakos, Heart Biometrics: Theory, Methods and Applications, In Biometrics: Book 3, J. Yang, Eds., Intech., 2011, pp.199-216.

