Dimensionality Reduction for Anomaly Detection in Electrocardiography: A Manifold Approach

Zhinan Li*, Wenyao Xu*†, Anpeng Huang*, and Majid Sarrafzadeh*†
**Joint Research Institute in Science and Engineering by Peking University and UCLA,*
Peking University, China
†*Wireless Health Institute, University of California, Los Angeles, California, 90095, USA*

***Abstract*—ECG analysis is universal and important in miscellaneous medical applications. However, high computation complexity is a problem which has been shown in several levels of conventional data mining algorithms for ECG analysis. In this paper, we presented a novel manifold approach to visualize and analyze the ECG signal. According to regularity of the data, our algorithm can discover the intrinsic structure and represent the streaming data with a *1-D* manifold on a *2-D* space. Furthermore, the proposed algorithm can reliably detect the anomaly in ECG streaming data. We evaluated the performance of the algorithm with two different anomalies in wearable applications: for the anomaly from heart disorders such as apnea, arrythmia, our algorithm could achieve up to 90% recognition rate; for the anomaly from the ECG device, our algorithm could detect the outlier with 100%.**

***Keywords*-Electrocardiography, Dimensionality Reduction, Manifold, Locally Linear Embedding**

I. Introduction

Heart disease is one of dominant causes of death and its research always gains lots of attention due to the difficulty of detection, diagnosis, treatment and recovery. It is reported that almost 1 million Americans die of heart disease each year, which adds up to 42% of all deaths in 2010. With the unhealthy food and habit in modern life, heart disease does not just kill the elderly but becomes increasingly universal in the young people. The American Heart Association conservatively predicted that the costs of heart disease in United States will achieve 800 billion dollars a year by 2030.

Currently, the heart disease detection and diagnosis still rely on the analysis of Electrocardiography (ECG). The traditional ECG device has *three* to *twelve* electrodes, which connect to different parts of human body. Each electrode will measure the change of electrical voltage or current caused by the heart beat over time. Therefore, heart activity can be described by this kind of multi-dimensional time series. Whenever the heart of the patient does not work normally, there will be an anomaly reflected on the ECG waves. Therefore, medical professionals can diagnose the heart health status by ECG from the patients.

However, Electrocardiography is difficult to well-understand without plenty of professional training and clinical experiences. Currently, there are two major methods to analyze ECG signals in terms of the granularity of the observation. One is called *local* feature or *fine-grained* analysis. According to the signal shape, researchers partition the one-period ECG wave into several local featured segments, referred as *P-QRS-T* complexes model. Each segment has its own specific shape and is significantly different from others. Whenever one of the ECG segments becomes abnormal, the medical professionals could analyze the abnormal part and draw the medical conclusion according to the specific anomaly. For example, the *R-R* interval on ECG from a patient with apnea is obviously different from that from normal people. This local feature analysis is easy to observe and understand, and has been widely used in ECG analysis. [1] proposed a *1-D* signal searching method to find the inquiry pattern in ECG. [2] developed motif detection algorithm to discover the repeated signal pattern for ECG anomaly detection. [3] addressed the motif discovery and inquiry searching on multivariate cases. The other method is called *global* feature or *course-grained* analysis. Instead of looking at the specific features on the local part of ECG, global feature analysis tries to recognize the abnormal signals based on the overall observation. For example, [4] described a wavelet based method on ECG for heart disease diagnosis. [5] formulated the ECG analysis problem as artificial neural network (ANN) to extract the important features.

With the development of wearable medical techniques, medical devices could be non-invasively deployed on the patients without affecting their daily life. [6] presented a few of wearable ECG devices for heart attack prevention. Instead of that, cardiac arrests have to visit the hospital periodically for heart examination, their ECGs can be measured with portable devices by 24/7. However, due to the limited power and computation ability on wearable devices, it is impossible to deal with the ECG raw data directly for analysis. [7] discussed the randomness and determinism in ECG and pointed out that ECG might be an low-dimension embedding in a high-dimensional spaces. Inspired by this observation, we proposed a manifold based method to cluster the different ECG signals or detect the anomaly. Different from the discussed methods on the above, our method could reduce the data dimensionality and process the data on a lower dimension space. After the dimension reduction, the computation overhead could be reduced dramatically. To the

978-0-7695-4698-8 2012

DOI 10.1109/BSN.2012.12

best of our knowledge, this is the first work to detect the anomalies in ECG by exploiting its manifold structure.

The remaining paper is organized as follows. Section II will describe background and related dimension reduction technology. Section III will elaborate the proposed algorithm in details. In Section IV, we will evaluate our method on real ECG signals. Conclusion and future work will be discussed in Section V.

II. Background

A. Dimension Reduction Techniques

Dimension reduction techniques (DRT) has been widely investigated in the past few decades. The main idea of DRT is to represent the high-dimensional raw data on an intrinsic low dimensional space. The application is either to decrease the computational cost for raw data or visualize the raw data in a human visible way. According to its methodological theory, DRT can be categorized into multiple axes, such as generative method v.s. discriminative method, linear method v.s. non-linear method, global method v.s. local method. Principle Component Analysis (PCA) is the most well-known DRT, which extracts the *global* features of raw data to *linearly* reduce the dimension in a *generative* way. Because of the linearity of computation model, PCA is extremely efficient and could be used in real-time applications. For example, [8] proposed a PCA-based method to identify the human beings via their ECG signals.

Another famous DRT is Locally Linear Embedding (LLE) [9]. Contrary to PCA, LLE tries to preserve the data structure in a non-linear method according to local features of raw data. Compared to PCA, the significant advantage of LLE is that it could discover the non-linear underlying structure, such as manifold, in the data. For example, LLE is able to unfold S-Roll [9] without destroying its original structure. LLE also has lots of applications in the real world, such as face recognition , speech recognition and human movement classification, where the raw data are manifold and embedded in high-dimensional spaces. Furthermore, LLE is a kind of unsupervised techniques, and no training process is needed to unveil the intrinsic data structure.

B. Intrinsic Low Dimension Embedding of ECG

Real ECG data are high-dimensional and seem difficult to analyze and predict due to the high complexity and internal uncertainty. [7] discussed the determinism and randomness of ECG signals and proved that ECG could embed in a two-dimensional time-delay embedding space. As illustrated in Fig. 1, ECG can be segmented into several parts (A-G), and each part will be mapping onto a set of points in two-dimensional space. For example, as shown in Fig. 1, periods similar to Part A on the left figure will be mapped as stochastic points in Region A on the right figure. Afterwards, these points constitute a trajectory as a *1-D* manifold in *2-D* X-Y coordinates. We can see that, in spite of that the forming trajectory is fuzzy, its boundary and trend are deterministic and predictable. Based on this observation, we proposed LLE based recognition algorithm to discover the intrinsic non-linear structure of ECG for anomaly detection.

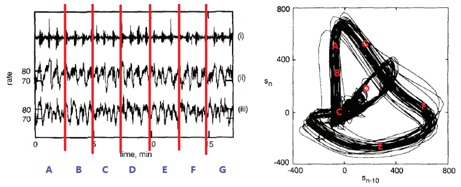
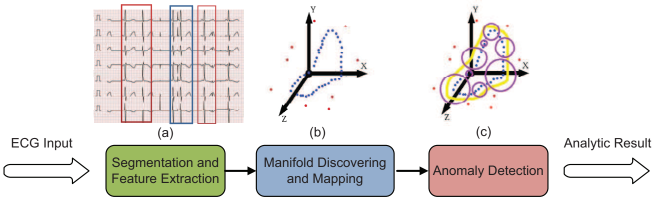


Figure 1. Manifold Structure of ECG (by courtesy of [7])

III. Manifold Based Anomaly Detection

In this section, we will present a manifold based approach to ECG anomaly detection. Fig. 2 shows the steps of the algorithm in flow chart. There are three phases in the proposed algorithm: segmentation and feature extraction, manifold structure discovering and mapping, anomaly detection and recognition. In the remaining part of this section, we will elaborate each steps in the algorithm.



Y
X
Z
Y
X
Z
(a)
(b)
(c)
ECG Input
Segmentation and Feature Extraction
Manifold Discovering and Mapping
Anomaly Detection
Analytic Result

Figure 2. The Flow of Manifold Approach to Detection Anomaly: Fig. 2(a) presents the ECG with anomalies, in which the blue rectangular marks the normal part, and the red rectangular marks the abnormal part. Fig. 2(b) shows the intrinsic 2-D manifold representation of the ECG. Fig. 2(c) is recognition results of our proposed algorithm.

A. Segmentation and Feature Selection

In the time series processing, there are two popular methods to segment the signal waves. One is fixed window cell size, and the data will be segmented equally with a fixed length. The other is period based cut, and the segment lengths are not necessary to be the same. In order to preserve the original structure of ECG, we will segment ECG with its distinct periodical feature, *R-R* interval.

Because of the variational lengths of *R-R* intervals, the similarity evaluation can not be easily integrated on most of manifold learning frameworks. There are some related research work on ECG local feature selection, such as *P-Wave*, *QRS*. For the sake of computation efficiency, we perform feature extraction on each segment with the statistical features on every sampling channel. In this paper, we choose *six* statistical features to represent the nature of ECG, including *arithmetic mean*, *standard deviation*, *derivative* *mean*, *derivative variance*, *correlation mean*, *correlation variance* [10]. In this way, each segment of sampling data from multiple channel will be transformed as a sequence:

$$X = \{x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{16}, ..., \\ , ..., x_{n1}, x_{n2}, x_{n3}, x_{n4}, x_{n5}, x_{n6}\}; \tag{1}$$

where n is the number of electrical electrodes of ECG device. For instance, if the device has 7 electrodes, the feature number of each segment will be 42.

B. Manifold Structure and Mapping

Our method to map the sequence X to a low dimension space is based on LLE framework [9]. LLE is an unsupervised algorithm and could reconstruct the data nonlinearly while preserving the locality. After the computation, the similar segments will be clustered automatically in a manifold form on the new low dimensional space. In general, there are *three* steps in the algorithm, which will be discussed in the following part.

1) K Nearest Neighbor Searching: The first step is to search K-nearest neighbors for each segment. All the nearest neighbors along with the basic segment will be grouped together. In the searching process, we use Euclidean distance to evaluate the similarity between segments. There are two ways to determine group size K in the searching procedure. One is using fixed integer. For example, we search the 5 nearest points and identify them as the neighbors. The other way is to identify the neighborhood by a threshold value in distance metrics. In this method, any point within a given radius will be recognized as the neighbor. Normally the topology of embedding will be well-preserved over a range of neighborhood size. We will discuss the algorithm performance variation in terms of group size K value selection in the experimental section.

2) Weighted Reconstruction With Nearest Neighbors: The second step is to fit each segment with its nearest neighbors. Assume that an arbitrary segment x with K nearest neighbors x_i, and its reconstruction error e can be formulated as:

$$e = \|x - \sum_{i=1}^{K} w_i * x_i\| \tag{2}$$

where w_i denotes the reconstruction weight from the component x_i. The optimization process is to minimize the reconstruction error of all segments by setting the weight w_i values. There are two attributes of the problem to ensure it well-imposed: 1) exclusiveness: the weight w_i of x is zero if x_i is not in the nearest neighbor list of x; 2)normalization: the sum of the weights of nearest neighbors should be 1. Therefore, we can rewrite the problem in the following format:

$$E = \sum_{j=1}^{N} \|x_j - \sum_{i=i}^{N} w_{ij} * x_{ij}\| \tag{3}$$

We can see that Eq.3 represents the reconstruction problem as a closed least square form, in which the weight w_{ij} could be solved efficiently [9].

3) Low Dimensional Embedding Construction: The third step is to construct the corresponding embedding in a low dimensional space. Based on the calculation results from the second step, the intrinsic geometrical structure of segments is characterized by w_{ij}. There is an assumption that the neighborhood relation in the high dimension space should be preserved in the low dimensional embedding manifold, and the nearest neighbor group should be with the same set, and the corresponding reconstruction weights w_{ij} will not get changed either. Based on this assumption, the embedding construction is to search the low dimensional representation y of x by minimizing the following error $E\prime$:

$$E\prime = \sum_{j=1}^{N} \|y_j - \sum_{i=i}^{N} w_{ij} * y_{ij}\| \tag{4}$$

where the weights w_{ij} are the computation results from Section III-B2, and the objects y_j are the low dimensional manifold. We can also notice that Eq. 4 is in a quadratic form and the embedding optimization process is comparably efficient to finish. Furthermore, all the manifold points y_i will be computed globally and simultaneously, and no local optima will affect the construction result.

Eq. 3 indicates that low dimension construction is only based on the locality of the high dimension data. This means that the computed manifold y_i can be translated with an arbitrary displacement without affecting Eq.4. Moreover, LLE states the computed manifold y_i can be rotated by an arbitrary angle without affecting Eq.4. This geometric attribute can be represented and formulated in the following two equations:

$$\sum_{i=1}^{N} y_i = 0 \tag{5}$$

$$\frac{1}{N} \sum_{i=1}^{N} y_i \cdot y_i = 1 \tag{6}$$

Therefore, manifold construction problem becomes an eigenvalue problem [9], in which we select the matrix rank to have the desired manifold dimension.

C. Anomaly Detection and Recognition

With the computing results from Section III-B, ECG segments have been mapped on a low dimensional manifold with a significant boundary. Fig. 3 shows an example of manifold construction results. The data here are from a patient with arrhythmia. The blue dots denote the regular ECG segments, and the red dots denote the abnormal ECG parts. The figure illustrates the blue dots constitute a *1-D* manifold (the yellow trajectory), and red dots distribute in all the space without any close form. We can envision that a trivial nearest neighbor (NN) search is possible to detect anomaly if the annotated set is large enough. However, any unknown point should be compared with every annotated regular segment with some threshold value. It is obvious that NN will become extremely low-efficient if the data scale is too large. Furthermore, the accuracy is not guaranteed due to the irregularity of manifold shape. In this phase, instead of dealing with raw data, we presented the manifold trajectory with dominated points (DP) to increase the efficiency of NN search. The algorithm is presented in Algorithm 1. Therefore, the searching space could be reduced. Note that our proposed algorithm is universal for any kind of manifolds, and the trajectory is not necessary to be a closed form.

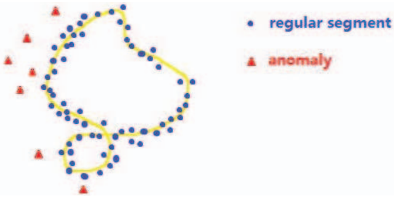


Figure 3. ECG Manifold with Regular Segments and Anomalies

IV. Evaluation

In this section, we discuss the performance of our proposed manifold approach to ECG anomaly detection. For integrity of the evaluation, we wish to perform the experiments comprehensively. In terms of ECG anomaly, there are two common resources to generate. The first is patient anomaly, where the abnormal ECG segments are from patients with unhealthy status, such as apnea, arrythmia. The second is device anomaly, where the abnormal ECG signals are from the improper use of ECG devices. For example, the electrodes of ECG device might be poorly even incorrectly attached on the body. Especially, the second cause is more common in wearable medical applications. We will address these two issues in the remaining part.

A. *Evaluation on Patient Anomaly*

To fairly evaluate the results of anomaly caused by heart disease, we use the online public ECG arrythmia database from PhysioBank Archive Index [11] as the benchmark. In this dataset, ECG is sampled with *three* channels, and all the segments in ECG have been annotated with *normal* or *anomaly* as the ground truth. According to the availability of the data, we use the benchmarks from *eight* patients with arrythmia for the experiments.

For the sake of comprehensive evaluation, we try different setups of group size K which is introduced in Section III-B1, and investigate the impact of group size K on the algorithm performance. Based on the fundamental of LLE, we understand that if K value is too small, the manifold construction in Section III-B3 has insufficient searching freedom; if K value is too large (more than the input dimensionality), the locality of raw data described in Section III-B2 will loose the unique definition [9]. Given the fact that the input dimensionality is 72, we evaluate the algorithm with K from 5 to 70, and the integer interval is 5. Fig. 4 shows the influence of K setup for the anomaly detection. From the experimental result (see Fig. 4), we can observe that the recognition rate will reach the maximal between 20 to 25, which is higher than 90% recognition rate.

Algorithm 1 Anomaly Detection

```
/* Step 1: Data Annotation and Setup */
Manifold regular segments $Y = (y_1, y_2, \cdots, y_n)$ from Section III-B;
Define a threshold value $T$ for importance evaluation;
Define the data set $Z$ to save generated DP;

/* Step 2: DP Generation */
$index = 1$;
for $i = 1$ to $n$ do
  if $y_i > y_{i-1}$ and $y_i > y_{i+1}$ then
    if $y_i / y_{index} > T$ then
      make $y_{index} \in Z$;
      increase $index$;
    end if
  end if
  if $y_i < y_{i-1}$ and $y_i < y_{i+1}$ then
    if $y_i / y_{index} > T$ then
      make $y_{index} \in Z_{DP}$;
      increase $index$;
    end if
  end if
end for

/* Step 3: NN Comparison with DP */
Recognize unknown input $\chi$ an anomaly or not;
while $i \leq dim(Z)$ do
  if $dist(\chi, z_i) \geq max(|z_i - z_{i-1}|, |z_i - z_{i+1}|)$ then
    $\chi$ belongs to anomaly;
  end if
end while
```

B. *Evaluation on Device Anomaly*

To learn the device anomaly, we performed pilot study in the lab with ECG device. The pilot study includes *three* subjects to evaluate the algorithm performance on anomaly caused by incorrect deployment of ECG electrodes. We simulated the misuse condition, and the experimental procedure is in this way: firstly, the subject wears the ECG device in the correct way, and we record the measurements as the ground truth. And then one of the electrodes will

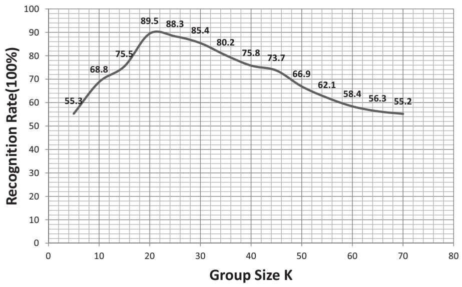


Figure 4. The Relation Curve with K Value and Recognition Rate

be poorly connected to the subject, and the corresponding measurements are labeled as *abnormal data*. For each section, the duration is about 3 minutes. And we iterated the experiments with each subject for *five* times.

Fig. 5 shows the two ECGs. The first is from correct measurements, and ECG is normally recorded. The second is from incorrect measurement, and the shape of ECG looks abnormal and random. When we performed the manifold embedding, these data were reconstructed under *2-D* coordinates. From Fig. 5, we can see that the data from green dots (correct measurements) are clustered together with a trajectory and far away from red dots (incorrect measurements). The experimental result shows that this kind of outliers (red dots) can be well-distinguished by our algorithm with 100% rate.

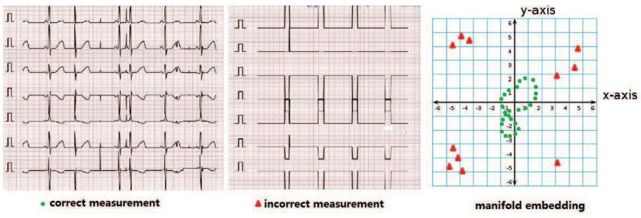


Figure 5. The Relation Curve with K Value and Recognition Rate

V. Conclusion and Future Work

In this paper, we introduced a manifold based approach to ECG anomaly detection. By taking the advantage of the regularity of ECG, the proposed method could explore the intrinsic signal structure and represent the ECG segments on a low dimensional space. The normal ECG segments will constitute a manifold, and the anomaly could be detected automatically. The experimental result shows that the proposed algorithm achieves high recognition rate for anomaly detection from two different resources: abnormal heart status (arrhythmia) and incorrect manipulation of ECG device (such as weak contact). In the future work, we plan to run clinical studies in the hospital to evaluate the performance of the algorithm. Also, in the view of promising experimental results, we could consider other applications with this proposed technique, such as disordered motif detection, patient identification.

Acknowledgement

This study is supported in part by National Science and Technology Major Project under contract No. 2012ZX03005013 and National Key Technology R&D Program under contract No. 2011BAZ03220-007.

References

[1] B. Huang and W. Kinsner, "ECG frame classification using dynamic time warping," in *IEEE International Conference on Electrical and Computer Engineering*, (Toronto, Canada), pp. 463–466, OCT 2002.

[2] A. Mueen and E. Keogh, "Online discovery and maintenance of time series motifs," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (New York, NY, USA), pp. 13 – 22, Aug. 2010.

[3] M. Potse, A. Linnenbank, and C. Grimbergen, "Software design for analysis of multichannel intracardial and body surface electrocardiograms," *Computer Methods and Programs in Biomedicine*, vol. 69, pp. 225 – 236, Nov. 2002.

[4] J. P. Martłnez, R. Almeida, S. Olmos, A. P. Rocha, and P. Laguna, "A Wavelet-Based ECG delineator: Evaluation on standard databases," *IEEE Transactions on Biomedical Engineering*, vol. 51, pp. 570 – 581, Nov. 2004.

[5] S. C. Saxena, A. Sharma, and S. C. Chaudhary, "Data compression and feature extraction of ECG signals," *International Journal of Systems Science*, vol. 28, pp. 483 – 498, Dec. 1997.

[6] S. Led, J. Fernandez, and L. Serrano, "Design of a wearable device for ECG continuous monitoring using wireless technology," in *Annual International Conference on Engineering in Medicine and Biology Society*, (Pamplona, Spain), pp. 1 – 5, Sept. 2004.

[7] H. Kantz and T. Schreiber, "Human ECG: nonlinear deterministic versus stochastic aspects," in *IEE Proceedings of Science, Measurment and Technology*, (Los Angeles, CA , USA), pp. 279 – 284, Nov. 1998.

[8] L. Beil, O. Pettersson, L. Philipson, and P. Wide, "Ecg analysis: a new approach in human identification," *IEEE Transactions on Instrumentation and Measurement*, vol. 3, pp. 808 – 812, Jun. 2001.

[9] L. Saul and S. Roweis, "Think globally, fit locally: unsupervised learning of low dimensional manifolds," *Journal of Machine Learning Research*, vol. 4, pp. 119 – 155, Dec. 2003.

[10] W. Xu, M. Zhang, A. A. Sawchuk, and M. Sarrafzadeh, "Co-Recognition of Human Activity and Sensor Location via Compressed Sensing in Wearable Body Sensor Networks," in *IEEE Conference on Body Sensor Networks*, (London, UK), May 2012.

[11] PhysioBank Archive Index. http://physionet.org/physiobank/database/(Checked on 03/18/2012).