Exploring Manifold Learning Methods For Automated Arrhythmia Analysis

Yan Yan, *Student Member, IEEE,* and Lei Wang, *Senior Member, IEEE*

***Abstract*—The abstract goes here. demo file is intended to serve as a “starter file” for IEEE journal papers**

***Index Terms*—IEEE, IEEEtran, journal, LaTeX, paper, template.**

I. Introduction

An era of big data in healthcare is now under way, decades of progress in digitalizing medical records accumulate vast amounts of medical data, simultaneously mobile healthcare and wearable sensor technologies offer healthcare data from larger population coverage. The noninvasive, inexpensive and well-established technology of electrocardiographic signal in mobile health or personal health has the greatest popularity in heart function analysis. Automated electrocardiography classification provides indispensable assist in long-term clinical monitoring, and a large number of approaches have been proposed for the task, easing the diagnosis of arrhythmic changes as well as further inspection, e.g., heart rate variability or heart turbulence analysis.

Lots of algorithms had been proposed for the classification and detection of electrocardiography signals. The electrocardiography classification or detection task had been divided into two parts: the feature extraction process and classifier. Sim- ple classifier such as linear discriminants [2] and kNN [3], more complex classifiers like neural networks [47], fuzzy inference engines [7,8], hidden Markov model [9,10], independent component analysis [11] and support vector machine [3,12,13] were also adopted by lots of researchers. Beyond the classifier, the performance of a recognition system highly depends on the determination of extracted electrocardiography features. Time domain features, frequency domain features, and statistical measures features for six fundamental waves (PQRSTU) had been used in feature extraction process [14]. Time domain features like morphological features include shapes, amplitudes, and durations were adapted primarily in [1517], frequency domain features like wavelet transformation were widely used [18], [19] stationary features like higher-order statistics also had been developed. Principal component analysis [20] and Hermite functions [21] have been used in electrocardiography classification and related analysis technologies as well. Almost every single published paper proposes a new set of features to be used, or a new combination of the existing ones [1]. The results from these algorithms or models were not amenable to expert labelling, as well as for the identification of complex relationships between subjects and clinical conditions [22]. But for the ambulatory electrocardiography clinical application, as well as the normal application in daily healthcare monitoring for cardiac function or early warning of heart disease, an automated algorithm or model would have significant meaning. The application of artificial intelligence methods has become an important trend in electrocardiography for the recognition and classification of different arrhythmia types [22]. The data explosion puts forward the new request to the method of data processing and information mining. Over the past decades computational techniques proliferated in the pattern recognition field, simultaneously the applications in electrocardiography recognition, detection and classification for relevant trends, patterns, and outliers. Most of the literatures in the electrocardiography classification task were focused on the supervised learning methods, as in unsupervised learning methods were infrequently used, which needs a lot of effort in labelling data. The MIT-BIH database [23] was the most widely used data in the classification and detection algorithm developments, while mass unlabelled electrocardiography data had been ignored due to the supervise learning approaches essential. Unsupervised learning methods become crucial in mining or analyzing unlabelled data, as the unlabelled electrocardiography data accumulated. Unsupervised learning-based approaches and the application to electrocardiogram classification in literatures mainly include clustering-based techniques [21,24,25], self-adaptive neural network-based methods [26,27] and some hybrid unsupervised learning systems [28].

II. State of the Art

Unsupervised off-line learning has seen very little attention in 3D model retrieval. The purity proposed by Bustos et al. [Bustos04] can also be considered as a weak form of unsupervised off-line learning. Purity is an estimate of the performance of a shape descriptor determined by using a pre-classified training database. Bustos used the purity to weight distance obtained from multiple shape descriptors. Classical methods for unsupervised learning of subspace includes PCA and MDS, both of which are quite effective if the feature points lie on or near a linear subspace of the input space. However, if the subspace is non-linear, these methods do not work well. Many non-linear methods have been proposed for unsupervised learning of subspace; Self-Organizing Map (SOM) and Kernel-PCA are some of the wellknown examples

Y. Yan was with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, and University of Chinese Academy of Sciences, China, e-mail: yan.yan@siat.ac.cn.

L. Wang was with Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences.

Manuscript received ; revised.

[Haykin99]. Recently, a class of geometrically inspired nonlinear methods, called manifold learning has been proposed for learning the m-manifold of measured feature vectors. Some of the examples of manifold learning algorithms are Isomap [Tannenbaum00], Locally Linear Embedding (LLE) [Roweis00], Laplacian Eigenmaps (LE) [Belkin03], Locality Preserving Projection [He03], and others. Manifold learning has been applied to many problems, including articulated figure motion analysis, multimedia data retrieval, and others. As a side note, these manifold learning algorithms have connections to reconstruction of a surface-based a 3D model from a point set, that is, estimation of a 2D manifold embedded in 3D space (e.g.,[Alexa01]). Analysis of 3D point set shape by using local meshing and spectra of mesh Laplacian, an approach used to watermark 3D point set models [Cotting04, Ohbuchi04] have resemblance to LE and LLE. Most of the effort so far on learning in the area of shape-based 3D model retrieval has been on interactive relevance feedback based approaches [Elad01, Leifman05, Novotni05]. Elad et al. is among the first to apply Support Vector Machines (SVM) learning in an online learning setting to improve 3D model retrieval [Elad01]. Leifman et al. [Leifman05] performed Kernel Principal Component Analysis (Kernel PCA) for an unsupervised learning of a feature subspace before applying a relevance feedback technique that employs Biased Discriminant Analysis (BDA) or Linear Discriminant Analysis (LDA) on the learned subspace. Novotni et al. [Novotni05] compared several learning methods, SVM, BDA, and Kernel-BDA, for their retrieval performance in a relevance feedback setting.

III. Methodology

The arrhythmia analysis problem can be generally considered as one time series based classification problem from the viewpoint of data analysis. Consider a set of n electrocardiography based time series data set $\boldsymbol{X} = \{\boldsymbol{x}_i\} \subset \mathbb{R}^d$, is given as the samples. Heartbeats in $\boldsymbol{X}$ include both labelled and unlabelled samples, the task is to estimate the labels of unlabelled data with some methods based on the labelled and relative methods. Here we adopt several manifold learning based techniques to embed the high dimensional input data into some lower dimension subspace. With the techniques some data properties were preserved in the embedded subspace (feature space). Then we train a classifier in the feature space using the labelled data, and use the classifier to classify the unlabelled data. From which we can get the arrhythmia information. Here we introduce several manifold learning techniques, the detailed theory and explanation can be accessed from the related literatures.

A. PCA

Principal Component Analysis (PCA) is a linear dimensionality reduction method, which try to find a linear subspace of lower dimension keeping the largest variance compare to the original feature space. PCA is by far the most popular unsupervised linear technique. The most important task in PCA method is to find the principal components of a data set, which can be implemented by finding a linear basis with reduced dimensionality for the input data sets, with the amount of variance in the data is maximal.

Mathematically, consider data set $\{\boldsymbol{x}_i\} \subset \mathbb{R}^d$, the embedded subspace with PCA can be denote as $\{\boldsymbol{y}_i\} \subset \mathbb{R}^k$ (d and k are the dimension numbers). Then the problem of finding the "best" k principal components can be transfer to finding k-dimensional subspaces that minimize the orthogonal distances. Solve

$$\underset{\mathcal{S}}{\operatorname{argmin}} \sum_{i=1}^{d} \|\boldsymbol{x}_i - P_{\mathcal{S}}(\boldsymbol{x}_i)\|_2^2 \tag{1}$$

where $P_s(\cdot)$ is the projection onto subspace $\mathcal{S}$. Solution of this problem is to find a linear mapping $\boldsymbol{M}$ from the original space to subspace $\mathcal{S}$, which maximizes the cost function $trace(\boldsymbol{M}^T cov(\boldsymbol{X})\boldsymbol{M})$). Then the data low-dimensional data features $\boldsymbol{y}_i$ of data points $\boldsymbol{x}_i$ in original space can be computed by mapping $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{M}$. A detailed theory analysis and tutorial can be found in [1] and [2] respectively.

B. MDS

Multidimensional Scaling (MDS) is a typical form of nonlinear dimensionality reduction method. MDS algorithm tries map each object some metric space such that the intra-point distances are preserved as well as possible. MDS had been widely used in data analysis, data representation & visualization, and manifold learning.

Mathematically, consider a MDS mapping $\boldsymbol{\Phi}$ from data set $\mathcal{X} = \{\boldsymbol{x}_i\}$ to a target space $\mathbb{R}^n$, such that the distance condition holds:

$$\|\boldsymbol{\Phi}(\boldsymbol{x}_i) - \boldsymbol{\Phi}(\boldsymbol{x}_j)\|^2 = d^2(\boldsymbol{x}_i, \boldsymbol{x}_j) \tag{2}$$

blabalaall

C. Isomap

MDS used the distance definition based on Euclidean space, which neglect the distribution of neighboring data points. While in many case, the high-dimensional data lies on or near a curved manifold, MDS may lead a mistake like in the Swiss roll dataset. The reason is that the MDS method consider tow data points as near points whereas the distance over the manifold is much large than the inner distance. Isomap resolves this problem by using a pairwise geodesic distance instead of Euclidean based distance.

D. Kernel PCA

Kernel PCA is a reformulation of linear PCA in a high-dimensional space constructed using kernel functions. Kernel PCA computes the principal eigenvectors of the kernel matrix rather than the covariance matrix in the origin PCA method. Apparently, constructing the kernel space transfer the linear based PCA into a nonlinear mapping.

E. Maximum Variance Unfolding (Semidefinite Embedding)

From the

IV. Material And Experiments

A. Data Collection

Acquiring and storing ECG data were the base for an analyzing task. Errors might creep into an analysis at any possible stage. Thus, not only the acquisition hardware system but also the transmission and storage should be carefully designed. As for the signal acquiring process, different kinds of sample rates might be involved, for common ECG acquisition device the sample rate would be 128Hz, 250Hz, 340Hz or 500Hz even higher. In murine studies, a sampling rate as high as 2kHz is considered sufficiently. Arbitrary resizing would be an ideal procedure to handle with the different sampling rate from a different data source to build the datasets for analysis, which would be adopted in the experiment to keep data consistency. Data Pre-processing Before the segmentation and feature extraction process, the ECG signals were preprocessed. As in the procedure of collecting ECG signals, in addition to the ECG signals, the baseline wander (caused by Perspiration, respiration and body movements), power line interference and muscle noise were recorded as well, which had been described in lots of literature [41]. When the filtering methods were proposed and adopted in the preprocessing, the desired information should not be altered. The ECG typically exhibits persistent features like P-QRS- T morphology and average RR interval, and non-stationary features like individual RR and QT intervals, long-term heart rate trends [22]. Possible distortions caused by filtering should be quantified in these features. The filtered ECG signals then were segmented into individual heartbeat wave- forms depends on the detected R peaks in a classification task. The ECG segmentation can be seen as the decoding procedure of an observation sequence regard- ing beat waveforms [42]. Dynamic time warping [43], time warping [44], Bayesian framework [45], hidden Markov models [42], weighted diagnostic distortion [46], morphology and heartbeat interval based methods [47] and genetic methods [48] had been used in this subtask. The state accuracies were close to 100%, which would be accurate enough in most online and offline applications.

B. ECG Arrhythmia

After the segmentation for the ECG records, we got plenty of ECG waveform sam- ples with variety categories. Since different physiological disorder may be reflected on the different type of abnormal heartbeat rhythms. For the task of classification, it is quite important to determine the classes would be used. In the early liter- ature, there were no unified class labels for an ECG classification problem. The MIT-BIH Arrhythmia Database was the first available set of standard test material for evaluation of arrhythmia detectors; it played an important role in stimulating manufacturers of arrhythmia analyzers to compete by objectively measurable per- formance. The annotations in the open database for the ECG categories adopted the ANSI/AAMI EC57: 1998/(R)2008 standard AAMI (2008), which recommended to group the heartbeats into five classes: on-ectopic beats (N as the Figure ?? (a)); supraventricular ectopic beat (S as the Figure ?? (b)); ventricular ectopic beat (V as the Figure ?? (c) and (d)); fusion of a V and a N (F); unknown beat type (Q). These classes or labels have been widely used in the ECG classification tasks re- lated literature (Table ?? illustrated). The normal beat, supraventricular ectopic beat and the ventricular ectopic beat categories were used much more frequently while the unknown beat type were abandoned because of its clinical valueless.

C. Data Source

Data from the ambulatory electrocardiography database were used in this study, which includes recordings of 100 subjects with arrhythmia along with normal sinus rhythm. The database contains 100 recordings, each containing a 3- lead 24-hour long electrocardiography which were bandpass filtered at 0.1-100Hz and sampled at 128Hz. In this study, only the lead I data were adapted after preprocessing in the classification task. The reference average heart beats for each sample has 97,855 beats for the 24-hour long recording, and the reference arrhythmia average is 1,810 beats which were estimated by a commercial software (this statistics aim to indicate the existence for arrhythmia samples, which should not be consider as a experiment preset). The MIT-BIH Arrhythmia Database [23] contains 48 half- hour recordings each containing two 30-min ECG lead signals (lead A and lead B), sampled at 360Hz. As well only the lead I data were used in the proposed method. In agreement with the AAMI recommended practice, the four recordings with paced beats were removed from the analysis. Five records randomly selected were used to verify the real time application. The remaining recordings were divided into two datasets, with small part of which were used as the training set of the fine-tuning process. The MIT-BIT Long-term Database is also used in this study for training and verification, which contains 7 long-term ECG recordings (14 to 22 hours each), with manually reviewed beat annotations and sampled at 128Hz. Similarly, the 7 recordings were divided into two datasets, with part used as the fine-tuning training set. A description of the labelled datasets are illustrated in Table 1 .

D. Data Preprocessing

Similar to the routine of electrocardiography classification task, the workflow con- sists of the stages of prepossessing, processing and the classifying. The prepossessing stage related technologies are not the focus of this study, so the classical methods for prepossessing were adapted, and just a brief introduction of the details would be mentioned. In the preprocessing stage, filtering algorithms were adapted to remove the arte- fact signals from the ECG signal. The signals include baseline wander, power line interference, and high-frequency noise. For the unlabelled database of ambulatory ECG and the MITBIH LT database, the Lead I data were extracted and a resample from 128Hz to 360Hz procedure was adopted for data consistency. Before the segmentation procedure from the long-time monitoring ECG signals- Heartbeat Detection: For the heartbeat detection, the MIT-BIH database and unlabelled database, the positions of R waves are determined. The provided fiducial points of R wave had been used as the basis of wave segmentation. The details of the implementation of R wave detection would not be described in this study, and a reference for the R wave detection algorithms had been explored in [58]. In the heartbeat segmentation process, the segmentation program of Laguna [59] was adapted, which also had been validated by other related work [2]. The segmentation process was focus on the Lead I of the recordings. After the segmentation for the ambulatory ECG database, three parts of heartbeat samples listed in Table 1 were acquired for the classification task. As for the pretraining, fine-tuning for our proposed task and comparison, we divided all the samples into three groups: the pretraining group as dataset A, the fine-tuning group as dataset B and test group as dataset C (illustrated in Table 2). Samples are chosen randomly from the original AR and LT database, the details of the sample class would be described in the experiment result analysis. As the algorithm of the deep structure training illustrated, both unsupervised learning and supervised learning are involved in the training process. The pre-training mainly used the unlabelled data to train the autoencoder parts, which only need to set the outputs equal to the inputs. Training data adopted in the unsupervised learning step include the whole samples from the ambulatory electrocardiography database and parts of the MIT-BHI database samples. In the supervised learning step, the MIT-BHI database samples with labels were adopted.

E. Data Embedding

1) PCA: : Consider the settings in Section III.A, we may have the following PCA algorithm 1

Algorithm 1 Principal Component Analysis

```
Input: Data Set $X = [x_1, x_2, \dots, x_n]^T$
Output: $y = x^n$
  $y \leftarrow 1$
```

2) Kernel PCA:

3) MDS:

4) Isomap:

F. Out-of-Sample Extending

G. Performance Assessment

V. Results and Discussion

Subsection text here.

VI. Conclusion and Future Work

In this paper, we explored an approach to improve shape-based 3D model retrieval performance through the distance measure that adapt to the database by means of learning based dimension reduction algorithm. Experimental comparison of efficacy of five dimension reduction algorithms showed that LE and LLE, both of which are local, non-linear dimension reduction methods, work better among the set of methods we tried. While the efficacy varies depending on the shape features, the unsupervised learning-based dimension reduction approach can be quite effective overall in improving retrieval performance of a wide range of shape features, especially if it is combined with the multiresolution shape comparison approach we have previously proposed [Ohbuchi03]. A possible future work would be an addition of relevance-feedback mechanism on the learned manifold to adapt to a users preference and/or a search occasion, in addition to the database to be searched for.

VII. Conclusion

The conclusion goes here.

Appendix A
Proof of the First Zonklar Equation

Appendix one text goes here.

Appendix B

Appendix two text goes here.

Acknowledgment

The authors would like to thank...

References

[1] H. Abdi and L. J. Williams, “Principal component analysis,” *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[2] J. Shlens, “A tutorial on principal component analysis,” *arXiv preprint arXiv:1404.1100*, 2014.

Yan Yan received the B.Eng. degree and the MSc in Instrument Engineering at the Harbin Institute of Technology in 2010 and 2012 respectively. He worked as research assistance at the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences from 2012 to 2014. He is working on his PhD in Computer Science started from 2014. His research interests are digital signal processing, machine learning, and control system.

Lei Wang received the B.Eng. degree in information and control engineering, and the Ph.D. degree in biomedical engineering from Xian Jiaotong University, Xian, China, in 1995 and 2000, respectively. He was with the University of Glasgow, Glasgow, U.K., and Imperial College London, London, U.K., from 2000 to 2008. He is currently a full professor with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. He has published over 200 scientific papers, authored four book chapters, and holds 60 patents. His current research interests include body sensor network, digital signal processing, and biomedical engineering