Project Requirement Task: The objective of the project is to find a good model for classifying ECG data. It is required that at least two classification methods should be used and performance comparisons between the methods should also be provided. Report: The report should cover the following contents: 1.Introduction to the methods used for classification 2.The procedure of data processing 3.How to employ model selection methods to find good models 4.Conclusions Data: The data for the project can be download from ftp://lqzhang:public@public.sjtu.edu.cn/StatisticalLearning2015/ . Submission Deadline: January 14th , 2016

Statistical Learning Project Report For ECG Disease Classification

苗富（1141209100）

PART I. OVERVIEW

Cardiovascular disease, including heart disease and stroke, remains the leading cause of death around the world. Yet most heart attacks and strokes could be prevented if some method of pre-monitoring and pre-diagnostic can be provided. The electrocardiogram (ECG) plays a key role in monitoring and preventing heart attacks. This project is about ECG pattern recognition.

The ECG consists of three basic waves: the P, QRS, and T(Fig. 1). These waves correspond to the far field induced by specific electrical phenomena on the cardiac surface, namely, the atrial depolarization (P wave), the ventricular depolarization (QRS complex), and the ventricular repolarization (T wave). ECG signal does not look the same in all the leads of the standard 12-lead system used in clinical practice. They usually change over different leads.

MIT-BIH Arrhythmia Database is widely used in ECG pattern recognition.(www.physionet.org/physiobank/databaselmitdb/). But this project, the dataset can be downloaded freely in here. (ftp://lqzhang:public@public.sjtu.edu.cn/StatisticalLearning2015/). The datasets consists with seven directory and for each directory there are some files, stored in mat file. Each mat file some number (about 20) matrix. Usually, the matrix is shaped with different length of columns and 12 rows. The 12 rows denoted one single heart beat time series in some lead. So a matrix is a single unit for a single heart beat. A mat file is several heartbeat for the same person. (Hao Zhang and Li-Qing Zhang, 2005) had noted that how they construct a new database from MIT-BIH Arrhythmia Database. And I adopt the same way of database construction.

- Use the first lead.
- For each array, there are some samplings from the original single heart beat. I sample 250 data-point time series with 100^{th} point as the maximum value of the original time series. And the left is uniformly sampled from the

original series.

- Make 7:3 division from the total dataset into train dataset and test dataset.

PART II. PCA AND SVM

Principal Components Analysis (PCA) is an exploratory multivariate statistical technique for simplifying complex data sets. Given n observations on m variables, the goal of PCA is to reduce the dimensionality of the data matrix by finding r new variables, where r is less than m. Principal components project high dimensional data into the subspace spanned by the eigenvectors with the r largest eigenvalues while remaining mutually uncorrelated and orthogonal. Each principal component is a linear combination of the original variables.

Support Vector Machine (SVM) is a widely used supervised learning algorithm. We start from the simple case of two linearly separable classes. We assume that we have a data set D ={ (xi, yi)} labeled examples, where yi takes value from set {-1, 1},and we wish to determine, among the infinite number of linear classifiers that separate the data, which one will have the smallest generalization error. Intuitively, a good choice is the hyperplane that leaves the maximum margin between the two classes, where the margin is defined as the sum of the distances of the hyperplane from the closest point of the two classes. If the two classes are non-separable we can still look for the hyperplane that maximizes the margin and that minimizes a quantity proportional to the number of misclassification errors. The trade off between margin and misclassification error is controlled by a positive constant C that has to be chosen beforehand.

Since it is unlikely that any real life problem can actually be solved by a linear classifier, the technique has to be extended in order to allow for non-linear decision surfaces. This is easily done by projecting the original set of variables x in a higher dimensional feature space. We call it as Reproduced Kernel Hilbert Space(RKHS). Though, the dimension space may be infinite and the project function can be very complex, a well-defined QP problem can still be efficiently solved. And we call this as “Kernel Trick”.

PART III. EXPERIMENTAL RESULT

A. Datasets

There are various ECG available databases available for researchers to use. We obtain our data set from MIT-BIH Arrhythmia. The data has been illustrated in detail in Part I.

Table 1.Dataset Class Distribution

Label Name	Number of Sample
Electrical axis left side	214
Left bundle branch block beat	226
Left ventricular hypertrophy	189
Normal	1076
Right bundle branch block beat	214
Sinus-bradycardia	153

Table 2.Basic Statistics Of Data

	1	2	3	4	…	248	249	250
count	2072	2072	2072.	2072	….	2072	2072	2072
mean	-0.0013	-0.0011	-0.0010	-0.0010	….	0.31648	0.34482	0.36712
std	0.01292	0.01283	0.01259	0.01251	….	0.19933	0.20421	0.20786
min	-0.1012	-0.0751	-0.0701	-0.0701	….	-0.0859	-0.0714	-0.0137
25%	-0.0060	-0.0061	-0.0061	-0.0061	….	0.16985	0.19618	0.21847
50%	-0.0012	-0.0009	-0.0009	-0.0009	….	0.27042	0.29923	0.32545
75%	0.00224	0.00233	0.00255	0.00257	….	0.44475	0.47495	0.50013
max	0.1324	0.1234	0.1261	0.1388	…	1.0555	1.0896	1.0896

B. Data Processing Using PCA

A great amount of data is not efficient to perform a pattern recognition process. The data set usually has noise that intervenes in the SVM training process. So we applied PCA to the data set. There are two motivations for PCA Preprocessing. One is that we need visualization to have a rough idea about the data. The other is that the size of data is 2072 and the dimension is 250, and we need dimension reduction to make the algorithm efficient, and we also need to denoise the original data.

The most significant component is listed in table3. The 20 principle components represent 97% of the total dataset energy. The others components have lower energy and are not significant to the signal composition. So we discard the low energy components with no significant loss in the information structure of the data set.

Table 3.Principle Component Distribution

Principle Component	1	2	3	4	5	6	7	8	9	10	20
Relative Contribution	0.615	0.113	0.078	0.034	0.028	0.019	0.018	0.010	0.009	0.007	0.002
Density Contribution	0.615	0.728	0.806	0.841	0.869	0.889	0.907	0.918	0.927	0.934	0.970

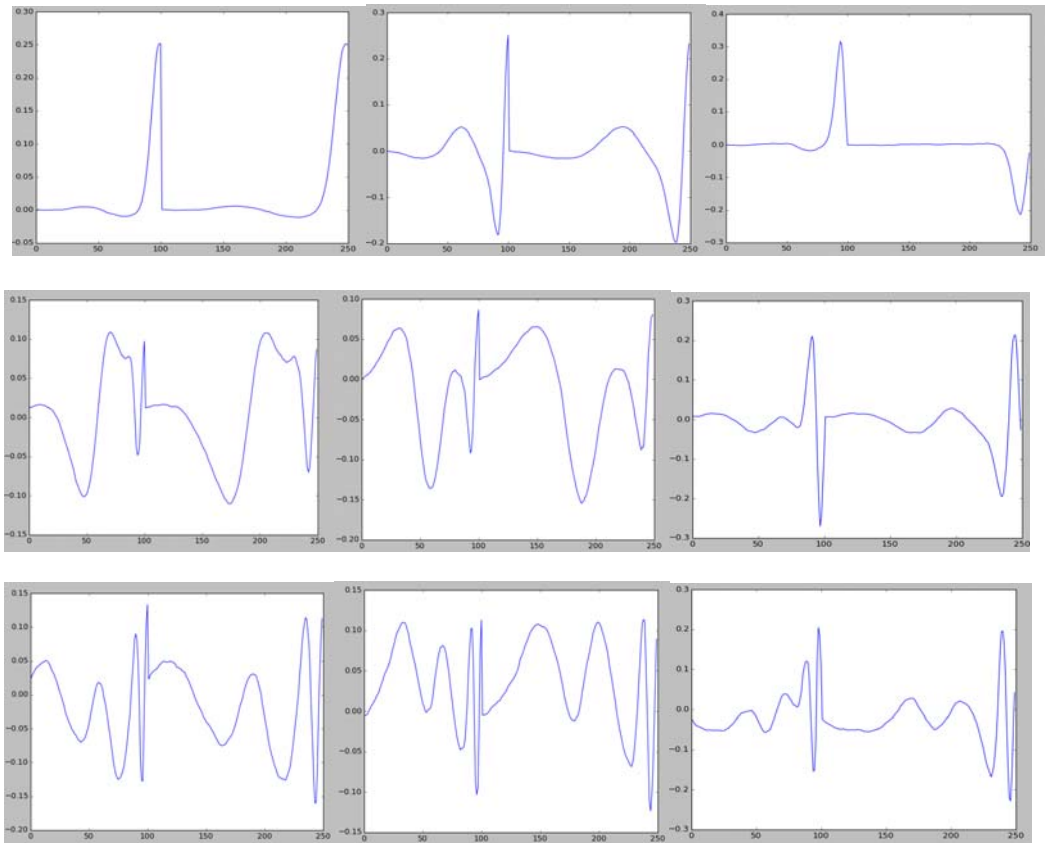

Fig1. The 9 components with highest energy

After the principal components are extracted, and the subspace is decided, we can project our data onto that subspace and thus reduce the dimensionality of our classification problem. We first project two types of data onto the space generated by the 3 most significant components.

It is widely accepted a heart beat of normal person should center around some point. If we use PCA dimension reduction technique, we can easily verify this idea. In Fig2, I plot the Five Abnormal heart beat point and the normal heart beat point respectively. From fig2, we can see that they are not linearly separated but also appear in some pattern. The center pattern gives us a intuitive motivation for using Gaussian Kernel Support Vector Machine to perform the classification.

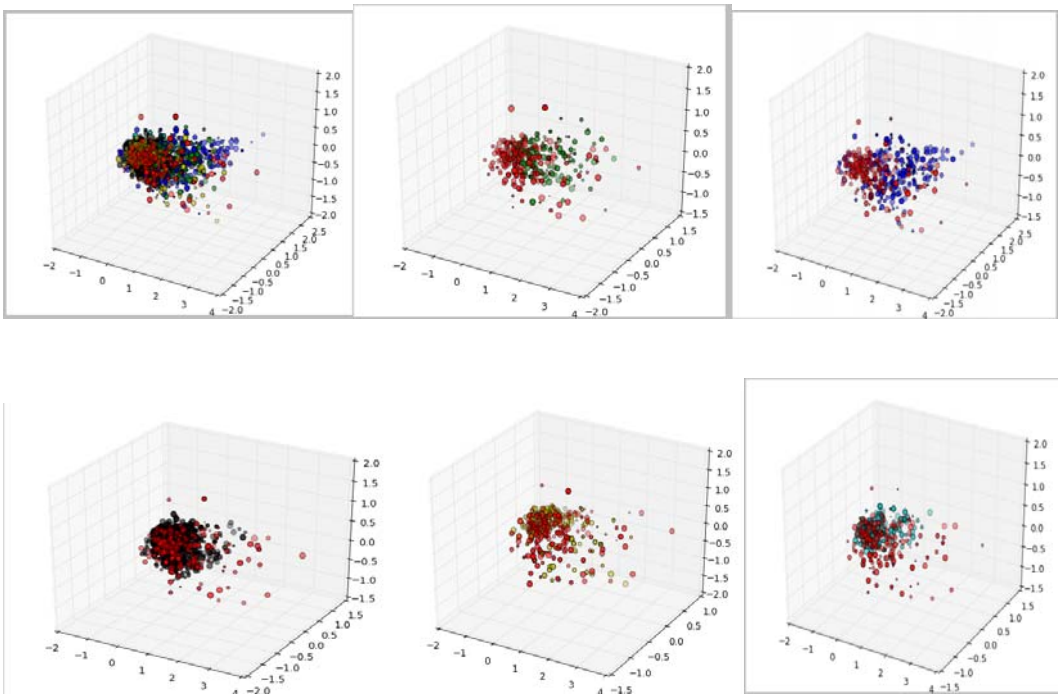

Fig2. The red point represent the component point of the Normal case, while the other represent other Abnormal case

C. Support Vector Machine

We choose RBF as the kernel function. Because the RBF kernel nonlinearly maps samples into a higher dimensional space, so it, unlike the linear kernel, can handle the case when the relation between class labels and attributes is nonlinear. We fix C at the value of 50. By specifying different value of gamma, we get the classification accuracy as in the Table 4,5.

Table 4. Classification Accuracy of SVM

	=0.5	=1	=1.5	=2	=2.5	=3
10 components	0.67	**0.73**	0.69	0.70	0.71	0.65
20 components	0.68	0.64	0.69	0.71	0.66	0.68
30 components	0.68	0.69	0.69	0.65	0.68	0.69

Table 5. Comparison of Classification Accuracy Between Different Model

	SVM+Gaussian Kernel	SVM+Linear Kernel	LR + L2	LR+ L1
10 components	**0.73**	0.56	0.58	0.54
20 components	0.64	0.61	0.56	0.55
30 components	0.69	0.55	0.60	0.59
40 components	0.67	0.62	0.63	0.61

IV. CONCLUSION

It is very important to identify arrhythmia and provide a method of machine pre-diagnosis. And in this project, I try some way of automatically recognize the disease type. And in our dataset, I obtain 73% accuracy in 6 Types disease recognition by use 10 principle component and Gaussian Kernel SVM. More advanced preprocess technique and model can be adopted to make further improvement.

The Python Code is in

https://github.com/miaofu/HeartDiseaseRecognitionFromECG/tree/master

The Data is in ftp://lqzhang:public@public.sjtu.edu.cn/StatisticalLearning2015/

V. REFERNCE

[1]. Zhang H, Zhang L Q. ECG analysis based on PCA and support vector machines[C]//Neural Networks and Brain, 2005. ICNN&B'05. International Conference on. IEEE, 2005, 2: 743-747.

[2]. Zhao Q, Zhang L. ECG feature extraction and classification using wavelet transform and support vector machines[C]//Neural Networks and Brain, 2005. ICNN&B'05. International Conference on. IEEE, 2005, 2: 1089-1092.

[3]. Jiang X, Zhang L, Zhao Q, et al. ECG arrhythmias recognition system based on independent component analysis feature extraction[C]//TENCON 2006. 2006 IEEE Region 10 Conference. IEEE, 2006: 1-4.