Learning Systems
DA8008 (HT12) Halmstad University

Classification model for Electrochardiograms

Mingkun Yang, Yuantao Fan

1 Introduction

This report is focus on Constructing a model To tell if a patient suffers from Transmural Ischemia (TI) or not. The input data is based on the signal from a 12 channel electrochardiogram (ECG).

Data given includes 300 observation which is from 12 ECG channel called V1, V2, V3, V4, V5, V6, aVL, I, -aVRII, aVF, and III. Each channel has 26 feature which are denoted I dur, Q dur, R dur, S dur, Rp dur, Sp dur, K dur, I ampl, Q ampl, R ampl, S ampl, Rp ampl, Sp ampl, K ampl, QRS area, QRS dur, Tmaxampl, Tminampl, timeTmax, timeTmin, ST ampl0, ST ampl20, ST ampl40, ST ampl60, ST ampl80, ST ampl100. The 'dur' features are the durations between different parts of the ECG and the 'ampl' are the amplitudes at these points. The physicians believes that the most important features are from 19-26 channel.

In this report, Finisher index for each variable is calculated. Variable is ranked according to fisher index. Linear classifier including Gaussian linear classifier and logistic regression classifier is constructed by inputting Features of importance. Multi-layer perception model is set up by using the best input of kNN-classifier. 1 to 10 node is tested, MLP input variable is improved by using forward selection and backward elimination. Important data, model and performance is present in the result.

2 Methodology

In this section, Method implemented for model construct is presented, including Data transformation, Fisher Index estimation, cross-validation, Gaussian linear classifier, logistic regression classifier, k-nearest neighbour classifier, multilayer perceptron learning using the best result from kNN classifier. Forward selection and backward elimination.

2.1 Fisher Index

Calculate Fisher Index is a useful method to rank the importance of data. In this case, Fisher index is calculated on every feature. The Fisher Index can be define as:

$$FI(k) = \frac{(\mu_{k,1} - \mu_{k,2})^2}{(N_1 - 1)\sigma_{k,1}^2 + (N_2 - 1)\sigma_{k,2}^2}$$

The feature has higher fisher index means that it's a feature of importance and it's good for classification.

2.2 Cross-validation

Cross-validation is applied to improve the ability of generalization. In this case, 10 fold validation is implement for classify improvement and model estimation. The original sample is randomly partitioned into 10 subsamples. Feature of importance is selected and is packaged. Then a combination of these feature is used, it's similar to backward elimination that every combination of the original set is tested, and thus compare the MSE of every combination. Pick the best combination at last.

2.3 Linear Classifier

In this case, Gaussian linear classifier and logistic regression classifier is used for this problem, feature with importance according to Fisher index is used as input of the model. Cross-validation and combination package is applied to improve the generalize ability and estimation the performance.

2.4 k-nearest Neighbour classifier

kNN classifier is implemented by inputting important feature, using cross validation and combination package. In this case, 1,3,5 of k value is tested, the model with best performance is chosen to compare with other classify model.

2.5 Multilayer perceptron model

MLP model is constructed by using the best input of kNN classifier, hidden node from 1 to 10 is tested. Forward selection and backward elimination is applied to find the best input combination. The best model is selected to compare the performance with other classifier.

3 Data

In this data is quite complicated, cause there is 12 channel, each of them has 26 feature. The 26 feature can be denoted I dur, Q dur, R dur, S dur, Rp dur, Sp dur, K dur, I ampl, Q ampl, R ampl, S ampl, Rp ampl, Sp ampl, K ampl, QRS area, QRS dur, Tmaxampl, Tminampl, timeTmax, timeTmin, ST ampl0, ST ampl20. The most important feature believed by physicians are #19-26 feature of each channel.

First, scatter the 'Dur' data with 'ampl' data, but it seems there is no significant difference between two classes, figure 3.1 shows that the 'dur' with 'ampl' data in channel 1-4.

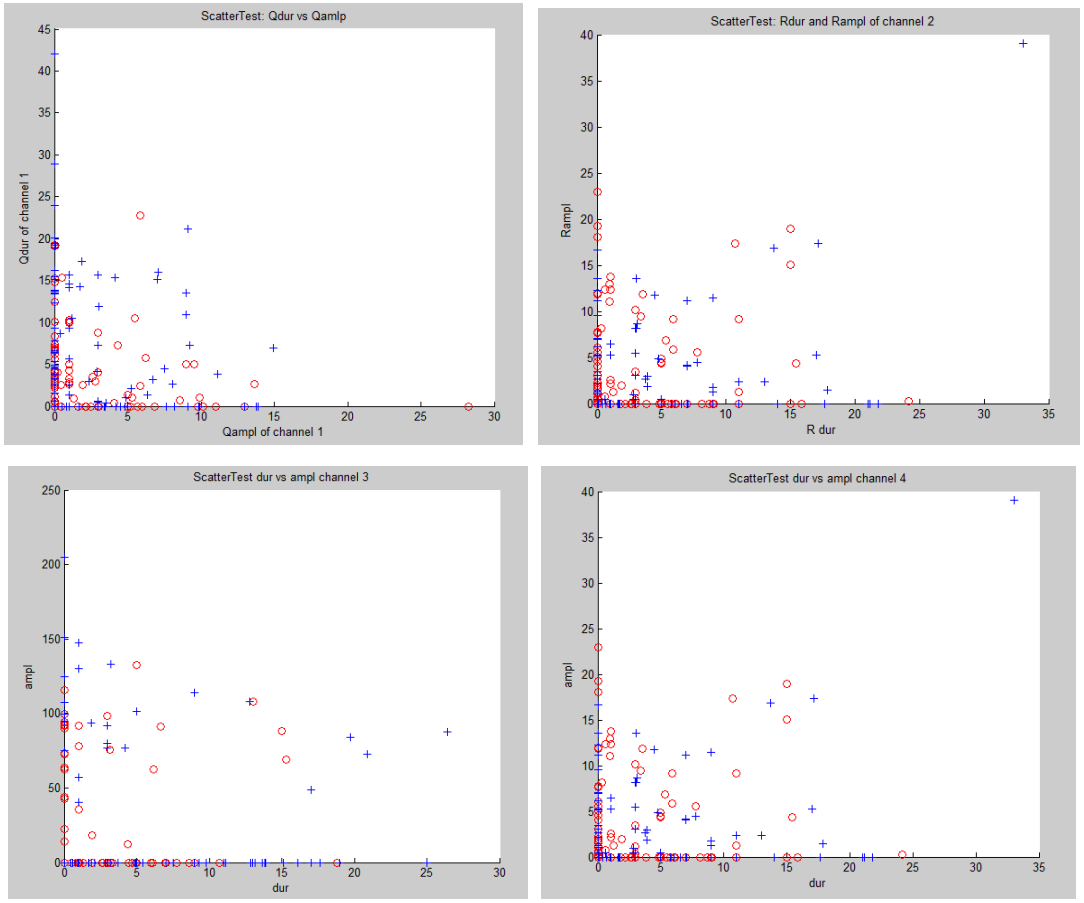

Figure 3.1 'dur' vs 'ampl' in channel 1-4

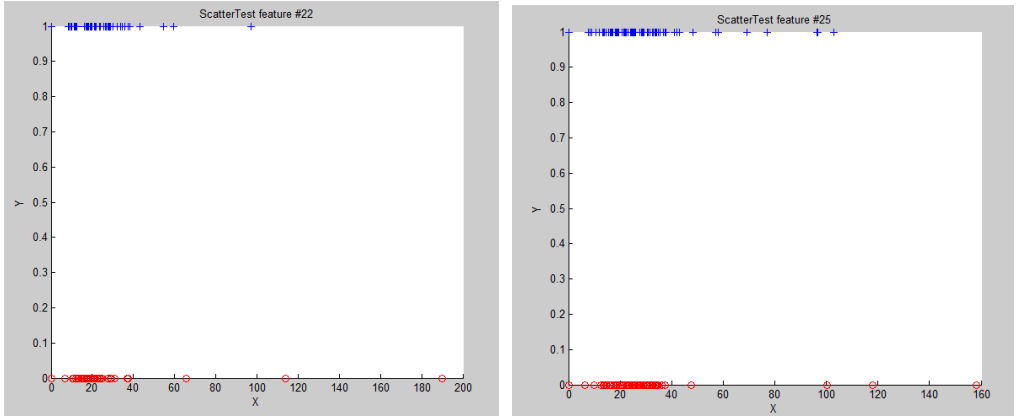

Figure 3.2 Feature recommend by physicians #22 and #25 of channel 1

These scatters are quite chaos. The feature recommend by physicians seems also chaos. So Fisher Index is calculated in order to find feature with great importance. The feature with high rank will use as our model's input.

4 Result

In this section, Result of Fisher Index, Gaussian linear classifier, logistic regression classifier, k-nearest neighbour classifier, multilayer perceptron learning method is presented.

4.1 Fisher Index

Fisher index is calculated for searching the best feature of the classify problem. Feature #202, #310, #298, #286, #203, #299, #311, #274 is ranked as high 8 on the Fisher index list. These feature will be used as input for several classifier.

4.2 Linear Classifier

In this case, Gaussian linear classifier and logistic regression classifier is used for this problem, feature ranked as high 8 is used for input. Cross-validation and combination package is applied to improve the generalize ability and estimation the performance.
For Gaussian linear classifier, the best MSE using 10 fold validation is 0.2600 when using input feature #298 and #274, which means that 26% of the points are misclassified. Figure 4.1 shows the sample result.

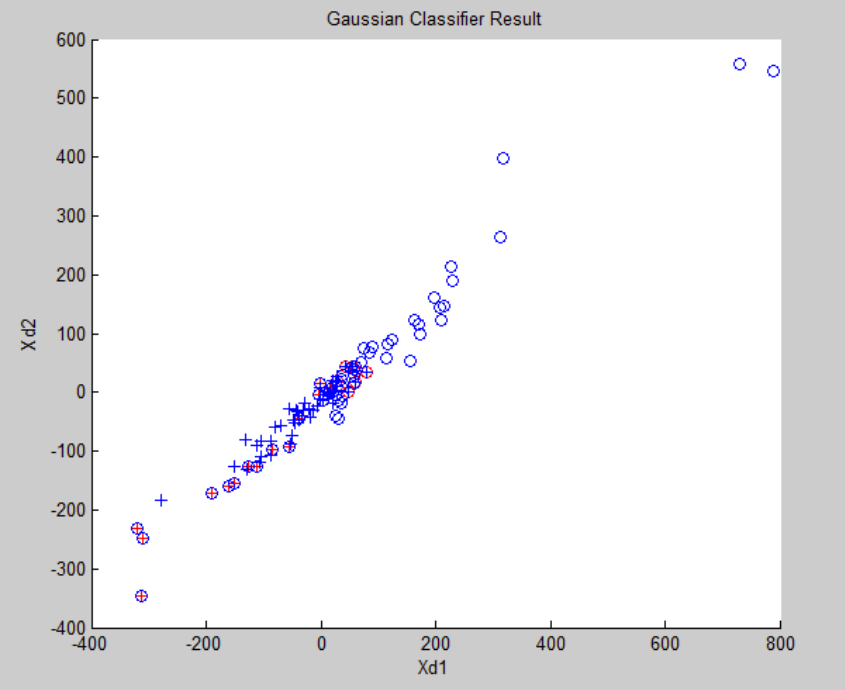

Figure 4.1 Gaussian Classifier Result

In figure 4.1, the points that red plus and blue circle with the same location is the misclassification point.
For Logistic regression classifier, The best MSE is 0.26 while input feature is #202

#203 #299 #274, the result is shown in figure 4.2.

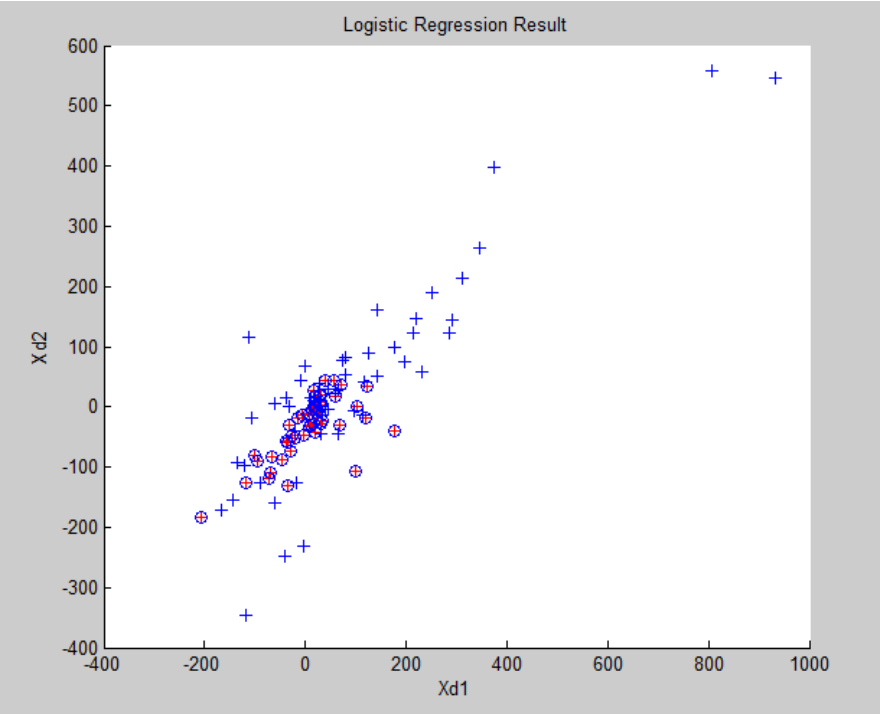

Figure 4.2 Logistic regression Classifier Result

4.3 k-nearest Neighbour classifier

kNN classifier is implemented by inputting important feature, using cross validation and combination package. In this case, 1, 3, 5, 7, 10 of k value is tested. For the best result, it k = 10, the best MSE is 0.2450 while using input #310. The result is shown in figure 4.3.

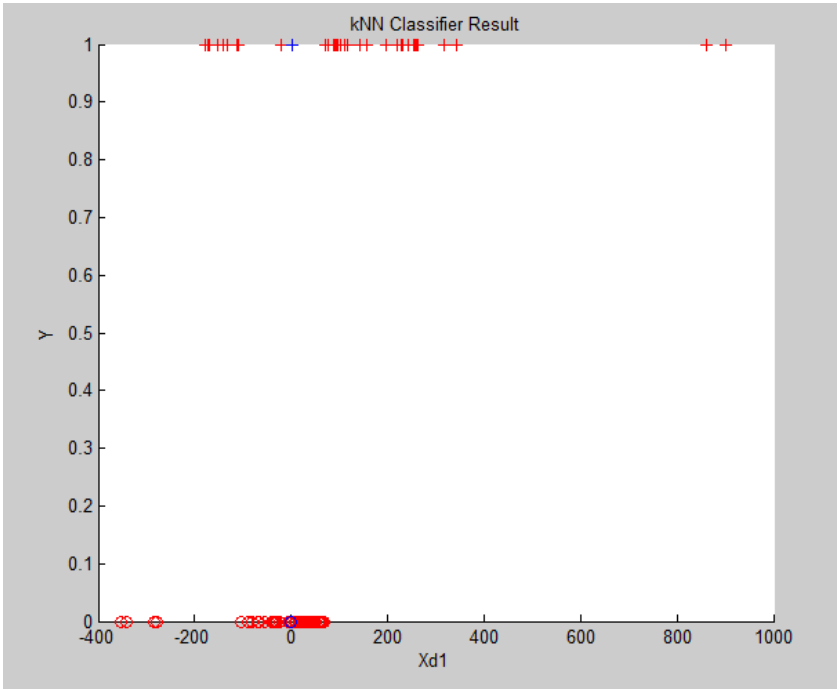

Figure 4.3 kNN Classifier Result

2.4 Multilayer perceptron model

MLP model is constructed by using the best input of kNN classifier, hidden node from 1 to 10 is tested. Forward selection and backward elimination is applied to find the best input combination. For kNN classifier, feature #310 is the best input. According to the requirement, we put feature #310 as our MLP's input and MSE = 0.2960, the result shows in figure 4.4.

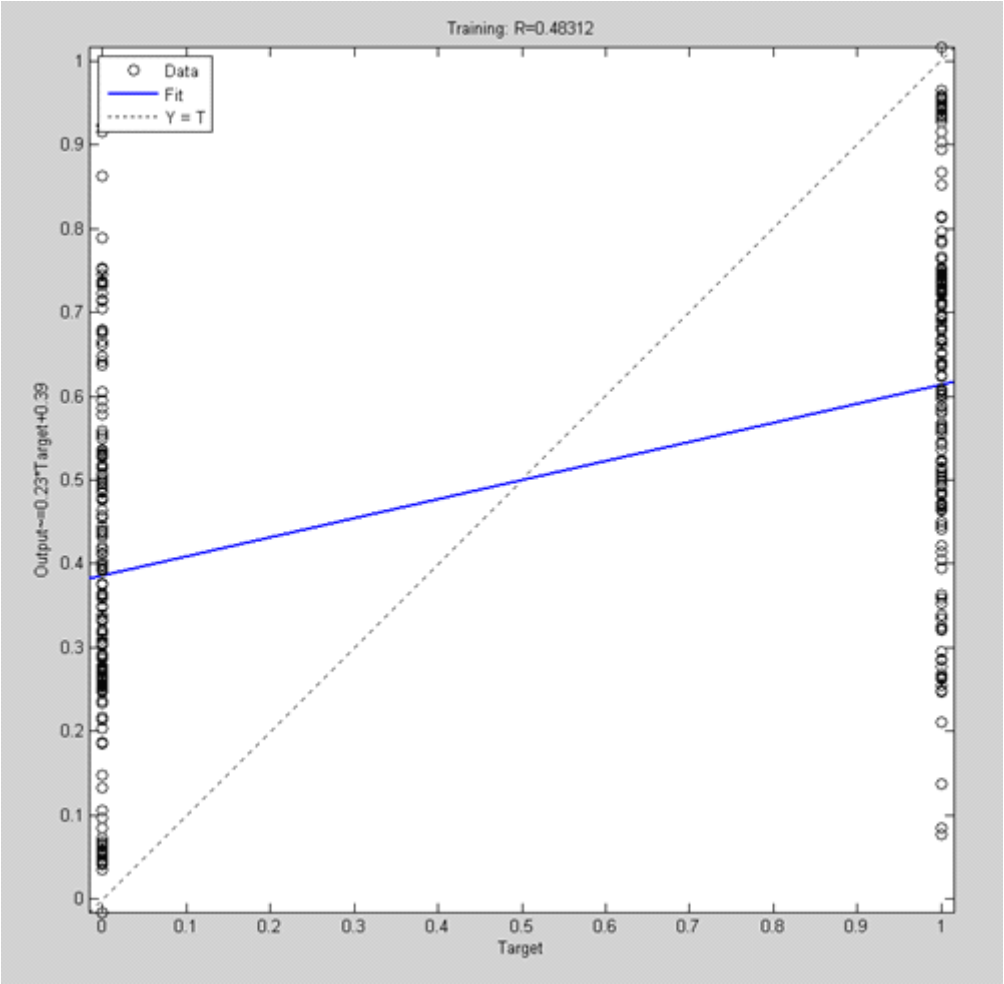

figure 4.4 Performance of MLP using feature #310

1 to 10 node is tested, and 10 nodes has the lowest MSE. Then we use the feature ranked as high 5 according to Fisher index, applied with cross validation and package. The result shows that choose input feature #310 and #203 has the best performance, MSE = 0.2650, the result shows in figure 4.5. We choose this model to be our best MLP model.

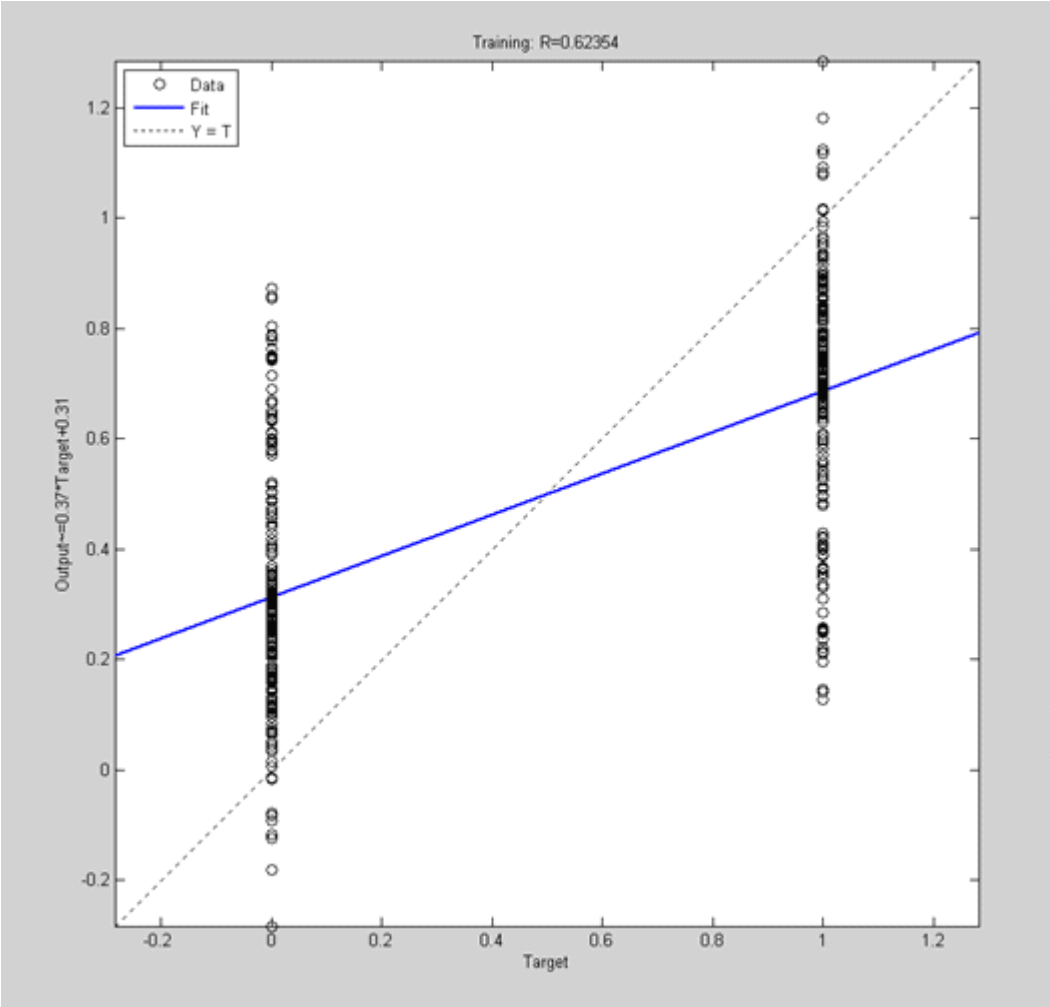

figure 4.5 Performance of MLP using feature #310 and #203