Different Techniques used to Improve the Performance of a Classifier of the Twelve-Lead Electrocardiogram

P de Chazal

University College Dublin, Dublin, Ireland
University of New South Wales, Sydney, Australia

Abstract

This study investigated the automatic classification of the Twelve-lead electrocardiogram (ECG) into different pathophysiological disease categories. The ECG database used in this study contained 926 recordings with 500 records classified with 100% accuracy and 426 records classified with 75% accuracy. Each record contained a simultaneously recorded 12-lead ECG of 8-10 seconds duration. Each record is classed as normal; left, right or biventricular hypertrophy; anterior, inferior or combined myocardial infarction. A baseline classifier was trained using a single beat from the 500 classified recordings and resulted in a 7-way classification test-set accuracy of 55%. The following techniques were used for improving the classification performance: 1) multi-beat data, 2) regularisation of the covariance matrix and, 3) utilisation of inaccurately classified data in the training process. Combining these three techniques resulted in a classifier with a test-set accuracy of over 75%.

1. Introduction

The classification of the electrocardiogram (ECG) into different pathophysiological disease categories is a complex pattern recognition task. Computer based classification of the ECG can achieve high accuracy and offers the potential of affordable mass screening for cardiac abnormalities. Successful classification is achieved by finding patterns in the ECG that discriminate effectively between the required diagnostic categories. Conventionally, a typical heart beat is identified from the ECG and the QRS, T and possibly P waves are characterised using features such as magnitude, duration and area. Classification is then achieved on the basis of these measurements. Measurements based on QRS, T and P sections can vary enormously even among normals and can lead to misclassification.

In this study, we take a simple feature set – samples of the ECG signal – and look at optimizing the performance of a linear discriminant based classifier. An advantage of this representation is that the approximate QRS detection point is the only cardiac characteristic point required. By eliminating the need to find other characteristic points a significant amount of computation is saved.

The following techniques were trialed for improving the classification performance: 1) use of multi-beat data, 2) optimisation of performance by regularisation of the covariance matrix and, 3) the utilisation of the inaccurately classified data in the training process.

A database of modest size was employed hence a cross-validation scheme was used to estimate the performance of the different techniques.

2. Methods

The ECG database used throughout this study contains 926 records that have been classified into seven classes: normal (NOR); left (LVH), right (RVH) and biventricular (BVH) hypertrophy; anterior (AMI) inferior (IMI) and combined (MIX) infarction. This 100% accurate validation was based on ECG independent clinical information, consisting of data derived from cardiac catheterisation, coronary angiography, left ventriculography, echocardiography and results from coronary care and cardiac surgery. In addition, all records have been classified by a committee of computer programs and the accuracy of these classifications was determined to be 75%. We do not have the 100% accurate classifications of the full data set. Instead, we have these classifications for 500 of the records and the 75% accurate classifications for the other 426 records.

The 500 fully classified records contain 155 NOR and 345 abnormal cases. The abnormal cases comprise 79 LVH, 21 RVH, 25 BVH, 77 AMI, 111 IMI and 32 MIX cases. All records contained between eight to ten seconds of digitally sampled (500 Hz sampling rate) data from simultaneously recorded twelve-lead ECGs. The twelve-lead ECG contains four redundant four leads. Because of the redundancy in the twelve-lead ECG set we chose to process leads I, II, and V1 to V6 as these leads represent a non-redundant set.

0276-6547/01 $17.00 © 2001 IEEE

Computers in Cardiology 2001;28:525-528.

2.1. ECG pre-processing

The ECG is filtered with a 0.5 - 40 Hz linear phase digital bandpass filter to remove unwanted baseline drift and powerline interference. QRS complexes were detected with a multi-lead detector [1].

2.2. Feature set

A feature set was obtained for this study directly from the sample values of each ECG lead. After bandpass filtering and QRS detection [1], samples were obtained for each heart beat from the ECG using two data windows. The first data window extended from 150ms before the QRS detection point to 150ms after the detection point. Within this data window the ECG was resampled at 80 Hz to provide 24 features. The second data window extended from 150ms after the QRS detection point to 600ms after the detection point. Within this data window the ECG was resampled at 20 Hz to provide 9 features. A total of 33 features were used to represent each of the eight leads. In addition age and sex were included in the feature set. Thus, for each QRS detection point a feature set was generated containing 266 features.

2.3. Inaccurately classified data

The 426 records with inaccurate classifications were utilised in the training of the classifier. These records were not included in the testing phase.

2.4. Classification performance estimation

When developing a classifier it is important to be able to estimate the expected performance of the classifier on data not used in training. The available data must be divided into independent training and testing sets. There are a number of schemes for achieving this and the most suitable for the size of data set used in this study, is *n*-fold cross validation [2,3]. This scheme randomly divides the available data into *n* approximately equal size and mutually exclusive "folds". For an *n*-fold cross validation run, *n* classifiers are trained with a different fold used each time as the testing-set, while the other *n*-1 folds are used for the training data. The choice of *n* influences the ratio of data used for training/testing with an optimal value of *n* in the range 5-20. Cross validation estimates are generally pessimistically biased, as training is performed using a subsample of the available data.

The randomising process was "stratified" so that all the folds contained the same relative proportions of normals and the six disease conditions. Studies have shown that stratification of the folds decreases both the bias and the variance of the performance estimate [3].

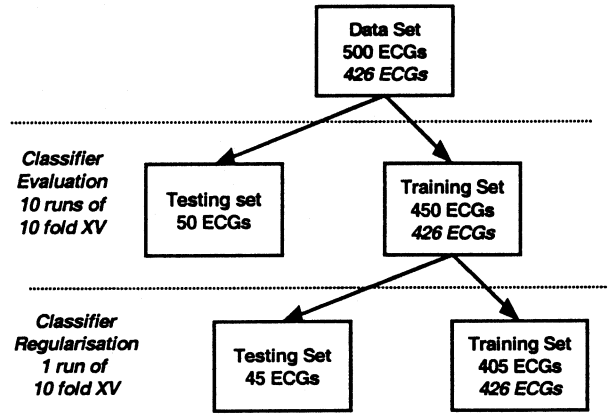

Figure 1: Splitting of the data using the cross-validation scheme. A double loop of cross-validation is utilised. See text for details. The italicised entries in some of the boxes indicate the inclusion of the 426 ECG records classified with 75% accuracy.

Cross validation estimates are highly variable and depend on the division of the data into folds. A decrease in the variance of the performance estimate may be achieved by averaging results from multiple runs of cross validation where a different random split of the training data into folds is used for each run. For this study ten runs of ten-fold cross validation were employed. Figure 1 depicts how the data was split for this study.

In this study we report the overall classification accuracy and the individual class sensitivities. The overall accuracy is the percentage of total cases correctly classified. A class sensitivity is the percentage of cases correctly classified of that class. The specificity is the sensitivity of the normal class.

2.5. Classifier

Linear discriminants were used as the classifier model for this study. They provide a parametric approximation to Bayes rule [2], so in response to a set of input features the output of the classifier is a set of numbers representing the probability estimate of each class. The final classification is obtained by choosing the class with the highest probability estimate. Linear discriminants partition the feature space into the different classes using a set of hyper-planes. Optimisation of the model is achieved through direct calculation and is extremely fast relative to other models such as neural networks.

Technique										
Multi-Beat Data	Partially Accurate Data	Covariance Regularisation	Acc(%)	Sensitivities (%) NOR	LVH	RVH	BVH	AMI	IMI	MIX
			55.2	74	49	15	24	51	60	23
x			69.6	90	62	20	33	70	75	33
	x		69.6	93	53	24	28	72	78	23
		x	72.5	94	57	36	43	71	78	38
x	x		72.6	94	59	30	39	73	79	33
x		x	75.4	95	63	44	47	70	82	43
	x	x	72.8	95	56	42	48	71	77	33
x	x	x	74.4	95	58	42	51	72	81	38

Table 1: The test-set overall accuracy performance and class sensitivities for the different combinations of performance enhancing techniques. All results are percentages and determined from ten runs of ten-fold cross validation.

		True Classification						
		NOR	LVH	RVH	BVH	AMI	IMI	MIX
Predicted Classification	NOR	95	1	1	0	0	3	0
	LVH	6	63	2	7	5	13	5
	RVH	30	0	44	21	0	5	0
	BVH	16	20	13	47	4	0	0
	AMI	11	6	0	1	70	4	8
	IMI	8	7	0	1	1	82	1
	MIX	1	2	4	2	36	13	43

Table 2: The confusion matrix of the best performing classifier. All results are percentages and determined from ten runs of ten-fold cross validation.

2.6. Regularisation

The performance of a classifier can often be improved by reducing the effective number of parameters of the model [2]. For linear discriminants this can be achieved by shrinking the covariance matrix ($\boldsymbol{\Sigma}$) towards the identity matrix $\mathbf{I}$ using

$$\boldsymbol{\Sigma}(\alpha)=(1-\alpha)\boldsymbol{\Sigma}+\alpha\mathbf{I} \qquad 0\le\alpha\le 1.$$

This is only appropriate if the training data has been rescaled so the variance of each feature is equal to one. When $\alpha=1$ then $\boldsymbol{\Sigma}(\alpha)=\mathbf{I}$ which results in a special form of LDA where the features are assumed to be statistically independent (and hence no covariance). In practice, various values of α in the range 0 to 1 are trialed and the classifier performance determined. The value of α that optimises the performance is chosen. In this study we have used the classification accuracy as the performance measure. The value of α was chosen to optimise the test-set accuracy determined from a single run of ten-fold cross-validation.

Optimisation of the classifier parameters required two nested loops of cross-validation to obtain unbiased estimates of classifier performance [3]. The inner loop was used to measure the performance of the classifier using different values α. Although these performance figures used test-set values, these values are optimistically biased as they were used during selection of α. The outer cross-validation loop was used to test the classifier configuration that was selected by the inner loop (see Figure 1).

2.7. Multi-beat classification

For multi-beat classification of an ECG record, the classifier processes the feature information of each beat separately and finds a set of probabilities for each beat. To obtain the final classification, the probabilities for each class are averaged across the beats and the class with the highest average probability estimate chosen.

During the training phase, feature data is obtained from each beat and treated as separate training examples. By using diagnostic information from all beats, more efficient use of the available ECG diagnostic information is made.

3. Results and discussion

Table 1 shows the overall accuracy and class sensitivities for the different combinations of performance enhancing techniques. The baseline classifier did not use any performance enhancing techniques and resulted in an overall cross-validated test set accuracy of 55.2%. The training set accuracy for this classifier was 99.7% and this clearly demonstrated that the classifier was overfitting the training data. This was expected as the number of input features (266) was large compared to the number of training cases (450).

All of the performance enhancing techniques significantly improved the classifier performance both in terms of overall accuracy and individual class sensitivities. Using multi-beat data boosted classification accuracy to 69.8%. On average there were 9.7 beats of data per record and thus much more data was used in estimating the classifier parameters. Inclusion of the 426 partially accurate records in the training data resulted in an accuracy of 69.8%. Regularisation of the covariance matrix resulted in an accuracy of 72.5%. Using multi-beat data in conjuction with the partially accurate data produced an accuracy of 72.6%. This was an improvement on the result when using either technique alone.

The best performing classifier resulted from using multi-beat data and co-variance regularisation and the accuracy was 75.4%. The specificity was 95% and the class sensitivities for the hypertrophy classes were 63%, 44% and 47% for LVH, RVH and BVH respectively. For the myocardial infarction classes the sensitivities were 70%, 82% and 43% for AMI, IMI and MIX respectively. The poor result for RVH, BVH and MIX classes is probably due to the fact that these are the smallest classes in our database.

The best performing classifier using single beat data achieved an accuracy of 72.8%. Both partially accurate data and covariance regularisation were utilised to derive this classifier. Using all techniques simultaneously resulted in accuracy of 74.4%.

Table 2 shows the confusion matrix of the best performing classifier. This provides insight into how the classes are being misclassified. From a practical viewpoint the most significant information is the number of disease classes classified as normals. This is given by the off-diagonal entries in the 3rd column (headed 'NOR') of the table. For example the RVH class is misclassified as NOR 30% of the time; other poorly performing classes include the BVH and AMI classes which are misclassified as NOR 16% and 11% respectively. Other notable results are the misclassification of the MIX class. Most of the misclassification are as IMI cases (13%) and AMI (36%) cases. These misclassifications are partially correct as all of these classes are types of myocardial infarction. A similar partially correct result is seen among the hypertrophy classes. The BVH class is misclassified as RVH cases and LVH cases at a rate of 13% and 20% respectively. The RVH class is misclassified as BVH cases 21% of the time.

Other authors [4,5,6] have attempted a similar ECG classification task using other ECG databases. Overall accuracy results vary between 66.3% and 77.4%, but because of the different proportion of classes in their databases a direct comparison of overall accuracy is not possible. Nevertheless, the results achieved in this project are favourable.

4. Conclusion

We compared the classification performance of different techniques used to enhance the performance of a classifier of the twelve-lead ECG. A simple feature set using samples values of the eight independent ECG leads was used and linear discriminants used for classification. The best performing combination of techniques was a regularised classifier processing multi-beat ECG data and resulted in an overall accuracy of 75.4%.

The final structure for the proposed classifier is very computationally efficient and easily lends itself to real-time implementation. After detection of each R-wave, a linear discriminant classifier processes the ECG samples. A classification is found for each heart beat and the final classification found by combining the individual classifications.

References

[1] de Chazal P, Celler BG. Automatic Measurement of the QRS Onset and Offset in Individual ECG Leads. In: IEEE Engineering in Medicine and Biology Conference 1996. IEEE Computer Society Press. 1996.

[2] Ripley BD. Pattern Recognition and Neural Networks. Cambridge University Press. 1996

[3] R. Kohavi. A study of cross validation and bootstrap for accuracy estimation and model selection. In: 14th Int. Joint Conference on Artificial Intelligence 1995:1137-1143.

[4] Willems JL, Lesaffre E. Comparison of multigroup logistic and linear discriminant ECG and VCG classification. Journal of Electrocardiology 1987; vol. 20(2):83-92.

[5] Willems J. Comparison of Diagnostic Results of ECG Computer Programs and Cardiologists. In: Computers in Cardiology 1992. Los Alamitos: IEEE Computer Society Press 1992:93-96.

[6] Bortolan G. Willems JL. Diagnostic ECG classification based on neural networks. Journal of Electrocardiology, 1993;vol. 26(Suppl):75-79.

Address for correspondence.

Philip de Chazal
Department of Electronic and Electrical Engineering,
University College Dublin, Dublin 4, IRELAND
philip@ee.ucd.ie