A Comparison of the ECG Classification Performance of Different Feature Sets

P de Chazal, RB Reilly

University College Dublin, Dublin, Ireland

Abstract

This study investigates the automatic classification of the Frank lead ECG into different disease categories. A comparison of the performance of a number of different feature sets is presented. The feature sets considered include wavelet-based features, standard cardiology features, and features taken directly from time-domain samples of the ECG. The classification performance of each feature set was optimised using automatic feature selection and choosing the best classifier model from linear, quadratic and logistic discriminants. The ECG database used contains 500 cases classed into seven categories with 100% confidence. Using multiple runs of ten-fold cross-validation, the overall seven-way accuracy of different feature sets and classifier model combinations ranged between 60% and 75%. The best performing classifier used linear discriminants processing selected time-domain features. This is also found to be the simplest and fastest classifier to implement.

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 measurements such as magnitude, duration and area. Classification is then achieved on the basis of these measurements.

Alternative representations of the diagnostic information of the ECG offer a number of advantages. Previous studies [1] have shown that it is possible to classify using features extracted from the wavelet transform of ECG signals and achieve comparable diagnostic accuracy to the standard cardiology features. 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.

In this study a system was established for evaluating the diagnostic ability of feature sets. For each feature set, the classification performance was optimised by automatic feature selection and choosing the best classifier model from linear, quadratic and logistic discriminants.

The feature sets evaluated include wavelet-based features, standard cardiology features, and features taken directly from time-domain samples of the ECG. Both single- and multi-beat classifications were considered.

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

2. Methods

For this study the Frank lead ECG was used.

2.1. ECG pre-processing

The ECG is sampled at 500 Hz then 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 [2].

2.2. Feature sets

Five feature sets were used in this study and derived using different techniques. All the sets had the age and sex of the subject as common members.

Wavelet transform (WT): this feature set was derived from the coefficients of the discrete wavelet transform and the methodology has been described previously [1]. Briefly, for all detected QRS complexes a data window containing the P-QRS-T complexes was isolated using the ECG samples in the range 200ms before the R-wave maximum points to 400ms after the R-wave maximas. The isopotential value was subtracted, and the data window multiplied with a Hanning window. A seven-level discrete wavelet decomposition of each data window was

0276-6547/00 $10.00 © 2000 IEEE

Computers in Cardiology 2000;27:327-330.

calculated using the Haar wavelet. We previously found that the wavelet choice and decomposition level had little influence on the classification accuracy [1]. The signal information of details 1 and 2 were discarded, as the frequencies covered by these levels were higher than frequency content of the ECG. There were 76 features per lead, hence including the age and sex measures there were 230 features per beat.

Standard cardiology features (CARD): This feature set was derived from standard QRS features and is described fully in [3]. For each QRS detection, the associated QRS onset and offset was determined. Features were then derived from the scalar leads e.g. QRS duration; vector loops e.g. XY area; and 3-D loop e.g. planarity of the QRS plane. In all, there were 229 features in this set.

Time-domain samples (TD): This feature set was derived directly from the sample values of each ECG lead. After bandpass filtering, the ECG was resampled at 80Hz and samples in the range 400ms before the R-wave maximum points to 600ms after the R-wave maximum were obtained. There were 254 features in this set.

Features were generated for a representative beat of each ECG for all feature sets. In addition, two more feature sets were generated using all beats of each ECG for the wavelet (WVm) and time-domain (TDm) features sets. This facilitated comparison of single-beat versus multi-beat classification.

2.3. Classifiers

A supervised training technique was used to derive all classifiers. In supervised training, a classifier model that maps the input features to the required output classes is chosen. The model has a set of adjustable parameters that are optimised using training data. For this study four classifier models were considered. All of the models provide a parametric approximation to Bayes rule [4], so in response to a set of input features the output of each 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. In the following analysis d represents the number of input features and c the number of classes.

Linear discriminants (LDA) 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. The number of parameters in the model is $(d+2c+1)*d/2$.

Quadratic discriminants (QDA) provide a generalisation of linear discriminants and partition the feature space using a set of hyper-quadratics. Again, optimisation of the model is achieved through direct calculation and is extremely fast. The number of parameters is significantly more than linear discriminants and is $c*d*(d+3)/2$. For large training sets this model will outperform linear discriminants but it is frequently outperformed by LDA for smaller sets.

Both linear and quadratic discriminants assume the feature data has a Guassian distribution for each class.

Linear-logistic discriminants (lin-LOG) impose fewer conditions on the feature space partitioning than the linear/quadratic discriminants. The model assumes the feature data has a class distribution belonging to one of the family of exponential distributions. This family includes many of the common distributions such as the Gaussian, binomial, Bernoulli and Poisson as special cases. Direct optimisation of the model parameters is not possible and an iterative numerical optimisation technique is required. The number of parameters to optimise is $(d+1)*(c-1)$. The number of classes is generally less than the number of features. Thus, for the classifier models considered here, this model has the smallest number of parameters.

Nonlinear-logistic discriminants (nl-LOG): Similar to linear-logistic models except even greater flexibility in the feature space partitioning. When implemented as a neural network (see below) the number of parameters to optimise is $(d+1)*h+(h+1)*(c-1)$ where h is the number of hidden units. The number of hidden units controls the flexibility of the feature space partitioning.

Both of the logistic discriminant models were implemented with feed-forward neural networks. A softmax output stage was used and the (negative) log-likelihood error function minimised. Hidden units were used in the non-linear logistic model. Optimisation of the parameters (weights) of both networks was achieved with a gradient-descent algorithm with an adaptive learning rate and momentum constant. Training was stopped when the successive iterations no longer resulted in a significant reduction in the error function. The weights of hidden units were optimised with the back-propogation algorithm.

Although the logistic discrimination models impose fewer conditions on the feature partitioning, in practice, linear discriminants perform as effectively for ECG classification [5].

2.4. Feature selection

The performance of most classifier training algorithms is degraded when one or more of the available features are redundant or irrelevant. Redundant features occur when two or more features are correlated whereas irrelevant features do not separate the classes to any useful degree. The classification performance of a given set of features may often be improved by searching for a subset of the features with higher performance. Finding this optimal subset is generally computationally intractable for anything apart from small feature sets. This is because the number of possible subsets rises exponentially with size of the feature set. In practice a sub-optimal heuristic search such as stepwise procedure is used [4]. A stepwise procedure for feature selection was used in this study.

	Testing Performance				Training Performance			
Feature Set	LDA	QDA	lin-LOG	nl-LOG	LDA	QDA	lin-LOG	nl-LOG
CARD	69.3	60.2	68.3	60.3	76.5	85.5	78.9	99.8
WV	71.1	63.8	70.8	64.2	78.8	88.3	82.7	99.8
TD	71.2	63.3	70.6	64.7	78.0	88.5	81.3	99.9
WVm	73.1	67.3	73.0	68.0	78.3	83.9	80.6	97.0
TDm	74.7	67.9	74.7	69.5	79.1	84.5	81.7	97.8
Free Parameters	*272*	*1190*	*108*	*390*				

Table 1: The overall testing and training set accuracies, derived from ten runs of ten-fold cross validation, of combinations of feature sets and classifier models. See sections 2.2 and 2.3 for abbreviations. The top 3 rows of results derive from classifying using a single beat from the ECG. Rows 4 and 5 are derived from multi-beat classifications. The italicised row shows the number of free parameters in each classifier model.

When comparing the subsets, the best performance measure to use is the classification performance but again computational restrictions prevent this being implemented. We have used Wilk's Lambda, which is a measure of class separation to measure the performance of the subsets. A low value of Wilk's Lambda indicates good separation of the classes and indicates probable high classification performance. Hence feature selection involves finding a subset with the lowest value of Wilk's Lambda.

2.5. 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.

2.6. 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 [6]. 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 [6].

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.

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.

2.7. Implementation

The work for this project was performed on a 300 MHz Pentium II PC running MATLAB version 5.3. All algorithms for feature selection, classifier training and data partitioning have been developed inhouse. Approximately ten minutes of processing time was required to perform cross-validation, feature selection and to train linear discriminant classifiers for the multi-beat feature sets.

3. Results

The ECG database used throughout this study contains 500 records with 155 normal (NOR) and 345 abnormal cases. The classification of every record is known with 100% certainty based on ECG independent clinical information. The abnormal cases comprise left (LVH), right (RVH) and bi- (BVH) ventricular hypertrophy; and anterior (AMI), inferior (IMI) and combined (MIX) myocardial infarction. The numbers of each class are shown the bottom row of Table 2. Each case contained

		Sensitivities						
Feature Set	Acc	NOR	LVH	RVH	BVH	AMI	IMI	MIX
Conv	69.3	83	58	40	42	69	78	39
WV	71.1	89	63	24	42	68	82	25
TD	71.2	89	62	27	39	69	81	33
WVm	73.1	90	66	21	39	72	83	36
TDm	74.7	90	72	29	42	73	83	38
Size		*155*	*79*	*21*	*25*	*77*	*111*	*32*

Table 2: The test-set overall accuracy performance and class sensitivities for feature sets with the best (LDA) classifier. The italicised row shows the size of each class.

between eight and ten seconds (approximately ten heart beats) of digitally sampled data from simultaneously recorded Frank lead ECGs.

The feature sets were processed in an identical way. For every run of every fold of cross-validation, automatic feature selection was applied to each feature set and 17 features were identified that maximised the classification performance. These features were then used to train the classifier and the classification performance on the testing-set determined. The nl-LOG models used 16 hidden units.

Table 1 shows the overall accuracy results for classifying each feature set with the different classifier models. Both the testing- and training-set results are shown. The LDA (69.3-74.7%) model was consistently the best performing classifier for all feature sets. It just outperformed the lin-LOG model (68.3-74.7%) and outperformed the QDA (60.2-67.9%) and nl-LOG (60.3-69.3%) by up to 9%. The extra flexibility of the QDA and the nl-LOG was seen in the higher training-set results (84.5-88.5%, 97.0-99.9% respectively) compared to the other models but this performance did not generalise well to unseen test data.

Table 2 shows overall testing-set accuracy and class sensitivities of the best classifier (LDA) on each feature set. The single-beat feature set results show that the wavelet (71.1%) and time-domain sets (71.2%) outperformed the cardiology set (69.3%). The multi-beat time-domain set was the best classifier (74.7%) and outperformed the equivalent wavelet set (73.1%) and both outperformed the single-beat sets by at least 1.9%.

The specificity of the best classifier was 90%. The sensitivities of the large abnormal classes, LVH, AMI and IMI were 72%, 73% and 83% respectively. The smaller abnormal classes, RVH, BVH and MIX didn't classify as well; the respective sensitivities were 29%, 42% and 38%.

Other authors [5,7,8] 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 ECG classification performance of different feature sets using different classifier models. The best performing combination was a linear discriminant classifier processing selected sample values of the ECG.

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 selected 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, Reilly R. Using Wavelet Coefficients for the Classification of the Electrocardiogram. In: Proceedings of World Congress on Medical Physics and Biomedical Engineering 2000:4 pages.

[2] de Chazal P, Celler BG. Automatic Measurement of the QRS Onset and Offset in Individual ECG Leads. In: Proceedings of the IEEE Engineering in Medicine and Biology Conference 1996. IEEE Computer Society Press. 1996:2 pages.

[3] de Chazal P, Celler BG. Selection of Optimal Parameters for ECG Diagnostic Classification. In: Computers in Cardiology. Piscataway: IEEE Computer Society Press. 1997:13-16.

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

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

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

[7] 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.

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

Address for correspondence.

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