A NOVEL, BATCH MODULAR LEARNING APPROACH FOR ECG BEAT CLASSIFICATION

Vijay P. Mani, Yu Hen Hu and Surekha Palreddy

Electrical and Computer Engineering Department
1415 Engineering Drive
Madison, WI 53706
hu@engr.wisc.edu

ABSTRACT

In this paper, we investigate a modular architecture for ECG beat classification. The feature space is divided into distinct regions and individual classifiers are developed for each region. We compare different combination strategies, and feature space partition strategies. We also describe a novel, batch modular learning method that can be used to incrementally improve the performance of the modular network.

1. INTRODUCTION

An ECG recording consists of a sequence of spiky beats each representing one contraction of the heart. By analyzing the type of the ECG beats, and the accompanying rhythms, a trained electro-cardiologist can diagnose probable causes of anomalies in the patient's heart.

ECG beat classification is a difficult pattern classification problem. The difficulties stem from many factors, including large dimension of the feature space, large amounts of the training samples, significant overlap between class boundaries and the ever-changing morphology with time.

In this paper we use modular architecture to distinguish normal heartbeats from those of premature ventricular contraction (PVC) beats.

Modularity is a manifestation of the principle of divide and conquers, so that we can solve a complex computational task by dividing it into simpler subtasks and combining their individual outputs. In modular architecture, multiple classifier modules are developed in parallel; each dedicated to classify a portion of the entire feature space. Then, an integrating unit is developed to combine the output of all modules to make a final decision. In communication, this technique is known as *sensor fusion* [23], and in machine learning, as *stack generalization* [20]. In artificial neural network paradigm terminology, this approach is known as *mixture of experts*[8], [9], [21], or *committee classifiers* [2], [4], [12], [17], [18].

A modular network offers several advantages over a simple neural network in terms of providing better performance and facilitating parallel learning.

(a) *Better Classification Performance*: By partitioning the feature space into localized regions, each module will have a simpler classification task than a monolithic classifier, and therefore a potential to offer better performance.

(b) *Parallel Learning*: Modular learning allows for parallel processing because individual modules are trained independently. This is extremely important when the size of training set becomes prohibitively large.

In this work, we focus on developing a modular network with multiple modules to handle large-scale ECG pattern classification. Our approach differs from our previous approach [7] in that we separate the development of individual modules and the integration unit into two separate phases of training. In the first phase, the feature space is partitioned into smaller regions by the integration unit, and a modular classifier is developed on each region. In the second phase, based on the outcome of each classifier, the integration unit is updated and the feature space is re-partitioned. This process is repeated until the performance saturates.

In this paper, we report experimental results comparing different combination methods which include the winner decides all method, and several fuzzy combination methods. We have also experimented with different partitioning methods, including random and clustering based approaches for the integration unit. Each modular classifier is realized with a Learning Vector Quantization (LVQ) [11].

Annotated ECG records from the MIT/BIH arrhythmia database are used for the experimentation.

2. MODULAR LEARNING

2.1 The Basic Approach

A basic modular network is shown in Figure 1. The individual classifiers can be of the same type or can be of different types. Each feature vector is presented to all modules. The gating network, (a.k.a. integration unit, or fusion center), whose inputs are also the feature vector, determines the K module classifier which should be given the responsibility to classify the present feature vector. The gating network's output is assigned to 1, for that classifier and 0 for the other classifiers. This way, the gating network functions like a classifier. For each input feature vector, it determines which classifier's output is to be passed and blocks the output from the other classifiers.

2.2 Partitioning Methods

In the mixture of expert network architecture, the overall output is a weighted sum of individual classifier's output:

$$z = \sum_{i=1}^{n} w(x,i)\,y(i) \qquad (1)$$

For pattern classification problem, we assume $0 \le y(i) \le 1$ and the target value $t \in \{0,1\}$. In [7], we proved the following theorem, which is stated here without proof:

Theorem 1. Denote $\Re_i$ to be the region in the feature space R such that the i^{th} modular classifier gives correct classification. Then the region C which the combined output is making correct classification is bounded by $\cup \Re_i$ — the union of $\Re_i$.

Clearly, the maximum benefit of a modular network can only be realized if all these $\Re_i$ are disjoint regions in the feature space. If one specific modular classifier is trained with training data exclusively within a region in the feature space, it is more likely that this module will correctly classify features within this region. Based on this observation, in this paper, we experiment with two partitioning methods. The first randomly partitions the feature space into equal parts. The second method uses the SOM_PAK [11] to generate cluster centers and the feature space is crisply partitioned based on these cluster centers. Individual LVQs are developed for each subspace.

2.3 Combining Multiple Classifiers

The outputs are combined with a weighting factor provided by the gating network (the integration unit). Below we show that a winner-takes-all strategy is optimal.

Theorem 2. Let $z,t \in \{0,1\}$, $0 \le y(i), w(x,i) \le 1$, and

$$\sum_{i=1}^{n} w(x,i) = 1 \qquad (2)$$

The solution to the constrained minimization problem:

$$\underline{Min}_{w} |t - z| = \underline{Min}_{w} |t - \sum_{i=1}^{n} w(x,i)\, y(i)| \qquad (3)$$

subject to the constraints $0 \le w(x,i) \le 1$, and $\sum_{i=1}^{n} w(x,i) = 1$, is

$$w(x,i) = \begin{cases} 1 & |t - y(i)| < |t - y(k)|, i \ne k; \\ 0 & Otherwise. \end{cases} \qquad (4)$$

Proof: Suppose that $|t-y(i)| < |t- y(k)|$ for $k \ne i$. If $w(x,i) < 1$, because the sum of $w(x,i)$ over all i equals to 1, and $w(x,i) \ge 0$, there must be at least an $i' \ne i$, such that

$$1-w(x,i) \ge w(x,i') > 0.$$

First, consider the case $t=1$ and $w(x,i) < 1$. Note that $1-y(i) < 1-y(k)$ implies $y(i) > y(k)$ for $k \ne i$. Now choose i' such that $y(i') > y(k)$ for $i', k \in \{k \mid k \ne i, w(x,k) > 0\}$. Then, $y(i) > w(x,i)y(i) + w(x,i')y(i') \ge z$. This leads to,

$$|t-z| = 1-z \ge 1-w(x,i)y(i)-w(x,i')y(i') > 1 - y(i)$$

In order words, $|t-z|$ is NOT minimized. Similarly, if $t=0$, then $y(i) < y(k)$ for $k \ne i$. Choose i' such that $y(i') < y(k)$ for $i', k \in \{k \mid k \ne i, w(x,k) > 0\}$. Then, $y(i) < w(x,i)y(i) + w(x,i')y(i') \le z$. Hence

$$|t-z| = z-0 > y(i) = y(i)-0$$

Again, $|t-z|$ is NOT minimized. Therefore, to minimize $|t-z|$, one must have $w(x,i) = 1$ for $y(i) > y(k)$, $i \ne k$. Q.E.D.

Theorem 1 guarantees the optimality of the *winner-takes-all* combination rule in a modular network as it minimizes the output error of the integration unit. Note that although we use the absolute error in the theorem, the same results can be proved using other norms of the error *t-z*.

In the past, many different combination criteria have been proposed for ensemble of classifiers. In this paper, we will experiment with a few of them and compare the results.

While the theorem 2 is optimal, it is not directly applicable to modular pattern classifiers having binary output $y(i) \in \{0,1\}$. In this paper, we adopt a remedy by weighing the output of the modular classifier by a scaling factor

$$S(x,i) = K/f(d(x,i))$$

where $d(x,i) = ||\mathbf{x}-c(i)||^2$ is the Euclidean distance between $c(i)$, the cluster center of the disjoint region within which the i-th modular classifier is trained, and **x**, the present testing feature vector. Three different choices of the function f() have been compared in this paper:

(a) Inverse distance: $S(x,i) = K/d(x,i)$

(b) Winner decides all: $S(x,i^*) = 1$, if $d(x,i^*) < d(x,k)$, $k \ne i^*$;

$= 0$, otherwise.

(c) Power method: $S(x,i) = K'/[d(x,i)]^{\mu}$, $1 < \mu < \infty$.

K and K' are scaling constant, chosen to satisfy eq. (2). The underlying heuristic in using $S(x,i)$ is that a modular classifier would in general give more accurate classification result when a sample is closer to the clustering center where most of its training samples locate. Note that $0 \le S(x,i)y(i) \le 1$. Hence, theorem 2 can be applied to choose the optimal weighting factor.

3. EXPERIMENTS AND RESULTS

3.1 ECG Data and Feature Vectors

The annotated ECG records from the MIT/BIH arrhythmia database [14] have been used in this study. This database has 48 records, each 30 minutes in length. The data were recorded in two channels (modified limb lead II and modified lead VI) of surface ECGs from long-term Holter recorders. They represent a variety of waveforms, artifacts, complex ventricular, junctional, and supraventricular arrhythmias, and conduction abnormalities. Data from 33 of the 48 records which contain normal beats and PVCs were used for this study. Classifiers were developed and evaluated using subsets of data from channel 1 of these 33 records sampled at 360 Hz.

Accompanying each record in the database is an annotation file in which each ECG beat has been identified by expert cardiologists. These labels, referred to as 'truth' annotations, are used to develop the classifiers and to evaluate the performance of the classifiers in the testing phase. Data is extracted in the form of feature vectors. Each feature vector has 9 elements. The first four feature elements are temporal parameters. The temporal features are the R–R interval between the current beat and the previous beat (RR1), between the previous beat and the one before it (RR0), between the current beat and the next beat (RR2), and the ratio of RR1 and RR2. These features are extracted for each individual beat in the database. A ratio of RR1 to RR0 provides an indication of an abnormal timing sequence and helps in identifying an abnormal beat. The next 5 feature elements are extracted based on morphology. The 'truth' annotations are appended to the feature vectors. Detailed descriptions of these features can be found in [16].

3.2 Experiment

We use the three-way cross validation method to improve the reliability of the results. The original data set is partitioned randomly into three subsets. We combine two of the subsets to generate the training data set, and use the third as the testing set. The subsets are rotated to yield three training-testing data set pairs. These results are averaged to obtain the final results.

To compare the performance of the multiple classifier approach, we conduct a base line experiment using a single LVQ to classify the entire feature space (the monolithic classifier).

The number of subspaces is empirically chosen to be five. The default parameters of the LVQ_PAK(ver 3.1) are used to generate the classifiers. The number of codebook vectors is chosen at 0.5% of the number of feature vectors in the set. It is observed that the performance curve tends to saturate with increasing values of the number of code book vectors and the saturation starts at 0.5% in this case. This method of choosing the number of codebook vectors for each LVQ, is not universal. It is highly subjective and is dependent on the distribution of the feature vectors in the feature space. To generalize the process we use the novel, batch modular learning method, to incrementally improve the performance of the modular network. The LVQs in this case are set to have a fixed number of codebook vectors. The incremental training moves the region boundaries to improve the classification results.

The majority rule is used to combine the individual outputs when the partitioning is done randomly. We use $\mu = 50.0$ for the power method. A Normal beat takes a class index of 0 and a PVC beat takes a class index of 1

3.3 Results Reporting

A two-class ECG beat classification problem can be regarded as a hypothesis testing problem with:

Null Hypothesis $\mathbf{H_0}$: The Beat is Normal

Alternate Hypothesis $\mathbf{H_a}$: The Beat is PVC

We compute both type I and type II errors from the experiments:

Type I Error (α) — probability of rejecting the Null hypothesis when it is true, also known as the false alarm rate

Type II Error (β)— probability of accepting the Null Hypothesis when it is false, also known as the miss ratio.

These error measures are tightly related to the AAMI reporting protocol which requires the statistics of the following:

Actual \ Classified	PVC Beat	Normal Beat
PVC Beat	TP(true positive)	FN(false negative)
Normal Beat	FP(false positive)	TN(true negative)

TP: Number of PVC beats classified correctly.
TN: Number of Normal beats classified correctly.
FP: Number of Normal beats classified as PVC beats.
FN: Number of PVC beats classified as Normal Beats.
Specifically,

$$\alpha = FP/(TP+TN+FP+FN)$$
$$\beta = FN/(TP+TN+FP+FN)$$

According to AAMI recommendation, s***ensitivity*** is the fraction of real events that are correctly detected. That is

$$Sensitivity\ (Se) = TP/(TP+FN)$$

Specificity is the fraction of non-events correctly rejected.

$$Specificity\ (Sp) = TN/(TN+FP)$$

We can re-write α and β as follows

$$\alpha = (1\text{-}Sp)*P_n;\ \beta = (1\text{-}Se)*P_p$$

Where, $Pn = (FP+TN)/(TP+TN+FP+FN)$ is the prior probability of observing a normal beat; and

$P_p = (FN+TP)/(TP+TN+FP+FN)$ is the prior probability of a PVC beat.

The values of sensitivity and specificity must ideally tend to one and the errors must be as close to zero as possible. However there is often a trade-off between the sensitivity and specificity, resulting in a similar trade-off between the type I and II errors.

3.3 Results

The results are summarized below.

Method	TP	FN	TN	FP	Se	Sp	α	β
Mono	5627	898	59892	5358	86.23	91.79	0.1251	0.0074
Eucli	4279	2246	59107	6143	65.58	90.59	0.3129	0.0086
Winner	5620	905	62460	2790	86.13	95.72	0.1261	0.0039
Power	5547	978	62643	2607	85.01	96.00	0.1363	0.0032
Rand	5926	599	55527	9723	90.82	85.10	0.0835	0.0135

Table 1. Using 0.5% for the number of codebook vectors

Method	TP	FN	TN	FP	Se	Sp	α	β
Eucli	1973	201	19138	2612	90.73	87.99	0.0843	0.0109
Winner	1858	317	20858	891	85.42	95.90	0.1325	0.0037
Power	1858	316	20858	892	85.44	95.90	0.1324	0.0037

Table 2. Using the novel, batch modular learning method.

5. SUMMARY

The results show that the modular approach yields better performance characteristics as compared to the monolithic case. Among the partitioning methods the clustering approach gives a higher specificity although we lose out a bit on the sensitivity. The novel, batch modular learning method gives comparable results and is preferred due to its universal applicability. Figure 2 shows the iterative improvement in performance for the three combination methods considered. The tendency of the classifier to do better on the specificity is due to the higher number of Normal beats, as compared to PVC beats, in the training data set. The number of Normal beats is ten times that of the PVC beats.

4. REFERENCES

[1] Bortolan G., Degani R., and Willems J. L., "Design of neural networks for classification of electrocardiographic

signals," *Proc. Annual Intl. Conf. IEEE Eng. Med. & Biol. Soc.*, no. pp. 1467–1468, 1990.

[2] Drucker, H., C. Cortes, L. Jackel, Y. LeCun, and V. Vapnik, "Boosting and other ensemble methods," *Neural Computation*, vol. 6, no. 6, pp. 1289-1301, 1994.

[3] Habboush, I., G. B. Moody, and R. G. Mark, "Neural networks for ECG compression and classification," in Proc. *Proceedings. Computers in Cardiology*, pp. 185-188, 1991.

[4] Hansen, L. K., and P. Salamon, "Neural network ensembles," *IEEE Trans. on PAMI*, vol. 12, no. 10, pp. 993-1001, 1990.

[5] Haykin S. *Neural Networks, Macmillan, 1994.*

[6] Hu Y. H., Tomkins W. J., Urrusti J.L., and Alfonso V.X. "Application of artificial neural networks for ECG signal detection and classification". Journal of Electrocardiology, 1994.

[7] Hu, Y. H., Palreddy, S., and W. J. Tompkins, "Patient Adaptable ECG Beat Classification using Mixture of Experts," in *Neural Network for Signal Processing V*, Ed(s)., IEEE, 1995.

[8] Jacobs, R. A., M. I. Jordan, S. Nowlan, and G. E. Hinton, "Adaptive mistures of local experts," *Neural Computation*, vol. 3, no. pp. 79-87, 1991.

[9] Jordan, M. I., and R. A. Jacobs, "Hierarchical mixtures of experts and the EM algorithm," *Neural Computation*, no. pp. 1993.

[10] Klingeman, J., and Pipberger, H. V., "Computer classification of electrocardiograms," *Comp. Biomed. Res.*, vol. 1, no. pp. 1, 1967.

[11] Kohonen, T., "The Self-Organizing Map," *Proc. IEEE*, vol. 78, no. 9, pp. 1464-1480, 1990.

[12] Krogh, A., and J. Vedelsby, "Neural network ensembles, cross validation and active learning," in *Advnaces in Neural Information Processing Systems 7*, Ed(s)., Cambridge MA: MIT Press, 1995.

[13] Linnenbank, A. C., Groenewegen, A. S., and Grimbergen, C. A.,, "Artificial neural networks applied in multiple lead electrocardiography: Rapid quantitative classification of ventricular tachycardia QRS integral pattern," *Proc. Annual Intl. Conf. IEEE Eng. Med. & Biol. Soc.*, no. pp. 1461-1462, 1990.

[14] Mark, R., and Moody R, "MIT-BIH Arrhythmia Database Directory", MIT, 1988.

[15] Mark R. and Wallen R. "AAMI Recommended Practice: Testing and Reporting Performance Results of Ventricular Arrhythmia Detection Algorithms". Association for the Advancement of Medical Instrumentation, AAMI ECAR-1987, 1987.

[16] Palreddy S. "ECG Beat Classification". Ph.D. Dissertation, University of Wisconsin-Madison, 1996.

[17] Seung, H. S., M. Opper, and H. Sompolinsky, "Query by committee," in *Proc. 5-th workshop on computational learning theory*, Ed(s)., San Mateo, CA: Morgan kaufmann, pp. 287-294, 1992.

[18] Tresp, V., and M. Taniguchi, "Combining estimators using non-constant weighting functions," in *Advances in Neural Information Processing Systems 7*, Ed(s)., Cambridge MA: MIT Press, 1995.

[19] Tsai, Y. S., Hung, B. N., and Tung, S. F., "An experiment on ECG classification using back-propagation neural network," *Proc. Annual Intl. Conf. IEEE Eng. Med. & Biol. Soc.*, no. pp. 1463-1464, 1990.

[20] Varshney, P. K., *Distributed Detection and Data Fusion*, Springer-Verlag, NY, 1997.

[21] Wolpert, D. H., "Stacked Generalization," *Neural Networks*, vol. 5, no. 2, pp. 241-259, 1992.

[22] Xu, L., and M. I. Jordan, "EM learning on a generalized finite mixture model for combining multiple classifiers," in Proc. *Proc. World Congress on Neural Networks*, Portland, OR, vol. IV, 1993.

[23] Yeap, T. H., Johnson, F., and Rachniowski, "ECG beat classification by a neural network," *Proc. Annu. Int'l Conf. IEEE Eng. Med. & Biol. Soc.*, no. pp. 1457–1458, 1990.

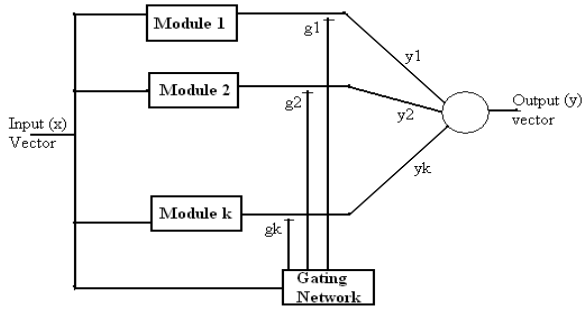

Figure 1. Modular Network

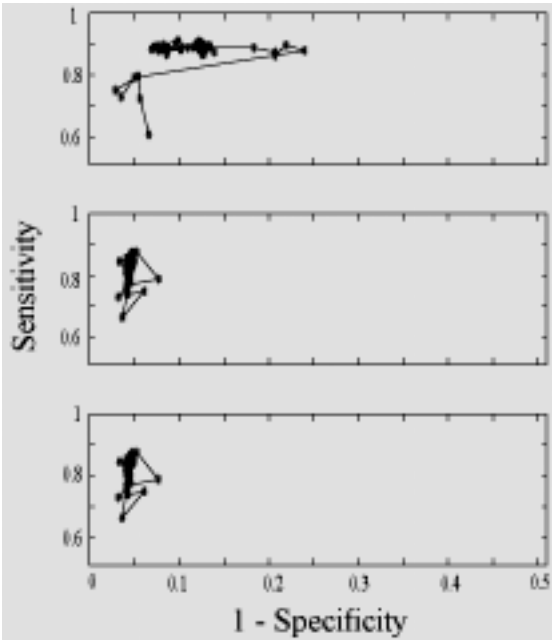

Figure 2. Performance curves for the Euclidean Distance, Winner Decides All and the Power Cases.