Modelling T-end in Holter ECGs by 2-Layer Perceptrons

W Bystricky, A Safer

Abbott GmbH & Co KG, Ludwigshafen, Germany

Abstract

Automated detection of T-end in high precision is required for ECG safety assessment of new chemical entities (NCEs). This task may be effectively accomplished by neural networks (NNs). Since it is scarcely known which configuration of NNs to choose for obtaining an optimal prediction, we explore a variety of layouts for a 2-layer perceptron.

Our training reference is the Physionet QT database with expert T-end annotations. The filtered and re-sampled signal from both channels spanning a variable time interval that contains the main part of the T wave is our model input. We investigate model variations by number of sampling points and hidden units. We train these models using Bayesian techniques and compare their properties by the evidence parameter, cross validation error, goodness of fit and the estimated prediction error.

While evidence and cross validation error favor medium sized models, residual standard deviation decreases to approximately 12 ms, whereas the estimated prediction error increases under growing model size. A medium sized 2-layer perceptron (15 sampling points over the T wave on individual channels, 15 hidden units) is suitable to describe expert annotations of T-end with a residual standard deviation of 15 ms. This configuration has promising generalization capabilities and can handle all T morphologies found in the training data set.

1. Introduction

Many attempts have been made to develop automatic measurements for waveform boundaries in ECG signals especially for T-end because of its significant clinical relevance in the QT issue (e.g. [1] [2]). While automatic methods have proved to be successful in case of undisturbed signals and well pronounced T waves they often fail when T wave morphology changes or where there is a small signal to noise ratio (either due to noisy signal or small T amplitude).

The neural network approach promises to work even in varying situations using expert knowledge for a quasi non-parametric T-end estimate. We fit a 2-layer perceptron neural network to the T-end triggers of a reference QT database[3]. For input we use the signal from both channels in a variable time interval that contains the T wave, thus avoiding information loss when regarding only one channel as it is often found in automatic algorithms.

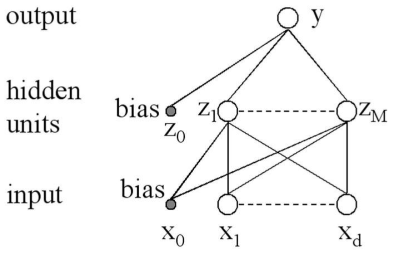

Figure 1. 2-layer perceptron

The Bayesian approach includes several advantages that have motivated us to apply it:

- because it allows to estimate the prediction error and confidence intervals can be assigned to each estimate,
- regularizer coefficients are determined exclusively on training data, and
- because it allows to compare different models.

The model properties are compared in terms of the evidence parameter, the cross validation error, goodness of fit and the estimated prediction error for the training data as well as for unseen data.

2. The neural network model

We use a 2-layer perceptron (see fig. 1) with logistic activation function in the hidden units and linear output activation. The output y can be written as:

$$y = \sum_{j=0}^{M} w_j^{(2)}\, h\left(\sum_{i=0}^{d} w_{j,i}^{(1)} x_i\right) \tag{1}$$

with activation function $h(a) = 1/(1 + exp(-a))$.

Bayesian techniques[1] are used to fit the model to given data. The error function to be minimized during the training process denotes as

$$\begin{aligned} S(\mathbf{w}) &= \tfrac{\beta}{2}\sum_{n=1}^{N}\{y(\mathbf{x}^n;\mathbf{w}) - t^n\}^2 + \tfrac{\alpha}{2}\sum_{i=1}^{W} w_i^2 \\ &=: \qquad \beta E_D + \alpha E_W \end{aligned} \tag{2}$$

[1] For details please refer to [4]. Notations are equivalent to this reference.

Here we assume that the target data t are generated by the smooth network function $y(\mathbf{x}^n; \mathbf{w})$ with additive zero-mean Gaussian noise of variance $1/\beta$. The regularizer coefficient α restricts the weights values. The weights part E_W assumes a prior distribution of the weights vector as:

$$p(\mathbf{w}) = \left(\frac{2\pi}{\alpha}\right)^{-W/2} exp\left(-\frac{\alpha}{2}\|\mathbf{w}\|^2\right) \qquad (3)$$

These weights as well as both hyperparameters are estimated during the Bayesian motivated training process that follows these steps:

1. We initialize $\alpha = 0.001$ and $\beta = 1$, and the network weights with random values drawn from $N(0; 0.01)$.
2. We train the network with a standard nonlinear optimization algorithm (conjugate gradient method BFGS [5]) to minimize the total error function $S(\mathbf{w})$.
3. After convergence or latest after 1000 cycles of the optimization algorithm we re-estimate the hyperparameters α and β using the update formula

$$\begin{aligned} \alpha^{new} &= \gamma/2E_W & (4) \\ \beta^{new} &= (N-\gamma)/2E_D & (5) \end{aligned}$$

where the quantity γ is defined by

$$\gamma \equiv \sum_{i=1}^{W} \frac{\lambda_i}{\lambda_i + \alpha} \qquad (6)$$

and $\{\lambda_i\}$ are the eigenvalues of $H = \beta\nabla\nabla E_D$, the Hessian of the unregularized error function.

4. If the changes of both hyperparameters are less than $|\triangle\alpha| < 0.01$ and $|\triangle\beta| < 0.1$ we stop the training process. Else the model weights are slightly perturbed by adding Gaussian noise drawn from $N(0; 0.01)$ and the optimization algorithm is restarted.

Having found the weights vector $\mathbf{w}_{MP}$ that maximizes the posterior distribution of $\mathbf{w}$ and the hyperparameters α, β we can approximate the distribution of the network output y given an input vector X as Gaussian with mean $y_{MP} \equiv y(\mathbf{x}; \mathbf{w}_{MP})$ and variance

$$\sigma_t^2 = \frac{1}{\beta} + \mathbf{g}^T\mathbf{A}^{-1}\mathbf{g} \qquad (7)$$

where $\mathbf{g} \equiv \nabla_{\mathbf{w}} y|_{\mathbf{w}_{MP}}$ is the derivative of the network output with respect to the weights and $\mathbf{A} = \nabla\nabla S_{MP}$ is the Hessian matrix of the total (regularized) error function. The first summand in equation 7 corresponds to the overall variability of the expert T-end triggers. The second summand describes the uncertainty of the model in predicting the given input vector. This value tends to become large for input data that are distant to the training data set or where T-end can only be measured inaccurately. So it will give us important information on how reliable a prediction may be.

Comparison of different network models may be obtained applying the evidence framework [6][4]. The evidence is the probability of observing the training data D given a specific model H_i. Its logarithm can be approximated by

$$\begin{aligned} \ln p(D|H_i) = -\alpha_{MP}E_W^{MP} - \beta_{MP}E_D^{MP} - \frac{1}{2}\ln|\,\mathbf{A}\,| \\ +\frac{W}{2}\ln\alpha_{MP} + \frac{N}{2}\ln\beta_{MP} + \ln M! + 2\ln M \\ +\frac{1}{2}\ln\left(\frac{2}{\gamma}\right) + \frac{1}{2}\ln\left(\frac{2}{N-\gamma}\right) \quad (8) \end{aligned}$$

3. The training data

All models are trained by use of the Physionet QT database [3]. This source consists of 105 fifteen-minute excerpts of two-channel ECG Holter recordings, chosen to include a broad variety of QRS and ST-T morphologies. Waveform boundaries for a subset of beats in these recordings have been manually determined by expert annotators using an interactive graphic display to view both channels simultaneously and to insert the annotations. All records are sampled at 250 Hz.

The input data are determined in the following way:

- We band pass filter[2] the signal to remove baseline shift and high frequency noise.
- For each beat we calculate a time interval $[t_1;\ t_2]$ that contains the main part of the T wave:

$$\begin{aligned} [t_1;\ t_2] = [160;\ min\{550\sqrt{0.001\cdot RR} + 100; \\ RR - 100;\ 800\}] \qquad (9) \end{aligned}$$

where RR is the time interval between the current and the next R trigger measured in milliseconds. The factor $\sqrt{0.001\cdot RR}$ comes from the well known Bazett correction formula for QT. The expression $min\{\cdots\}$ ensures that the following QRS complex is excluded from the time interval and that, in case of bradycardia (or missed following beats), the length of the time interval is restricted.

- The filtered signal is re-sampled at a fixed number of equally spaced points in the mentioned time interval using cubic spline interpolation.
- The sampling values of both channels are arranged one after the other.
- All sampling data from a single beat are standardized to zero mean and standard deviation = 1.

Target data are the manually annotated T-end triggers[3]. These values are post-processed to be of about unit size by:

$$t \mapsto \frac{t - t_1}{t_2 - t_1} \qquad (10)$$

[2] Bessel filter of order 4, 1 - 16Hz, applied bidirectionally

[3] q1c annotations in the Physionet QT database [3]

Some of the records could not be used for training:

- All 13 records from the MIT-BIH Supraventricular Arrhythmia Database.
- Records sel301, sele0203, sele0411 (systematic signal offset that disturbs the band pass filter process).
- Records sel35, sel37 (no T trigger)

Taking into consideration that only those beats can be used for training where the following beat is also available (for evaluation of RR), there remain 2349 beats for training.

4. Analysis

The following model configurations are investigated:

$$\begin{array}{rl} \text{\# Sampling Points} & \{11, 13, 15, 17, 19, 21, 25\} \\ \text{\# Hidden Units} & \{6, 7, 8, ... ,19 , 20, 22, 25\} \end{array} \qquad (11)$$

The smallest model (11 sampling points and 6 hidden units) has 145, the largest model (25 sampling points and 25 hidden units) 1301 adjustable model parameters. For any combination of sampling points and hidden units a model is trained in the manner described above. Log-evidence is calculated using formula 8. Cross validation error is determined as follows:

1. Training events are randomly assigned to one of 10 data sets.
2. The model is re-trained using nine of the ten data sets, starting from the fully trained model. The hyperparameters α and β are kept constant.
3. The re-trained model is tested on the omitted data set.
4. This re-training procedure is done for each of the ten data sets.
5. The final cross validation error E_V is calculated as:

$$E_V = \sqrt{\frac{1}{N} \sum_{n=1}^{N} \{y(\mathbf{x}^n; \mathbf{w}) - t^n\}^2} \qquad (12)$$

Goodness of fit is measured by the residual standard deviation. The prediction error for a given beat is estimated by

$$E_P^{(n)} := \sqrt{\mathbf{g}^T \mathbf{A}^{-1} \mathbf{g}}(t_2 - t_1) \qquad (13)$$

that is derived from formula 7. For each model this formula is applied to all training beats and furthermore to a subset of all annotated normal events in the database drawn randomly[4]. The distributions of the E_P's are compared by their median.

5. Results and discussion

Residual standard deviation decreases with model size (fig. 2) whereas prediction errors are increasing (fig. 3).

Evidence and cross validation errors are quite noisy measures. Nevertheless we can observe a broad optimum region between 12 and 19 hidden units and a slight preference of minor sampling points (fig. 4, 5).

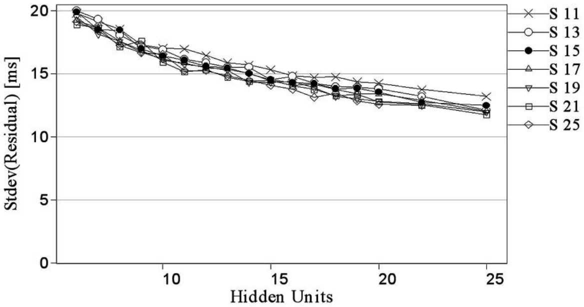

Figure 2. Standard deviation of residuals

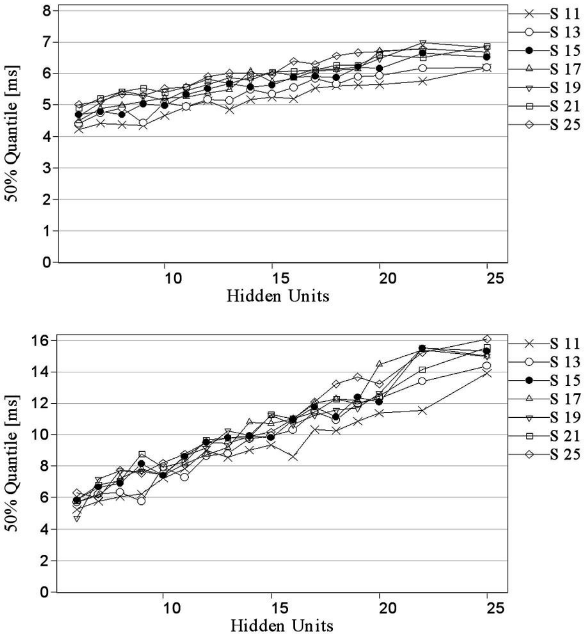

Figure 3. Prediction error for training data (above) and unseen data (below)

The reverse behavior of prediction error and goodness of fit reflects the problem with Occam's razor - that is the principle that we should prefer simple to complex models when the latter are not necessary to explain the data[7]. Both generalization measures offer similar answers to that problem: A neural network with the given configuration should use at least 12 hidden units and high re-sampling rates should be avoided.

Looking at a medium sized model with 15 hidden units and 15 sampling points we observe some differences between the data sources (table 1). Reference triggers from

[4] 10% of the original atr annotations of records we used for training

Source	N	M_R	SD_R	M_E
MIT	445	-0.63	16.6	7.15
MST	145	0.41	18.3	4.63
NSR	280	0.85	11.4	4.44
EST	879	-0.15	12.9	6.03
SDP	600	0.17	15.4	7.57

Table 1. Distribution of training data residuals and prediction error grouped by record sources. Medium model size (15 sampling points and 15 hidden units). MIT = MIT-BIH Arrhythmia Database; MST = MIT-BIH ST Change Database; NSR = MIT-BIH Normal Sinus Rhythm Database; EST = European ST-T Database; SDP = 'sudden death' patients from BIH; M_R = residual mean; SD_R = residual standard deviation; M_E = prediction error mean. All units in ms.

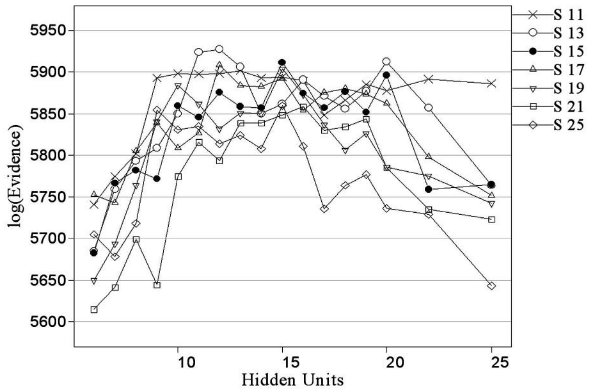

Figure 4. Log evidence (formula 8)

the MIT-BIH Normal Sinus Rhythm Database are fitted best, with a residual standard deviation (SD_R) of 11.4 ms and a mean prediction error of 4.44 ms. However, SD_R from MIT-BIH ST Change Database is about 1.5 times larger.

The given data are suboptimal with respect to train neural networks. Optimized training data should cover the entire input space and should be harvested with an emphasis on the most informative events. These may be identified as those with the largest prediction error [6]. Such active learning strategy promises to produce better prediction quality with less amount of training data.

Typically one should try to find the global optimum of a given model by repeating the optimization process with different starting conditions. Another approach is to build committees of networks from different runs[4]. Due to the large computing demand and dense screening net of model sizes we omitted these solutions. Our noisy results may be due to suboptimal fitting strategies, and give additional rise to investigate balance of evidence and cross validation parameters.

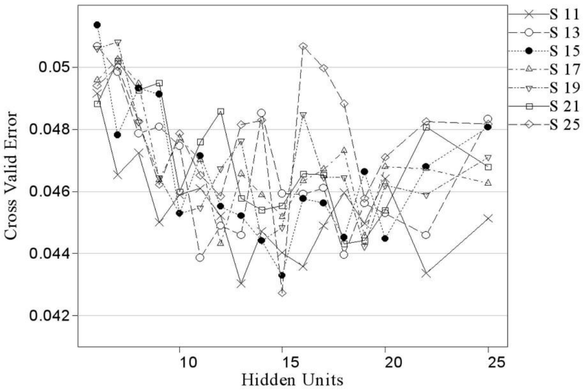

Figure 5. Cross validation error E_V (formula 12)

6. Conclusion

This analysis gives a first answer in the competition between goodness of fit and generalization in modelling T-end with a 2-layer perceptron. The investigated range of model sizes covers the optimal model size, thus providing a profound basis for further optimization of neural network strategies to automatize precise T-end assessment.

References

[1] Bystricky W, Safer A, Schweizer M, Melcher H. Beat-to-beat repolarization duration measurements from ambulatory 24 hour ecg analysis: Which is the most reliable method of evaluation. In Computers in Cardiology. 1999; 161–164.

[2] Jane J, Blasi A, Garcia J, Laguna P. Evaluation of an automatic threshold based detector of waveform limits in holter ecg with the qt database. In Computers in Cardiology. 1997; 295–298.

[3] Laguna P, Mark RG, Goldberger A, Moody GB. A database for evaluation of algorithms for measurement of qt and other waveform intervals in the ecg. In Computers in Cardiology. 1997; 673–676.

[4] Bishop CM. Neural Networks for Pattern Recognition. Oxford University Press, 1995.

[5] Numerical Recipes in C: The Art of Scientific Computing. Cambridge Univ Press, 1988-1992.

[6] MacKay DJC. Bayesian Methods for Adaptive Models. Ph.D. thesis, California Institute of Technology, 1992.

[7] Radford MN. Bayesian Learning for Neural Networks. Lecture Notes in Statistics. Springer, 1996.

Address for correspondence:

Dr. A. Safer
Abbott GmbH & Co KG
Knollstraße 50
D-67061 Ludwigshafen
anton.safer@abbot.com