LU TP 01-19

Case-Based Sensitivity Analysis for Artificial Neural Networks with Applications in Medicine

Mattias Ohlsson

Complex Systems Division, Department of Theoretical Physics
Lund University, Lund, Sweden

Lars Edenbrandt

Department of Clinical Physiology
Lund University, Lund, Sweden

To appear in *Proc. Neural Networks and Expert Systems in Medicine and Healthcare*, Milos Island, Greece (2001)

Abstract: It is often important to be able to explain the reasoning behind a machine learning algorithm, such as an artificial neural network, especially in the medical domain. In this paper we develop a case-based sensitivity analysis method for neural networks. The method uses a trained neural network committee in order to find both important and unimportant input variables for individual cases. The sensitivity analysis is formulated as combinatorial optimization problem, where the mean field annealing method is used as a tool for finding good solutions. The approach is tested on a problem from the medical domain; namely the problem of identifying patients suffering from acute myocardial infarction in the presence of left bundle branch block. We feel that the case-based sensitivity analysis developed here can be used to understand the complex functioning of a neural network and that the method can be applied on other problems from the medical domain.

INTRODUCTION

An Artificial Neural Network (ANN) is a powerful tool for classification problems in the medical field. When used in the medical domain it is often important that the classification tool is able to explain its reasoning. For classification models based on ANN this can be difficult because of their non-linearity. However, in order to reach an acceptance by the users it is important that one can understand its functioning and ultimately learn from its success. One approach to learn from an ANN is found in [1], where a case-based explanation method is developed that explains the reasoning behind the network by showing a set of similar cases.

The approach taken in this paper can be described as a case-based sensitivity analysis method, which aims at finding the important inputs for each case (e.g. patient). The notion of importance should be read as causal importance since the method will monitor changes in neural network response when manipulating the inputs. We are not looking at changes in the generalization error when removing inputs (predictive importance), which often requires retraining of the ANN. The method presented here operates on a trained neural network committee used for classification problems.

The methodology is illustrated with the real-world problem of identifying patients suffering from acute myocardial infarction in the presence of left bundle branch block. The only source of information is the 12-lead electrocardiogram (ECG) that, after pre-processing, is used as input to the neural network classifier, which classifies the ECG into either acute myocardial infarction or not. The objective for the case-based sensitivity analysis, for this particular problem, is to be able to answer questions like:

- What ECG measurements were important when making the classification?
- The network committee made an erroneous classification. How can one change the ECG in order to obtain a correct classification?
- What measurements can be omitted for this ECG without changing the output from the network committee?

The methodology is developed in the next section, where the neural network committee is defined and followed by a description of the case-based sensitivity analysis. The section is ended by a brief review of the mean field annealing method. The experiment section describes the medical classification problem on which the method is tested; followed by the result section. The paper is ended with a conclusion.

METHODS

Artificial Neural Networks

The starting point is a trained neural network committee with P members. The output from the committee is calculated as the mean of its members. Let $\mathbf{x} = (x_1, x_2, ..., x_N)$ be a given input vector, the committee output $y_{\text{com}}(\mathbf{x})$ is then given by,

$$y_{\text{com}}(\mathbf{x}) = \frac{1}{P}\sum_{k=1}^{P} y_k(\mathbf{x}) \tag{1}$$

where $y_k(\mathbf{x})$ is the output from the k:th neural network in the committee.

We will now study the effect of changing the input by an amount $d\mathbf{x}$, which results in a new committee output $y_{\text{com}}(\mathbf{x}+d\mathbf{x})$. The change of the i:th component of the input vector dx_i, can only take one out of M possible values and is expressed as

$$dx_i = \sum_{l=1}^{M} s_{il}\epsilon_{il} \tag{2}$$

where we have introduced a binary decision variable s_{il}, defined according to

$$s_{il} = \begin{cases} 1 & \text{if} \quad x_i \to x_i + \epsilon_{il} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Obviously, the variables must fulfill the following normalization (Potts) conditions:

$$\sum_{l=1}^{M} s_{il} = 1 \quad (i = 1, ..., N) \tag{4}$$

By this construction, the input vector $\mathbf{x}$ can be changed into one out of possible M^N new input vectors.

If we assume a feed-forward neural network with one hidden layer, the committee output for the modified input vector $\mathbf{x} + d\mathbf{x}$ is given by,

$$\begin{aligned} y_{\text{com}}(\mathbf{x}+d\mathbf{x}) &= \frac{1}{P}\sum_{k=1}^{P} y_k(\mathbf{x}+d\mathbf{x}) && (5) \\ &= \frac{1}{P}\sum_{k}\phi_o\left(\sum_j \omega_j^{(k)}\phi_h\left(\sum_i \tilde{\omega}_{ij}^{(k)}\left[x_i + \sum_l s_{il}\epsilon_{il}\right]\right)\right) && (6) \end{aligned}$$

where $\phi_o(\cdot)$ and $\phi_h(\cdot)$ are the activation functions for the output and the hidden layer, respectively. The input to hidden weights for the k:th network in the committee are denoted by $\tilde{\omega}_{ij}^{(k)}$ and the corresponding hidden to output weights are $\omega_j^{(k)}$ (see Fig 1). For a fixed set of weights and input vector changes (ϵ_{il}), y_{com} can be regarded as a function of the decision variables s_{il}. Next we will formulate two different cost functions that are to be minimized with respect to s_{il}.

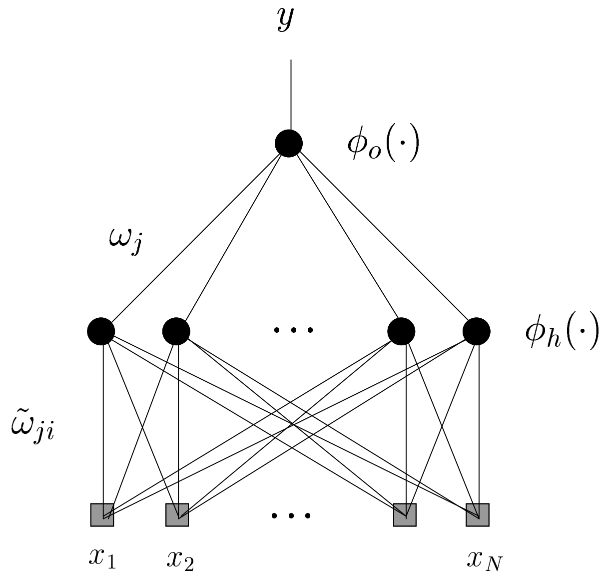

Figure 1: The neural network architecture used in this paper.

The Cost Functions

Depending on the objective with the case-based sensitivity analysis two different cost functions will be formulated. If we want to find unimportant input variables for a particular case, the input vector changes (ϵ_{il}) can be defined as:

$$\epsilon_{i0} = -x_i \tag{7}$$

$$\epsilon_{i1} = 0 \tag{8}$$

with $M = 2$. This implies that each input variable can be either "off or on". A suitable cost function to minimize is therefore,

$$E_1(S) = (y_{\text{com}}(\mathbf{x} + d\mathbf{x}) - y_{\text{com}}(\mathbf{x}))^2 + \alpha \left(\sum_{i=1}^{N} s_{i0} - N_{\text{off}} \right)^2 \tag{9}$$

where S denotes the matrix with elements s_{il}. The first term minimizes the difference between the modified and the original committee output. The second term is zero when the number of input variables in the "off" state is N_{off}. The set of variables $\{s_{il}\}$ minimizing this cost function determines the N_{off} input variables that, when deleted, makes the smallest change of the committee output. Since this is a case-based analysis one has to minimize this cost function for each case.

On the other hand if the objective is to find a set of input variables that are important for a particular case, another cost function must be minimized. Importance is now measured as causal importance, which means that we want to identify the input variables that, when changed, has the largest effect on the committee output. The cost function appropriate for this is,

$$E_2(S) = -\left(y_{\text{com}}(\mathbf{x}+d\mathbf{x}) - y_{\text{com}}(\mathbf{x})\right)^2 + \alpha\left(\sum_{i=1}^{N} s_{i0} - (N - N_{\text{mod}})\right)^2 \qquad (10)$$

The first term now maximizes the change caused by modifying the input $\mathbf{x} \to \mathbf{x}+d\mathbf{x}$. The input vector changes are defined in order to cover a small neighborhood around $\mathbf{x}$,

$$\epsilon_{il} = \frac{2l}{M}\delta_i \mathbf{x}_i \qquad (l = 1, ..., M/2) \qquad (11)$$

$$\epsilon_{il} = -\frac{2(l - M/2)}{M}\delta_i \mathbf{x}_i \qquad (l = M/2+1, ..., M) \qquad (12)$$

$$\epsilon_{i0} = 0 \qquad (13)$$

with appropriate values for δ_i. M should be large enough in order to avoid effects of the discretization. The second term in the cost function (Eq. (10)) forces the solutions to have N_{mod} modified input variables.

When minimizing either of the cost functions E_1 or E_2 the solutions found must conform to the Potts conditions of Eq. (4). Next we will see how the method of mean field annealing can be used to efficiently find good minima to the cost functions.

The Mean Field Annealing Method

Since the case-based sensitivity analysis is formulated as a combinatorial optimization problem we need an efficient procedure for minimizing $E_{1,2}$ with respect to the binary s_{il} variables and subject to the constraints of Eq. (4). This is a non-trivial minimization problem. Using some local updating rule will most often yield a local minimum close to the starting point, with poor solutions as a result. Simulated annealing (SA) [2] is one way of escaping from local minima since it allows for uphill moves in E. In SA a sequence of configurations is generated according to a stochastic algorithm, such as to emulate the probability distribution

$$P(S) = \frac{e^{-E(S)/T}}{\sum_{\{S'\}} e^{-E(S')/T}} \qquad (14)$$

where the sum runs over all possible configurations S'. The parameter T (temperature) acts as a noise parameter. For large T the system will fluctuate heavily since $P(S)$ is very flat. This implies that the generated configurations contain mostly poor and infeasible solutions. On the other hand, for a small T, $P(S)$ will be narrow, and the sequence of configurations will be strongly dependent upon the initial one and contain configurations only from a small neighborhood around the initial starting point.

In SA one generates configurations while lowering T (annealing), thereby diminishing the risk of ending up in a suboptimal local minimum. This is quite CPU-consuming, since one has to generate many configurations for each temperature following a careful annealing schedule (typically $T_k = T_o/(\log(1+k))$ for some T_o) in order to be certain to find the global minimum. In the *mean field annealing* (MFA) approach [3] the costly stochastic SA is approximated by a deterministic process. MFA also contains an annealing procedure. The original binary variables s_{il} are replaced by continuous mean field variables $v_{il} \in [0, 1]$, with a dynamics given by iteratively solving of the so-called *mean field equations* for each T.

An additional advantage of the MFA approach is that the continuous mean field variables can evolve in a space not accessible to the original binary variables. The mean field equations suitable for updating our mean field variables are:

$$v_{il} = \frac{\exp(u_{il}/T)}{\sum_{k=1}^{M} \exp(u_{ik}/T)} \tag{15}$$

where the variables u_{il} are given by,

$$u_{il} = \frac{\partial E}{\partial v_{il}} \tag{16}$$

Usually, in order to improve convergence, the derivatives in Eq. (16) are replaced by finite differences (see e.g. [4]). The algorithmic procedure for minimizing the cost function is summarized if Fig. 2:

EXPERIMENTS

The case-based sensitivity analysis method outlined above is applied on a classification problem from the medical domain. The task is to detect acute myocardial infarction (AMI) in electrocardiograms (ECGs) with left bundle branch block (LBBB) [5]. Early diagnosis of AMI is of vital importance. The 12-lead ECG is still the best and most readily available device for investigation of AMI, and since the presence of (LBBB) makes the electrocardiographic manifestations of acute myocardial transmural ischemia difficult to detect, the presence of LBBB is an issue of major diagnostic importance.

1. Initialize the mean field variables close to the value $1/M$ (randomly).
2. Set the start temperature to a large value (e.g. $T_o = 10$).
3. Randomly (without replacement) select one row of the V matrix, say row k.
4. Update all $v_{kl} (l = 1, ..., M)$ according to Eq. (15).
5. Repeat items 3-4 M times (such that all rows have been updated once).
6. Repeat item 5 until no changes occur (defined by e.g. $\frac{1}{MN} \sum_{il} |v_{il} - v_{il}^{\text{old}}| \leq 0.01$).
7. Decrease the temperature, $T \to \eta T$ (e.g. $\eta = 0.95$).
8. Repeat items 3-7 until all v_{il} are close to 1 or 0.
9. Finally, the mean field solution is given by the integer limit of v_{il}, i.e. for each row $i \quad (i = 1, ..., N)$ select the column l^* such that v_{il^*} is the largest element for this row. Let $s_{il^*} = 1$ and all other $s_{il} = 0$ for this row.

Figure 2: Summary of the mean field annealing algorithm.

Study Population

This study is based on ECGs recorded at the emergency department of the University Hospital in Lund, Sweden from July 1990 to May 1997. The final data set consists of 518 ECGs with LBBB configuration, of which 120 are recorded on patients with AMI and 398 on control patients. The AMI group consists of 74 ECGs recorded on males and 46 ECGs recorded on females. The control group is composed of 202 ECGs recorded on males and 196 ECGs recorded on females. Fig. 3 shows two ECGs from the study group, where the ECG on the right-hand side comes from a patient with a diagnosis of AMI.

Electrocardiography

The 12-lead ECGs are recorded by computerized electrocardiographs (Siemens-Elema AB, Solna, Sweden), with 7 measurements from each of 5 leads being selected for further analysis: QRS duration, Q, R, and S amplitudes, and three ST-T measurements (ST-J amplitude, ST amplitude 3/8 and positive T). The ST amplitude 3/8 was obtained by dividing the interval between ST-J point and the end of the T wave into eight parts of equal duration. The amplitude at the end of the third intervals was denoted ST amplitude 3/8. In total 35 measurements from the ECG are used as inputs to the neural network.

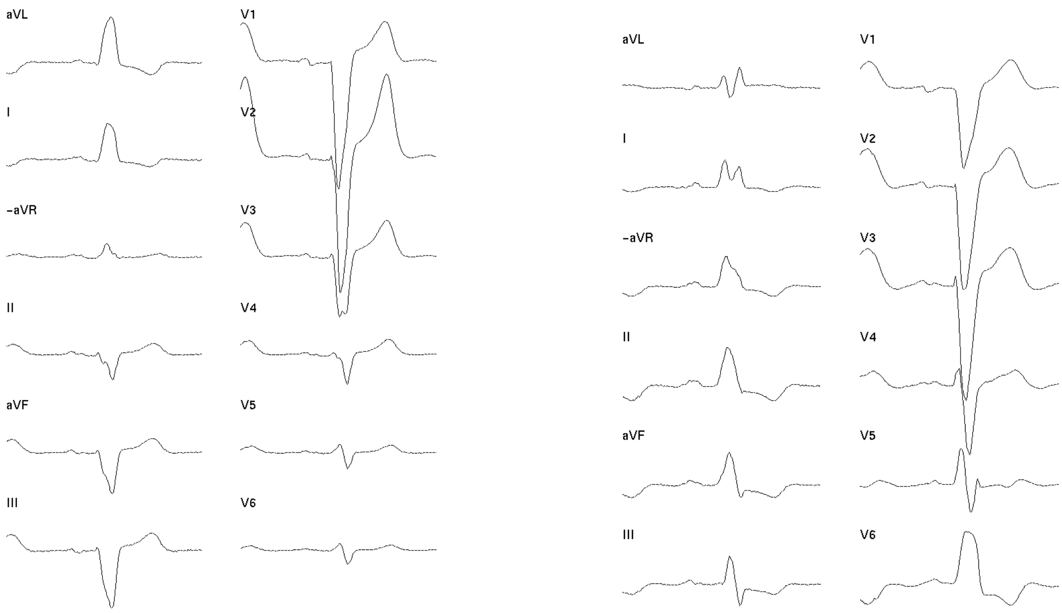

Figure 3: Example of ECGs with presence of left bundle branch block. (left) **No** diagnosis of acute myocardial infarction. (right) **With** a diagnosis of acute myocardial infarction.

Artificial Neural Networks

A standard feed-forward, one-hidden-layer neural network architecture is used in this study. The input layer comprises 35 nodes, one for each of the ECG measurements. The hidden layer contains 5 nodes and the output layer consists of one node, which encodes the output as to whether the patient suffers from AMI or not AMI. A weight elimination term [6] is added to the Kullback-Liebler error function in order to regularize the neural network. The regularization parameter is set using a 4-fold cross-validation scheme. Finally, a committee of 10 networks is trained using the full training set. The training set consists of a random selection of 478 ECGs, the remaining 40 ECGs are used as test cases for the case-based sensitivity analysis.

Case-based sensitivity analysis

In order to find important measurements for the different cases in the test set we use a cost function of the second type E_2 (see Eq. (10)). The number of input vector changes M is 11 and the maximum deviation δ_i for each input x_i is set to, $\delta_i = 0.05x_i$. The number of modified variables N_{mod} is 5, which means that the 5 most influential inputs will be detected. We are also looking for inputs that can be deleted and still get the same committee output, i.e. using a cost function of type E_1. The parameters used for this and the previous analysis are summarized in Table 1.

Cost function	M	δ_i	N_{mod} / N_{off}	α
E_1	2	-	$N_{\mathrm{off}} = 10$	≥ 1.0
E_2	11	$\delta_i = 0.05x_i$	$N_{\mathrm{mod}} = 5$	≥ 2.0

Table 1: Parameters used when analyzing the cases in the test set with a cost function of type E_1 or E_2. The value of the α parameter is increased from the value in the table until a valid solution is found.

1 RESULTS

It is interesting to monitor the development of the mean field variables during the minimization process (annealing). Fig. 4 shows such a set of mean field variables for one of the test cases, during the minimization of E_2. At high temperature (left part of the graph)

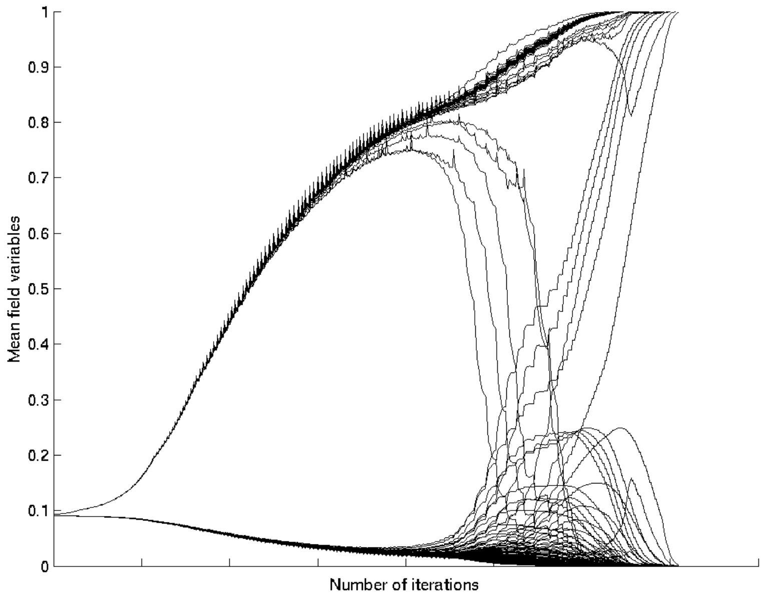

Figure 4: The development of the mean field variables during the minimization of E_2 for a case in the test set.

all variables (385) have approximately the value of $1/M$. However, as the temperature decreases, the variables converge to either one or zero.

The result of the case-based sensitivity analysis for some of the test cases is shown in Table 2. This table lists the 5 most important input measurements for 2 non-AMI ECGs and 2 from ECGs with diagnosis of AMI. Case 1 and 3 got the wrong classification from the network, while case 2 and 4 were correctly classified.

Case	Network output	True class	Important measurements
1	0.54	0	Q_{amp}-II, S_{amp}-II, ST_{amp}-V2, ST_{amp}-V3 , ST_{amp}-V4
2	0.034	0	QRS_{dur}-V4, S_{amp}-I, ST_{amp}-I, ST^{38}_{amp}-II, ST_{amp}-V2
3	0.028	1	QRS_{dur}-V2, QRS_{dur}-V3, S_{amp}-II, S_{amp}-V2, T^{+}_{amp}-V2
4	0.78	1	Q_{amp}-V3, ST_{amp}-II, ST_{amp}-V2, ST_{amp}-V4, T^{+}_{amp}-V4

Table 2: The important measurements for 4 cases in the test set, together with the network output and true class belongings.

Looking at the other cases in the test set one finds that the measurements from the ST-T interval are usually important, which is natural for the detection of AMI.

Looking at the results for the minimization of the cost function E_1, which aims at finding unimportant input variables one can deduce that the measurements QRS_{dur}-II, Q_{amp}-I, Q_{amp}-V4, R_{amp}-I and R_{amp}-V4 are not important for the classification of AMI. This conclusion is based on an average of the 40 test cases and there are variations between the different cases, which makes the cased-based sensitivity analysis useful.

CONCLUSIONS

We have developed a case-based sensitivity analysis method for artificial neural networks. It can be used in order to find both important and unimportant input variables for each case presented to the neural network.

Although not presented in this paper, the method can be used to find the minimal set of changes for a misclassified ECG, in order to obtain the correct classification. This application requires another representation of the ECG that enables a visual inspection of the inferred change. This can be of great benefit for the medical doctor when analyzing the ECG.

The case-based sensitivity analysis presented here is not limited to classification of ECGs, there is a wide range of problems both within and outside the medical domain, where case-based sensitivity analysis could be helpful.

ACKNOWLEDGMENTS

This study was supported by grants from the Swedish Medical Research Council (K2001-27X-09893-10B) and in part supported by the Swedish Foundation for Strategic Research.

References

[1] Caruana R., Case-Based Explanation for Artificial Neural Nets. Proc. Artificial Neural Networks in Medicine and Biology, Göteborg, Sweden, 303-308 (2000).

[2] Kirkpatrick S., Gelatt C.D. and Vecchi M.P., Optimization by Simulated Annealing. Science 220, 671-680 (1983).

[3] Peterson C. and Söderberg B., A New Method for Mapping Optimization Problems onto Neural Networks. Neural Systems 1, 3-22 (1989).

[4] Ohlsson M. and Pi H., A study of the mean field approach to knapsack problems. Neural Networks 10, 263-271 (1997).

[5] Olsson S. E., Öhlin H., Ohlsson M. and Edenbrandt L., Neural networks - a diagnostic tool in acute myocardial infarction with concomitant left bundle branch block. LU-TP 00-38 (submitted to Journal of the American College of Cardiology).

[6] Hanson S.J. and Pratt L.Y., Comparing biases for minimal network construction with backpropagation. In: Touretzky DS (ed) Advances in Neural Information Processing Systems, San Meteo CA, 177-185 (1989).