Exploring new possibilities for case based explanation of artificial neural network ensembles

Green, Michael; Ohlsson, Mattias

2008

Exploring new possibilities for case based explanation of artificial neural network ensembles

Michael Green and Mattias Ohlsson
Computational Biology and Biological Physics, Lund University, Sweden

Aim

- Investigate the possibility of explaining the decision of an artificial neural network ensemble, case by case, in a clinical setting.

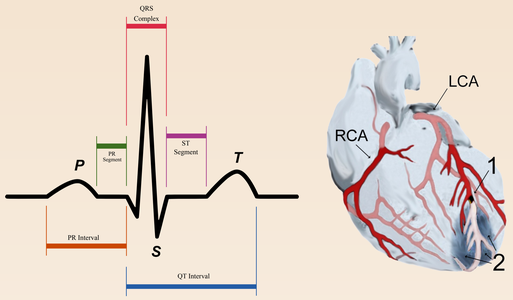

Figure 1. To the left is an Illustration of the different sections a lead from the standard electrocardiogram is divided into. The right picture shows what happens to the heart during a late stage of acute coronary syndrome.

Introduction

- Artificial neural network (ANN) ensembles is a powerful classification tool for many clinical problems.
- These tools has long suffered from lack of interpretability due to their complex nature. This has severely limited the practical usability of ANNs in settings where an erroneous decision can be disastrous.
- Several attempts have been made to alleviate this problem. Many of them are based on decomposing the decision boundary of the ANN into a set of rules.
- We explore and compare a set of new methods for this explanation process on one artificial data set, and one acute coronary syndrome data set consisting of 861 electrocardiograms (ECG) collected retrospectively at the emergency department at Lund University Hospital.

Sensitivity Method

- Analyze how sensitive the network decision is with respect to a given feature by modified partial derivatives.
- The derivative of the outer sigmoid function of each network in the ensemble is removed in order to avoid truncating effects. See Equation 1.

$$S_l(x) = \sum_{i=1}^{I} \sum_{j=1}^{J} \omega_{ij} g'_{ij}(x) \tilde{\omega}_{ijl}$$

Equation 1. Mathematical representation of the sensitivity method for a feature **l** in a patient **x**.

Euclidean Distance Method

- Find the shortest distance from data point D to decision boundary B by network inversion.
- Create a vector V = D - B and divide it into its components. These are marked by red arrows in Figure 2.
- The magnitude of each component describes the importance of the corresponding feature.

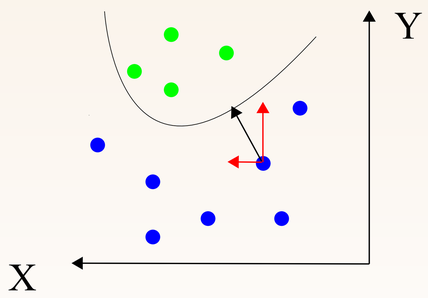

Figure 2. Illustration of the Euclidean distance method. The distance vector to the decison boundary is divided into its components i.e. the red vectors. The magnitude of these vectors determine the explanational potential for each feature.

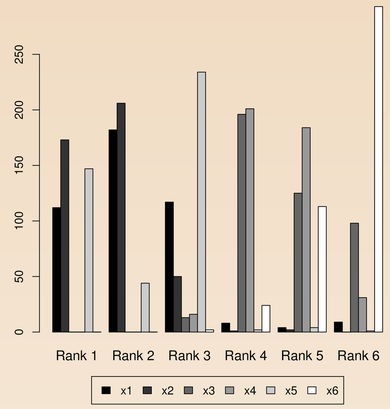

Figure 3. Distribution of selected input features within each rank for the artificial data set. For example: the first six bars tells us how many times variable 1,2,3,4,5 and 6 was voted the most imortant explanational feature for a given patient. The next six bars represent the second most important feature and so on. In this data set only features 1, 2 and 5 matters. The rest is noise.

Features Selected

- The distribution of features selected on the artificial data is consistent with their importance to the classification.
- The average overlap of the selected explanational features between physician and method increases with the predictive certainty of the ann ensemble.

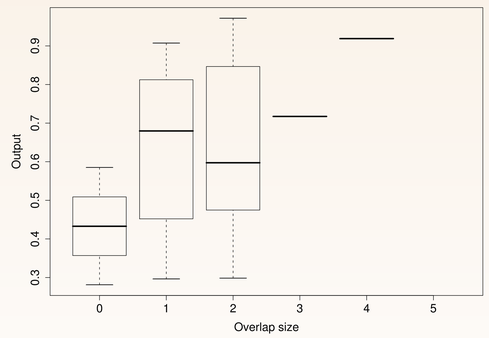

Figure 4. Illustration of the distribution of probability for acute coronary syndrome stratified on the number of overlapping selected features between a physician and the neural network ensemble. probabilities near 0.4 indicate uncertain predictions.

Explaining the ANN ensemble

- Our algorithms managed to extract 99% good explanations on the simulated data set.
- The overlap of selected explanational features between physician and method was similar to the overlap between physician 1 and physician 2.

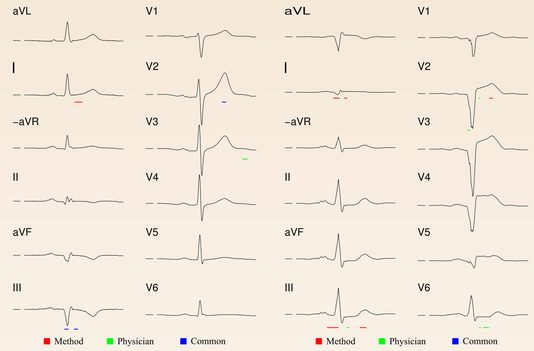

Figure 5. Illustration of the top five features indicating acute coronary syndrome as given by the ann ensemble and an experienced physician. The selected features are marked in red (ANN ensemble), green (physician) and blue (common interpretation) in the two ECGs.

Conclusions

- The algorithms has the potential to be used as a case-by-case explanatory aid when using ANN ensembles in clinical decision support systems.