3 An Implementation Example

The coronary attack is one of the most important and common reason for death all over the world. Since most of the deaths are from coronary heart disease, it is important to diagnose heart disease from simple clinical tests or determine whether a patient has risk factor after the coronary attack. This emphasizes the importance of an alternative method which may be helpful for early and accurate decision giving. Thus, so far many works have been realized on the prognosis of heart diseases using test results of the patients by researchers [8-11].

3.1 Echocardiogram Dataset

Echocardiogram dataset which takes place in the UCI repository of machine learning databases consists of some information about 132 patients who have suffered heart attacks at some point in the past. Some of the patients are alive after one year and some are not. The survival and still-alive variables, when taken together, indicate whether a patient survived for at least one year following the heart attack. The most difficult part of this problem is correctly predicting that the patient will not survive. This problem can be reduced by adding new samples to the dataset. The dataset has 13 raw attributes, however only 9 of them are used. All attributes are numeric-valued. The definitions of the 13 attributes are as follows:

1. Survival: the number of months patient survived (has survived, if patient is still alive);

It is possible that some patients have survived less than one year but they are still alive because all the patients had their heart attacks at different times. Thus, the second variable should be investigated to confirm this.

2. Still-alive: a binary variable (0: dead at end of survival period, 1: still alive);
3. Age at heart-attack: age when heart attack occurred;
4. Pericardial-effusion: Pericardial effusion is the fluid around the heart (0:no fluid, 1: fluid);
5. Fractional-shortening: a measure of contractility around the heart;
6. epss: E-point septal separation, another measure of contractility;
7. lvdd: left ventricular end-diastolic dimension. This is a measure of the size of the heart at end-diastole;
8. Wall-motion-score: It is a measure of how the segments of the left ventricle are moving;
9. Wall-motion-index: It is equal to the wall-motion-score divided by number of segments seen. Usually 12-13 segments are seen in an echocardiogram. This variable can be used instead of the wall-motion-score;
10. Mult: a derivate variable which can be ignored;
11. Name: the name of the patient;
12. Group: meaningless;
13. Alive after 1 year: Derived from the first two attributes (0: patient was either dead after 1 year or had been followed for less than 1 year, 1: patient was alive at 1 year).

Real-world data commonly contains instances with missing attribute values. The completion of the missing attribute values is one of the problems that most learning models have to handle. The echocardiogram dataset used in this work has also missing attribute values in both input and output attributes. In this work, the missing input attribute values are completed by a pre-processing method [12-13] which is done by replacing the missing attribute values with the average value of the attribute and the instances with missing output attribute values are discarded. Since these methods transform the echocardiogram data before it is given to the neural network model, the pre-processing method used are applied both in training and testing. After the pre-processing, 117 instances; 24 instances from the class who were alive at one year and 93 instances from the class who were either dead after one year or had been followed for less than one year have remained in echocardiogram dataset.

3.2 Cross-Validation Method

In this work, different neural networks were used to decide whether a patient will live one year after a heart attack using echocardiogram dataset. To estimate the accuracy of the neural network models included in this work, cross-validation, which "provides a nearly unbiased estimate" of the accuracy, is used. Cross-validation in its simplest form is the division of a dataset into two subsets and training the network with one of the subsets while testing it with the other subset [14]. As noted in [14], cross-validation estimates of accuracy can have a high variability especially with small sample sizes, such as in echocardiogram dataset. Thus, in this work, two-fold cross validation method is used in order to remove the potential imbalance in the class distributions. The echocardiogram data set (117 instances) is divided into two subsets which are denoted as A and B. Classification accuracies of 1st case, which means the subset A is used for training and subset B is used for testing and of 2nd case, which means the subset B is used for training and subset A is used for testing, are obtained. Also the averages of these two cases are found.

3.3 Applied Neural Network Structures

MLP network, which has configuration of 8 input neurons, 5 neurons in hidden layer, and 1 output neuron with learning rate, 0.1, was trained for 400 epochs. Tangent sigmoid and logarithmic sigmoid transfer functions were used in MLP training. The input values have been normalized between 0 and 1. MLP network models were trained with almost all network learning algorithms. Among all these algorithms, the one giving the best results for MLP network, which is BFGS quasi-Newton (trainbfg) learning algorithm, takes place in Table 1. In RBF network, spread value is chosen as 0.1 which gives the best accuracy. The spread values are chosen as 0.1 for GRNN networks, 1.9 for PNN networks and 0.1 for LVQ networks. MATLAB 7.0 Neural Network Toolbox is used in the simulation of the networks.

3.4 Accuracy Results of the Networks

The classification accuracies obtained in two cases and the average of these cases for MLP, RBF, PNN, GRNN and LVQ networks are given in Table 1.