Neural Network Approach to Incomplete Data Applied to Assessing Cardiac Health

Joanna Grabska-Chrząstowska

AGH University of Science and Technology, Krakow, Poland

Abstract

The project is based on data from the Cleveland Clinic Foundation Clinic, located in Cleveland. In the database, there are 13 variable: age, sex, type of chest pain, resting blood pressure, serum cholesterol, blood sugar levels, results of the resting ECG, maximum heart rate, angina, decrease the value of the ECG ST , slope of the ST segment on the ECG, number of large blood vessels, scintigraphy result. The patient is assigned to one of the two groups: healthy or sick (0 or 1).

Using a neural network MLP (Multi Layer Perceptron) with backpropagation learning method, for all 13 parameters almost 95% of correct classification of validation set was achieved. Unfortunately, even a best chosen neural network is not suitable for classification of incomplete data. With the help of genetic algorithm used to select the input group, the most important parameter was found. Maximum heart rate determines the classification of a fairly good result (71.5%). Most of the databases have this parameter. Other easily available parameters were added in order to improve the quality of classification. A choice of four parameters gives the best optimal results for test databases, within the limits of 80% positives, and for one of them even close to 90%. The results demonstrate the possibilities of neural networks to classify vectors of incomplete content.

1. Introduction

The classification problem with missing data is an important issue especially in the case of medical data. There is no standard procedure for collecting data for the particular disease. For this reason, for example, in assessing the degree of degradation of the heart muscle, several medical research centres established a group of parameters in order to create a knowledge database. Unfortunately, the records of the database are often incomplete. This paper aims at developing a general method for optimal selection of input data to neural networks in case of missing data. The project is based on data from the Cleveland Clinic Foundation, located in Cleveland, US [7]. The data was collected from accessible web base UCI Machine Learning Repository. In the database, there are 13 variables: age, sex, type of chest pain, resting blood pressure, serum cholesterol, blood sugar levels, results of the resting ECG, maximum heart rate, angina, decrease the value of the ECG ST, slope of the ST segment on the ECG, number of large blood vessels, scintigraphy results. The patient is assigned to one of the two groups: healthy - sick (0 or 1).

The problem of missing data in a database has been known for many years [1]. In most cases researchers concentrate on the attempt to replace the missing values. The most popular method is e.g. inserting the mean value of all cases for the given parameter. In medical use it is not an ideal method. Certainly, one may omit the missing data on condition that most data in the database is fully given. In medical use it may appear that a particular medical centre has a problem of using different procedures and it may not have a possibility of performing some tests. An interesting approach to the issue of missing data without imputation is presented in [2]. A neural network with cosine function was used to match the input data or omitting it in a particular use. The paper concentrates on such a choice of parameters that it ensures the lowest amount of missing data and the model matching the issue best.

Attempts to find optimal neural models have often been undertaken by other researchers [3-5]. Usually, in a particular case one should use other approaches.

This paper describes an attempt to find a universal method using, among other, a genetic choice of input data configuration implemented in the program Statistica [6].

The database called Cleveland was the reference database during the whole research.

A professional program Statistica by StatSoft was used to create neural network models.

2. Classification results for reference database and the other databases

Table 1 presents the cardinality of data in particular databases. Cleveland database is a full database and is a reference for all the results. If the medical data comprised

in that set are not fully credible it will influence all the results. It concerns particularly the assessment of the primary data, i.e. a group which a patient belongs to. In the databases creating a test group for the ideas mentioned in the paper all the missing data was replaced by the value 0, which means no signal stimulating the given value at a given entry for the neural network. Consequently, it can cause a move of the answer towards the answer 0, i.e. it may cause an incorrect classification of sick patients into a group with the answer 1 (i.e. it may increase the number of false negative cases).

Table 1. Database cardinality.

Database name	All	Healthy	Sick
Cleveland (reference database)	**297**	160	137
V.A.	**200**	51	149
Hungarian	**294**	188	106
Switzerland	**123**	8	115

The method presented in the paper should eliminate most of such cases by eliminating the parameters with the highest number of missing data.

The best results of classifying patients into two groups (hence two neural network outputs) were received for a neural network with 11 neurons in a hidden layer (with an exponential transfer function) and an output with a special exponential transfer function for classification networks (named softmax function). The whole set of 297 elements was divided into three subsets: the learning set - 208 (70%), the test set, which is helpful in the learning process - 30 (10%), and a validation set used to a final check-up of the effects of learning - 59 cases (20%). The chosen effects of looking for the best neural model are shown in Table 2.

Table 2. Result details for Cleveland data for classification neural networks (13 inputs).

Neural networks	Quality of sets [%]		
	training	test	validation
13 -> 5 -> 2	83,7	86,2	94,9
13 -> 10 -> 2	83,7	82,8	93,2
13 -> 11 -> 2	**88,0**	**89,7**	**94,9**
13 -> 15 -> 2	83,7	86,2	91,5

Other database tests with the selected network from Table 2 are presented in Table 3 as the reference point for further considerations.

Table 3. Results for Cleveland data and other databases for the best classification neural network (13 inputs).

Database name	Quality [%]	Sensitivity [%]	Specificity [%]
Cleveland (reference database)	**89,6**	86,9	91,9
V.A.	**47,5**	33,6	88,2
Hungarian	**72,1**	29,2	96,3
Switzerland	**56,9**	55,7	75,0

Only the Hungarian database has a considerably good quality but it is only an apparent success. Ascribing the value 0 to the missing data results in moving the existing results towards zero and prefers better recognition of healthy patients than the sick ones. It is proved by the values of sensitivity: 29.2 % (ability to identify sick patients) and specificity: 96.3 (ability to identify healthy patients).

3. Tests for chosen parameters

3.1. Analysis of important parameters

Analyzing the available parameters and the missing data in the subsequent databases leads to a conclusion that the last three parameters (#11-13): the slope of the peak exercise ST segment, number of major vessels (0-3) colored by fluoroscopy and diagnosis of heart disease (angiographic disease status) are very difficult to obtain. Due to that reason they do not exist in all the three test databases. On the other hand analysis of neural networks with a possibility of decreasing the number of inputs shows that all the three parameters exist in each combination of inputs with acceptable results.

Additionally maximum heart rate parameter (#8) and additionally sex (#2), chest pain type parameter (#3) and ST depression induced by exercise relative to rest (#10) seem to be important.

It should be added that the result of the reference database for only one parameter - the pulse at the input is not the worst at all - 71%. However it will not be enough and one should add combinations of other important parameters omitting the last three parameters, which the other three bases do not take into consideration.

3.2. Test results for a lower number of parameters

Considering the above the most obvious combination of parameters is the choice of the first 10 elements of the database. The results for such trials are shown in Table 4 V.A. database is still out of range – such a network does not recognize healthy patients sufficiently (8 correct classifications out of 51, i.e. 15.7%).

Table 4. Results for Cleveland data and other databases for the best classification neural network (first 10 inputs).

Database name	Quality [%]	Sensitivity [%]	Specificity [%]
Cleveland (reference database)	**82,5**	79,6	85,0
V.A.	**67,5**	85,2	15,7
Hungarian	**76,9**	77,4	76,6
Switzerland	**74,8**	76,5	50,0

Another set of parameters are 4 inputs: sex (#2), chest pain type parameter (#3), maximum heart rate parameter (#8) and ST depression induced by exercise relative to rest (#10). The results presented in Table 5 show an obvious decrease of the classification quality for Cleveland database (from the original 88% to 82.5%) but, at the same time, there is a betterment of the analysis quality for test databases. Although the specificity parameter for V.A. base is low (15.7%) but occurs at high sensitivity, i.e. almost unmistakable selection of sick patients. The healthy ones will, in the worst case, have to undertake extra examination. One should remember that hat V.A. database holds three times more sick patients than the healthy ones. In Switzerland database there are only 8 healthy patients so the specificity parameter (50%) seems to be of little importance.

Table 5. Results for Cleveland data and other databases for the best regression neural network (4 inputs).

Database name	Quality [%]	Sensitivity [%]	Specificity [%]
Cleveland (reference database)	**80,8**	78,8	82,8
V.A.	**76,0**	96,6	15,7
Hungarian	**75,2**	83,0	70,7
Switzerland	**88,6**	91,3	50,0

The undertaken tests revealed that the network, which has good perspectives, is the structure with three inputs (# 2, 3 and 8) i.e. sex (#2), chest pain type parameter (#3), maximum heart rate parameter (#8). However, the results shown in Table 6 depict that the lack of one more parameter worsened the test indicators.

Table 6. Results for Cleveland data and other databases for the best classification neural network (3 inputs).

Database name	Quality [%]	Sensitivity [%]	Specificity [%]
Cleveland (reference database)	**79,1**	74,5	83,1
V.A.	**73,5**	92,6	17,6
Hungarian	**74,1**	81,1	70,2
Switzerland	**87,0**	90,4	37,5

4. Discussion

Comparing the received results it must be stated that decreasing the number of inputs brought unexpectedly good results. However it is impossible to go under four parameters. Additional trials for one input did not qualify for wider discussion (Cleveland - 69.7%, V.A. 71.5%, Hungarian 56.5%, 79.7% of the classification).

Improving the results of tests for the databases with missing data ensued a drastic decrease of the number of the missing elements. Given that the results for the tested databases oscillate around 80%, i.e. they are similar to the results of the reference database, we can conclude that: firstly, the assessments of the experts if a patient has cardiac problems or not in particular centers do not differ from each other and secondly, the accepted assumptions were correct. Table 7 shows the comparison of results.

Table 7. Comparison of results.

	Quality for different numbers of input parameters [%]			
Numbers of param.	13	10	**4**	3
Database name	all	1- 10	**2,3,8,10**	2,3,8
Cleveland (reference database)	89,6	82,5	**80,8**	79,1
V.A.	47,5	67,5	**76,0**	73,5
Hungarian	72,1	76,9	**75,2**	74,1
Switzerland	56,9	74,8	**88,6**	87,0

6. Conclusions

The paper shows a general way of behavior in the case of classification in databases with a lot of parameters and a lot of missing data. It was shown that neural networks, apart from generally known benefits as classifiers, have big possibilities of classification with a limited number of inputs on condition that the user has a full reference database at his disposal (received e.g. after removing incomplete cases). In the case of cardiac data, after using the above presented method, it was possible to achieve a high ability of binary classification by a network with 4 (four) inputs. In the future one may aim at classifying heart degradation in the whole group of the sick (dividing the second class into four).

Acknowledgements

The work was supported by AGH-UST grant 11.11.120.612

References

[1] Gupta A, Lam MS. Estimating missing values using neural networks. The Journal of the Operational Research Society

1996; 47:229-38.
[2] Randolph-Gips M. A new neural network to process missing data without Imputation. . In Seventh International Conference on Machine Learning and Applications; San Diego, CA USA 2008; 756-62.
[3] Tadeusiewicz R. Using neural networks for simplified discovery of some psychological phenomena. In: Rutkowski L. (et al., eds.), Artificial Intelligence and Soft Computing. Springer-Verlag: Berlin – Heidelberg – New York, 2010;104–23.
[4] Szaleniec M, Tadeusiewicz R, Witko M. How to select an optimal neural model of chemical reactivity? Neurocomputing 2008;72: 241–56.
[5] Dudek-Dyduch E, Tadeusiewicz R, Horzyk A. Neural network adaptation process effectiveness dependent of constant training data availability. Neurocomputing 2009 72:3138–49.
[6] StatSoft homepage: http://www.statsoft.com
[7] Merz CJ, Murphy PM. UCI repository of machine learning databases 1996. http://archive.ics.uci.edu/ml/datasets/Heart+Disease

Address for correspondence:

Joanna Grabska-Chrząstowska
AGH Katedra AiIB, al. Mickiewicza 30,
30-059 Krakow, Poland
asior@agh.edu.pl