IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Volume-5, Issue-3 E-ISSN: 2347-2693

A Study on performance of UCI Hungarian dataset using missing value management techniques

R. Misir[1*], R.K. Samanta[2]

[1]Department of Computer Science, Vidyasagar University, Midnapur, India
[2]Department of Computer Science & Application, University of North Bengal, Darjeeling, India

[*]*Corresponding Author: rajeshmisir@gmail.com*

Available online at: www.ijcseonline.org

Received:02/Feb/2017 Revised: 11/Feb/2017 Accepted: 04/Mar/2017 Published: 31/Mar/2017

***Abstract*: This is to presents a study on performance of UCI Hungarian data sets using missing value management techniques. We used bootstrap algorithm with multiple imputation (MI), LOCF, Mean–Mode substitution and IV-for missingness on the reduct file of the dataset to use all 294 instances in the dataset for our experimental input. Five imputed files were generated from the original reduct file in MI technique where from we have taken the average result and created other input files as per requirements for each specified technique, which are studied using two most recognized but opposite in nature approaches for classification, viz. IBPLN and BBP among many of such learning algorithms in the literature [10], but the most well-known among them are back propagation [11], [12], ART [13], and RBF networks [14]. Accuracy for test cases of five imputed files varies from 89.79% to 99.00% by CCR measure, the most recognized benchmarking parameter for judging classification result and performance of the dataset.**

***Keywords*: Hungarian data sets, CARN, Amelia View, R Statistical platform, Boot strapping, Multiple imputation, LOCF, Mean–Mode substitution, IV-for missingness, online incremental back propagation, Batch back propagation, CCR.**

I. INTRODUCTION

Heart disease has become a much known term on this 21th century covering a large area of heart related issues and its malfunctioning, about which we get alarming reports from W.H.O reports [1]. So we are in search of early prediction for the purpose early prevention of heart diseases and followed by remedies to avoid fatal consequences. In its way to conquer over the disease by means of said early detection using artificially intelligent system development, UCI benchmarking data repository plays a major role in this field. Though several benchmarking datasets for the same are there, yet most research works in foregoing few years are focused on the Cleveland data set due its comparatively very less missing value attributes in the dataset, increasing its smoothness for the use in classification algorithms. But we have studied that nearly no significant work on Hungarian data sets so far.

However, we even found that due to huge missing value attributes in Hungarian data sets, it can't be used directly for classification purpose by simple correlation based feature selection (CFS) after removing the records/rows having missing value attributes from the dataset (complete deletion) and then applying classification algorithms on the reduct set, like C4.5, as several experimental established works of our own research activity on Cleveland dataset. So, on the basis of our previous experience on the Cleveland dataset, we modified our present pace as following.

We have got the reduct set using correlation based feature selection (CFS) before applying imputation or any other missing value management techniques. We then prepared the input file for experiment as per missing value management technique. The input file was fully randomized for experimental transparency. We took five separate imputed files for MI, by default output specification of the AmeliaView multiple imputation software. We tested five imputed files separately for their classification performance using classification techniques like Incremental Back propagation network (IBPLN) and Batch Back Propagation (BBP) in ANN environment and got average value of the output benching marking parameters. Accuracy of test cases of five imputation techniques varies between 89.79 (IV-for missingness) and 99.00 (M.I) percent by correct classification rate (CCR) measure. Further to mention here that these two classification algorithm approaches are quite opposite in nature, yet producing nearly equal good results for all of our applied imputation techniques.

Rest organization of the paper is as follows. An overview of earlier works, data sets descriptions, CARN project, R

Statistical platform, Amelia View software, LOCF, Mean–Mode substitution**,** IV-for missingness, online and batch back learning strategies are placed in section 2. Section 3 presents the application. At section 4, we presented the results. Conclusions are summarized in the end.

II. BACKGROUND

An overview of earlier works

On the Cleveland heart disease database and Statlog heart disease database wide varieties of classification algorithms have been attempted, with the outcome of good classification accuracies. Hybrid systems using suitable feature reduction techniques with suitable classification techniques showed better performances.

Hungarian heart disease dataset

Hungarian dataset was taken from UCI machine learning repository [1]. The Hungarian dataset contains 294 instances including a total of 491 missing values; we included all missing values in our study i.e. used full dataset. Original database has 14 columns and 294 rows each having eight categorical columns: sex, cp, fbs, restecg, exang, slope, thal, num and five numeric columns: age, trestbps, chol, thalach and oldpeak.

CARN project Preliminaries

The R Archive Network is a comprehensive network of ftp and web servers throughout the world that store up-to-date and identical versions of code and documentation for R, needs to use the CRAN mirror nearest to anyone's location of use to minimize network load. The IIT Madras is our CRAN mirror for this experiment. Details can be gathered from https://cran.r-project.org/

R programming language and software environment

An augmented implementation of the S is R programming language combined with lexical scoping semantics and followed by modifications by Ross Ihaka and Robert Gentleman at the University of Auckland at New Zealand. It is a programming language and software environment for statistical computing which is actually a GNU project. The source code for this software environment is primarily written in C, FORTRAN, and R itself. It is available freely under the GNU (General Public License) which supports graphics language and is massively used by statisticians and data miners for developing statistical software and data analysis, polls, data miners' surveys, and scholarly literature database studies. R has been identified by the FDA as suitable for interpreting data from clinical research.

Amelia View software

It is software named after the American lady pilot Amelia Mary Earhart, who disappeared during making a circumnavigation flight over the central Pacific Ocean of the globe in 1937 and runs on R-environment of CARN. It is used for multiple imputation in a probabilistic boot strapping environment is getting wide spread and works well for both numeric and categorical data.

LOCF

Another widely used imputation methods for missingness is the last observation carried forward (LOCF). Every missing value replaces with the last observed value for its own subject. Although simple, a biased estimate for the treatment effect is produced that underestimates the variability of the estimated result.

Indicator Variable for Missingness

For continuous predictors, indicator variables for missingness is a popular approach in the field of social sciences, where for each continuous predictor variable with missingness we include an extra indicator identifying which observations on that variable have missing data. As because it forces the slope to be the same across both missing-data groups, this technique tends to produce biased coefficient estimates for the other predictors included in the model, yet sometimes gives satisfactory results. It is a simple and often useful approach for categorical predictors, especially for unordered categorical predictors, to impute an extra category for the variable indicating missingness.

Mean–Mode substitution

Theoretically mean substitution is that the mean is a reasonable estimate for a randomly selected observation from a normal numerical distribution and the same for categorical/nominal distribution is mode. However, for not strictly random missing values, especially when a great inequality present in the number of missing values for the different variables, the mean substitution method may lead to inconsistent bias. As thumb rule we replace the missingness by the attribute mean, if the attribute is numeric. Otherwise, replace missingness by attribute mode (if the attribute is categorical).

III. CLASSIFICATION ALGORITHMS

IBPLN

The IBPLN is basically used in online training where network adjustment of weights and bias values for every training item based between the difference of training and target data outputs and the computed outputs, which can be described in high-level pseudo code as:

loop no-of-epochs times
for each training data item
compute weights and bias deltas for current item
adjust weights and bias values using deltas
end for

end loop

Batch back propagation (BBP)

Here, adjustment delta values are accumulated over all training items giving an aggregate set of deltas and then aggregated deltas are applied to each weight and bias. The high-level pseudo code is as:

loop no-of-epochs times
for each training data item
compute weights and bias deltas for current item
accumulate the deltas
end for
adjust weights and bias deltas using accumulated deltas
end loop

IV. APPLICATION

System description

Basically, in this study there are three stages: feature extraction and reduction phase by correlation and then imputation phase followed by randomization phase and lastly classification phase by incremental back propagation neural networks (IBPLN), and batch back propagation (BBP) algorithm. The schematic view of our system is as in Figure. 1.

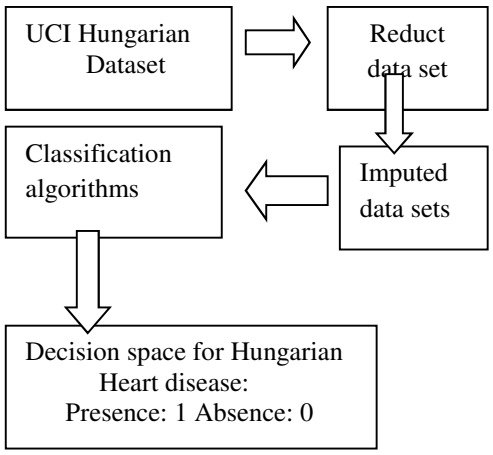

Figure.1. Schematic view of our system

Data preprocessing

Model development needs data processing as its primary requirement. So complete randomization of the reduct data sets before imputation had been done. We got eight reduced features as given in Table 1 using CFS. After randomization, we then generate imputed reduced datasets. The experimental data is now partitioned into three: Training set (68%), Validation set (16%), and Test set (16%).

Table 1. Reduced Hungarian heart disease attributes

#	Name of the	#	Name of the
1.	Sex	5.	exang
2.	cp	6.	oldpeak
3.	chol	7.	slope
4.	fbs	8	Thal

Network architecture

Choice of network architecture and feature selection are two initial tasks for using ANN. The hold-out *validation set* of data would be useful in making all these decisions successful [17]. Logistic function of the form $F(x) = 1/ (1+e^{-x})$ is used in the hidden and output nodes. Theoretically, logistic function as the activation function at one layered hidden network and output nodes can approximate any function arbitrarily closely, having sufficient hidden nodes [18]. So, we use one input layer, one hidden layer, and one output layer. We use the formula of Goa [19] modified by Huang et al [20] for fixing the number of hidden layer neurons. If *s* is the number of neurons, for *m* inputs and *n* outputs, then

$$s = \sqrt{(0.43mn + 0.12n^2 + 2.54m + 0.77n + 0.35)} + 0.51 \quad (1)$$

Here s = 6 after round off for our present study with m = 8, n = 2. For controlling overtraining, we retain the copy of the network with the lowest validation error.

V. MODELING RESULTS

Two classification algorithms using two combinations were implemented for five imputed datasets for MI, then average of five is calculated for MI and on single synthesized data sets for other methods, using Alyuda NeuroIntelligence. The summarized results are as given in Table 2.

VI. PERFORMANCE EVALUATION METHODS

As the performance measure, it is required to compute classification accuracy, test sensitivity, test *vs* pecificity and ROC. AUC close to one indicates more reliable diagnostic results for comparing classifiers in two-class problems, which is an important measure in biomedical researches for classification performance that is being used for assessesment of the performance of diagnostic tests.

VII. OBSERVATIONS OF EXPERIMENTAL RESULTS

The observations are noted below:

In our first attempt for using the dataset for classification purpose using very robust software like "Alyuda NeuroIntelligence 2.2 (577)", we have got the comment from software data analysis report "The dataset is insufficient to train neural network". It discarded 284 rows out of 294 due to 491 missing attribute values in those rows.

Hence it was impossible to accept the data set for further experimentation without any treatment on missing value attributes in a strongly suitable and obviously well accepted logical fashion. In this case, listwise deletion and pairwise deletion is fairly not acceptable choices. So we choose LOCF, Mean–Mode substitution**,** IV-for missingness and multiple imputation techniques for our target dataset and using the imputed datasets we get overwhelming results which range from 96.73 to 100 percent by CCR measure, from that "insufficient dataset before imputation", as evident from the above attached table. Anyhow, ML is giving the best result w.r.to CCR AND ACCURACY parameter, next comes MI and then comes Mean-Mode substitution, then Indicator Variable for missingness, then comes LOCF technique as per our observations.

Table 2. Modeling Results.

IMPUTATION METHOD	Classification Method	Test set (CCR%)	Test Specificity	Test Sensitivity	Test AUC	Comparison of ACCURACY.	
						TRAIN	TESTING
1. LISTWISE DELETION	NOT APPLICABLE : DATASET BECAME UNUSABLE						
2. PAIRWISE DELETION	NOT APPLICABLE : DATASET BECAME UNUSABLE						
3. MEAN-MODE SUBSTITUTION	IBPLN	91.07	80.83	33.11	0.9203705	89.28	91.08
	BBP	91.65	92.05	19.91	0.8872322	89.46	91.66
4. INDICATOR-VARIABLE FOR MISSINGNESS	IBPLN	89.79	86.64	26.62	0.8843445	83.40	89.79
	BBP	90.86	89.07	36.14	0.9047375	86.47	90.86
5. LAST OBSERVATION CARRIED FORWORD	IBPLN	89.84	81.54	27.14	0.8810333	91.83	89.85
	BBP	85.70	89.54	29.88	0.8790275	84.69	85.71
6. MULTIPLE(Amelia) IMPUTATION	IBPLN	97.94	99.87	36.35	0.9989142	99.43	98.65
	BBP	99	99.72	37.24	0.9989881	97.52	99.00
7. MAXIMUM LIKELY HOOD IMPUTATION	IBPLN	99.78	100	36.57	0.9983333	100	99.78
	BBP	99.86	99.86	36.51	0.99964538	99.64	99.86

VIII. CONCLUSIONS

We present here a work on a rarely and solely used Hungarian dataset for heart disease from benchmarking UCI data repository. The data set in its original form was not suitable for prediction and classification work by means of any well established algorithm of classification due its huge number of missing attribute values. Missing attribute values in the data set lead to uncertainty which increases with their number of occurrences in the dataset. In case of disease diagnosis for human endeavor, an incorrect result of machine supported diagnosis is related to the question life and death and hence for judgment of benchmarking parameters we compute their worst case value and also the maximum value to compute their frequency base average, which is mentioned above. Uncertain environmental parameters like missing value attributes should be handled with best care and priority without any compromise. From

this consideration we strongly argue for the multiple imputation boot strapping techniques as a solution to work in the uncertain environment regarding missing value attributes in the input platform. We also argue here for an initiation and experimentation on other datasets with such uncertainties in the same domain.

REFERENCES

[1] Rajesh Misir, R. K. Samanta, "Prediction of Heart Disease and Performances of Data Sets", Proceedings of Third International Conference on Computing and Systems 2016, pp. 7-10, January 21st-22nd, ISBN:978-93-85777-13-4.

[2] Rajesh Misir, Malay Mitra and R. K. Samanta, "A Study on Bench marking Parameters for Intelligent Systems", National Conference on Computational Technologies-2015,International Journal of Computer Science and Engineering, Volume-3, Special Issue-1. E-ISSN:2347-2693.

[3] R. Das et al., :Effective diagnosis of heart disease through neural network ensembles. Expert Systems with Applications, 369, 7675-7680 (2009).

[4] N. Cheung: Machine learning techniques for medical diagnosis. School of Information Technology and Electrical Engineering. B.Sc. Thesis, University of Queensland (2001).

[5] K. Polat et al.,: A new classification method to diagnosis heart disease: supervised artificial immune system (AIRS). Proc. of the Turkish Symposium on Artificial Intelligence and Networks (2005).

[6] Lecture Notes in Electrical Engineering 326, Springer India (2015).

[7] K. B. Nahato et al., : Knowledge mining from clinical datasets using rough sets and backpropagation neural network. Computational and Mathematical Methods in Medicine, Article ID 460189 (2015).

[8] W. McCulloch and W. Pitts, A logical calculus of the ideas immanent in nervous activity, Bulletin of Mathematical Biophysics, vol. 7, 115-133(1943).

[9] D. O. Hebb, *The Organization of Behavior, a Neuropsychological Theory,* New York, John Wiley, (1949).

[10] A. Roy, *Artificial Neural Networks- A Science in Trouble,* SIGKDD Explorations, vol. 1, issue 2, 33-38,(2000).

[11] D. E. Rumelhart, J. L. McClelland (eds.), *Parallel Distributed Processing: Explorations in Microstructures of Cognition*, vol. 1: Foundations, MIT Press, Cambridge, M.A., 318-362, (1986).

[12] D. E. Rumelhart, *The Architecture of Mind: A Connectionist Approach,* Chapter 8 in J. Haugeland (ed.), Mind_design II, 1997, MIT Press, 205-232,(1986).

[13] S. Grossberg, *Nonlinear Neural Networks: Principles, Mechanisms, and Architectures,* Neural Networks, vol. 1 , 17 - 61, (1988).

[14] J. Moody and C. Darken, *Fast Learning in Networks of Locally-Tuned Processing Units,* Neural Computation, vol. 1, 281-294, (1989).

[15] L. Fu, H. Hsu, and J. C. Principe, *Incremental Backpropagation Learning Networks,* IEEE Trans. on Neural Networks, vol. 7, no.3, 757-761, (1996).

[16] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, *Learning internal representation by error propagation,* in Parallel Distributed Processing: Explorations in the Microstructures of Cognition, MA, MIT Press, vol. 1. (1986).

[17] M. S. Hung, M. Shankar, M. Y. Hu, *Estimating Breast Cancer Risks Using Neural Networks,* J. Operational Research Society, Vol. 52, 1-10 (2001).

[18] K. Hornik, M. Stinchcombe, H. White, *Multilayer feedforward networks are universal approximator,* Neural Network, Vol.2, 359-366 (1991).

[19] D. Goa, *On structures of supervised linear basis function feedforward three-layered neural networks*, Chin. J. Comput., vol. 21, no. 1, 80-86 (1998).

[20] M. L. Huang, Y. H. Hung, and W. Y. Chen, *Neural network classifier with entropy based feature selection on breast cancer diagnosis*, J Med Syst, vol. 34, no. 5, 865-873(2010).

