AMSE JOURNALS-AMSE IIETA publication-2017-Series: Modelling C; Vol. 78; N°4; pp 467-477

Submitted Nov. 2016; Revised Jan. 2, 2017, Accepted July 15, 2017

A Study on UCI Hepatitis Disease Dataset Using Soft Computing

* Malay Mitra, ** R. K. Samanta

* Department of Computer Science and Application, University of North Bengal, Raja Rammohunpur, 734013, India. (malay.mitra68@gmail.com)

** Department of Computer Application, Siliguri Institute of Technology, Darjeeling – 734009, India (rksamantark@gmail.com)

Abstract

Accurate diagnosis for decision making in medical diagnosis is solicited for further treatment planning. Intelligent decision support system plays an important role for medical diagnosis as well as early detection of disease to survive. In intelligent model machine learning is achieved by searching a pattern in the available data set. For this reason, data preprocessing plays a vital role for better learning and analysis process. This work uses UCI Hepatitis disease data set. Missing data are managed by using multiple imputation. Feature extraction is done using rough set (RS) based techniques. Data preprocessing was the main focus to achieve better classification accuracy. Incremental Back Propagation Learning Network (IBPLN) and Levenberg Marquardt (LM) algorithms are used as classifier. The parameters – CCR, Sensitivity, Specificity and AUC are considered for performance prediction.

Key words

UCI Hepatitis data set, Feature reduction, Classification algorithm, Neural networks, Multiple imputation.

1. Introduction

Hepatitis disease, an inflammation of lever without pinpointing any specific reason [1], is one of the major diseases worldwide as many casualties are addressed. No major symptoms are there for people infected with hepatitis B or Hepatitis C, which is one of the main reasons for diagnostic errors for a physician. Decision has been made by matching the laboratory findings and other symptoms of current patient with the similar previous one, leaving a chance of incorrect diagnosis from physician's cognitive mistakes. For this reason, intelligent decision support system should be needed to minimize the complexity of hepatitis disease diagnosis. An intelligent automated system has been attempted in this work to classify hepatitis disease. Multiple imputation using EMB algorithm has been used for managing missing data. Features are selected using Rough Set (RS) feature reduction approach. Finally, we integrate the model, as multiple imputation is used, using two different approaches. A novel approach has been attempted in this work in comparison to previous works on diagnosis of hepatitis disease [1-5].

The data set used in this study is described in section 2. Section 3 presents missing data management. Section 4 describes feature selection technique. Artificial neural network along with two classification algorithms are described in section 5. Section 6 presents the model and data preprocessing. Results are shown in section 7. Finally, conclusions have been summarized.

2. Hepatitis Data Set

TABLE I. The attributes of hepatitis disease database of UCI.

Attribute	Attribute name	Attribute values
1.	Class	DIE, LIVE
2.	AGE	10,20,30,40,50,60,70,80
3.	SEX	Male, female
4.	STEROID	no, yes
5.	ANTIVIRALS	no, yes
6.	FATIGUE	no, yes
7.	MALAISE	no, yes
8.	ANOREXIA	no, yes
9.	LIVER BIG	no, yes
10.	LIVER FIRM	no, yes
11.	SPLEEN PALPABLE	no, yes
12.	SPIDERS	no, yes
13.	ASCITES	no, yes
14.	VARICES	no, yes
15.	BILIRUBIN	0.39, 0.80, 1.20, 2.00, 3.00, 4.00
16.	ALK PHOSPHATE	33, 88, 120, 160, 200, 250
17.	SGOT	13, 100, 200, 300, 400, 500
18.	ALBUMIN	2.1, 3.0, 3.8, 4.5, 5.0, 6.0
19.	PROTIME	10, 20, 30, 40, 50, 60, 70, 80, 90
20.	HISTOLOGY	no, yes

Table I presents UCI hepatitis database [6], containing 155 records with 20 attributes, among which one is class attribute having two values 'DIE' (32 cases) and 'LIVE' (123 cases). Multiple imputation using EMB algorithm is used to manage missing values present in the data set.

3. Missing Data Management

Since the data set contains only 155 samples and many samples contain one or more missing values, it should be addressed properly for better performance of the model. List-wise deletion and Pair-wise deletion methods are avoided as the number of samples are not sufficient to implement those. We did not use Regression imputation as it is complex to implement for multivariate data set. So we prefer multiple imputation using EMB approach to generate five completed data set from the data set with missing values.

The EMB algorithm is the combination of EM algorithm [7] with bootstrap approach to take draws from posterior. In each draw the data is bootstrapped and then run EM algorithm to find mode of the posterior for the bootstrapped data, which also gives fundamental uncertainty [8]. Finally, imputations are done using observed part, unobserved part, mean vector and covariance matrix with linear regression.

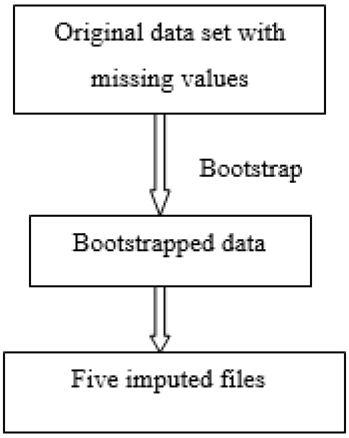

Fig.1. A schematic view of Multiple Imputation.

4. Feature Selection

Selection of features or feature reduction or extraction are important for classification as even the best classifier may perform poorly if the features are not well chosen [9][10]. Rough set based feature selection technique, presented by Pawlak in 1980's [11], was taken into consideration in this work.

Rough set theory is an intelligent technique which is used for discovering data dependencies, reducing redundancies, evaluating the importance of attributes, discovering patterns in data and also classifying objects. An information system I may be defined as I = (U,A), where U is the universe of discourse and is a non-empty set of finite objects; A be the finite non-empty set of attributes.

$$\kappa^{P} = | \underline{P}X | / | PX | \qquad (1)$$

where $X \in U$, be a target set that can be represented by preparing P-lower approximation ($\underline{P}X$) and P-upper approximation (PX).

5. Artificial Neural Network

A number of interconnected nodes, connected by connection links, each of which has a weight, constitute artificial neural network that plays an important role for developing intelligent model. One of the two tasks to develop ANN model is to decide the structure of the network which includes number of layers, number of nodes in each layer, the node function, type of network i.e. feedforward or feedback, and pattern of interconnection of nodes. Another task is to train the network using test data set which adjusts weights and threshold values. Donald Hebb [12] first proposed such learning rule. Hundreds of such algorithms are used in literature [13]. The most well known among them are back propagation [14][15], Art [16] and RBF network [17]. This work used IBPLN and LM algorithms as follows:

A. Incremental Backpropagation Learning Network

Incremental backpropagation learning network is based on the learning rule as given in equation 2, in which the network learns by backpropagation rule of Rumelhart et al. [18].

$$\Delta W_{ij}(k) = s(k)\ \eta \delta_{j}(k) O_{i}(k) \qquad (2)$$

Where s is scaling factor used for keeping weights within bounds, W_{ij} is the link weight from unit i to unit j, O_i is the level of activation at i-th unit, k stands for iteration number. The IBPLN works as given in literature [19].

B. Levenberg Marquardt (LM) Algorithm

LM algorithm is an iterative method that locates the minimum of multivariate function which is expressed as the sum of squares of non-linear real-valued functions [20][21]. The link weights for (k+1)th iteration is calculated from the k-th iteration as follows:

$$w(k+1) = w(k) - (J^{T} J + \lambda I)^{-1} J^{T} \varepsilon(k) \qquad (3)$$

Where J – the Jacobian matrix, λ – adjustable parameter, ε - error vector.

6. Structure of the Model

Fig. 2. shows the structure of the model to describe this study that contains four phases – Generation of five completed data sets to manage missing data; Feature selection using rough set based feature selection technique; Analysis phase using neural network algorithms namely Incremental back propagation learning network and Levenberg Marquardt algorithm. Finally, model has been integrated using two methods. Firstly, different performance prediction parameters are evaluated after analysis phase and the results are combined to obtain the final result. In other process, before analysis phase we integrate five imputed file to prepare a new file and analysis has been made on this file to obtain the result.

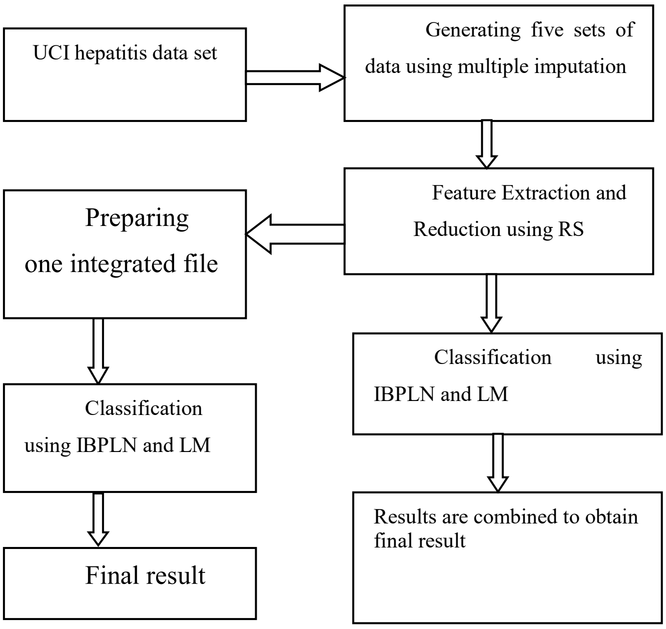
UCI hepatitis data set
Generating five sets of data using multiple imputation
Feature Extraction and Reduction using RS
Preparing one integrated file
Classification using IBPLN and LM
Classification using IBPLN and LM
Final result
Results are combined to obtain final result

Fig. 2. Block diagram of a system for hepatitis disease diagnosis.

A. Data Preprocessing

One of the important steps for model development is data preprocessing. UCI hepatitis disease data set contains a number of missing values. This work implements multiple imputation using EMB algorithm to generate five completed data sets from the original data set containing missing values. During imputation we apply certain range specifications for attributes after consulting a physician [22]. Rough set using ROSETTA [23] was implemented to extract reducts from all the five imputed data sets. Finally, seven common features are chosen as reduced features as given in Table II, present in all reducts. The data sets are partitioned into three: Training set (68%), Validation set (16%), and Test set (16%).

Table II. Reduced hepatitis attributes using RS

#	Name of the attributes
1.	AGE
2.	STEROID
3.	BILLIRUBIN
4.	ALK PHOSPHATE
5.	SGOT
6.	ALBUMIN
7.	PROTIEN

B. Network Architecture

Choice of network architecture is an important task in designing intelligent model. Validation data set is used to tune all network parameters except weights. Logistic function of the form $F(x) = 1/ (1+e^{-x})$ has been used in the hidden and output nodes. In this work we used one input layer, one hidden layer and one output layer as it is able to approximate any function arbitrarily closely, provided that the number of hidden nodes are large enough [24]. Number of hidden nodes are evaluated using the formula of Goa [25] modified by Huang et al. [26].

$$s= \sqrt{}\,(0.43mn + 0.12n^2 + 2.54m + 0.77n + 0.35) + 0.51 \qquad (4)$$

Where m is the number of input nodes, n is the number of output nodes. In this study, m = 7, n = 2; and therefore s = 6 after round off. So, for all combinations six hidden nodes are used in this work.

7. Modeling Results

In two different approaches, classification algorithms using two combinations RS + IBPLN and RS + LM were implemented in Alyuda NeuroIntelligence [27]. The supported system was Intel Core Solo T1350 CPU(1.86GHz, 533MHz FSB, 2MB L2 cache) with 512MB DDR2 RAM.

Table III gives the number of nodes used in input, hidden and output layers, number of epochs, number of retains and the number of records used in training, validation and test data sets. To control overtraining the copy of network with lowest validation error has been retained.

TABLE III. Network parameters applying to WBCD

Network structure			Epochs(retrain)	Numbers patterns		
II	HHL	OO		Training	Validation	Testing
77	66	11	2000(10)	109	23	23
77	66	11	2000(10)	109	23	23

A. Performance Evaluation Methods

Performance prediction of an ANN model can be done by evaluating correct classification accuracy (CCR), specificity, sensitivity and area under ROC curve (AUC) as these parameters are important measure of classification performance without knowing the distribution of data. These are evaluated by using the following formulae:

Accuracy = $\frac{TP+TN}{TP+FP+FN+TN}$ x 100% (5)

Sensitivity = $\frac{TP}{TP+FN}$ x 100% (6)

Specificity = $\frac{TN}{FP+TN}$ x 100% (7)

Where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives respectively. In two class problems AUC is considered as one of the important measure of classification performance and its value close to 1 indicates reliable diagnostic result [28].

B. Experimental Results

Table IV A and table IV B gives the results that we obtained from 100 simulations using two combinations RS + IBPLN and RS + LM on five imputed files. The results are then combined to obtain the final result. This technique is denoted as approach-I. In another approach denoted as approach-II, we integrated five imputed files into a single file. Analysis were made on this file using two combinations RS + IBPLN and RS + LM. Compiled results of 100 simulations for these approaches are shown in tables IV A, IV B, IV C and IV D.

Observations which are noted as follows:

- Out of two different approaches as mentioned above, RS +LM implemented on integrated file, as described in approach-II, performs better in terms of CCR, sensitivity, specificity and AUC.
- 100% classification accuracy was achieved as the best performance for all the methods described here. The lowest CCR is **86.96%** for the combination RS + LM in approach-II.

- Table V shows the comparative study of our results with other studies [4]. It is evident that the method RS + LM on integrated file outperforms all other methods.

TABLE IV A. Results of Approach – I Results from100 simulations

Methods	Imputation Number	Test set (CCR%)				Specificity			
		Highest (freq)	Lowest (freq)	Avg	Grand Avg	Highest (freq)	Lowest (freq)	Avg	Grand Avg
RS + IBPLN	IMP-1	100(4)	78.26(2)	88.51	91.13	100(45)	33.33(6)	82.38	81.45
	IMP-2	100(8)	78.26(2)	91.82		100(47)	25(1)	80.73	
	IMP-3	100(13)	73.91(1)	91.25		100(50)	25(1)	82.2	
	IMP-4	100(9)	78.26(2)	92.36		100(47)	20(1)	82.39	
	IMP-5	100(5)	78.26(6)	91.71		100(41)	25(4)	79.54	
RS + LM	IMP-1	100(11)	82.61(2)	92.61	**93.45**	100(59)	50(6)	90.22	**87.52**
	IMP-2	100(15)	86.96(13)	93.72		100(51)	33.33(1)	84.86	
	IMP-3	100(14)	86.96(8)	93.8		100(68)	33.33(1)	91.51	
	IMP-4	100(8)	86.96(14)	93.23		100(44)	33.33(3)	82.78	
	IMP-5	100(13)	86.96(10)	93.88		100(56)	33.33(2)	88.22	

TABLE IV B. Results of Approach-II Results from100 simulations

Methods	Imputation Number	Sensitivity				AUC			
		Highest (freq)	Lowest (freq)	Avg	Grand Avg	Highest (freq)	Lowest (freq)	Avg	Grand Avg
RS + IBPLN	IMP-1	100(12)	80(1)	90.08	93.07	100(12)	51(1)	89.02	91.03
	IMP-2	100(22)	83.33(3)	93.72		100(14)	53(1)	91.9	
	IMP-3	100(29)	80.95(1)	92.99		100(21)	54(2)	91.96	
	IMP-4	100(23)	77.27(1)	94.53		100(14)	56(1)	90.59	
	IMP-5	100(20)	78.95(1)	94.03		100(10)	70(1)	91.67	
RS + LM	IMP-1	100(20)	84.21(1)	93.41	**94.85**	100(18)	74(4)	90.18	**92.43**
	IMP-2	100(37)	83.33(3)	95.65		100(23)	70(1)	92.32	
	IMP-3	100(36)	85(2)	94.68		100(30)	52(1)	91.18	
	IMP-4	100(32)	85(2)	95.43		100(27)	71(2)	93.77	
	IMP-5	100(26)	83.33(1)	94.85		100(19)	79(1)	94.7	

Table IV C. Results of Approach-II Results from100 simulations

Methods	Test set (CCR%)			Specificity		
	Highest (freq)	Lowest (freq)	Avg	Highest (freq)	Lowest (freq)	Avg

RS + IBPLN	100(9)	73.91(1)	92.17	100(43)	28.57(1)	80.49
RS + LM	100(17)	86.96(1)	**94.61**	100(49)	33.33(1)	**85.17**

TABLE IV D. Results of Approach-II Results from100 simulations

Methods	Sensitivity			Area Under ROC Curve (AUC)		
	Highest (freq)	Lowest (freq)	Avg	Highest (freq)	Lowest (freq)	Avg
RS + IBPLN	100(34)	72.2(1)	94.78	100(20)	80(2)	94.7
RS + LM	100(41)	89.47(1)	**96.60**	100(34)	81(1)	**95.66**

Table V. Comparison of accuracies of different methods applied in hepatitis disease diagnosis

Hepatitis disease	KNN	Naïve Bays	SVM	FDT	PSO	CBR-PSO	RS + LM Approach-II (present study)
Best	89.86%	86.35%	90.31%	78.15%	89.46%	94.58%	**100%**
Lowest	70.29%	66.94%	65.22%	61.49%	75.35%	77.16%	**86.96%**
Average	83.45%	82.05%	86.92%	75.39%	82.66%	92.83%	**94.61%**

8. Conclusion

This work implements RS based feature selection techniques to extract features. Imputed files obtained as a result of multiple imputations were combined into a single file and Levenberg Marquardt algorithm was used as classifier to obtain a better result in comparison to other results including Neshat et al. [4]. Lowest performance is also recorded in this work as in medical diagnosis it should also be a judging parameter for performance prediction. Much of the previous literatures do not clearly specify about the result, whether it was a best simulation product or average of several simulations. So, the highest, lowest and average performance of the methods are presented in this work.

We suggest that a combination of seven reduced features (Age, Steroid, Billirubin, Alk Phosphate, SGOT, Albumin, Protien) obtained by implementing rough set based feature selection technique, might be important to doctors while taking final decision. The techniques used here could be applied to other diseases.

References

1. K. Polat and S. Gunes, Hepatitis disease diagnosis using a new hybrid system based on feature selection (FS) and artificial immune recognition system with fuzzy resource allocation, Digital Signal Processing, vol. 16, no. 6, pp. 889-901, 2006.
2. K. Rezaee et al., An Intelligent Diagnostic System for Detection of Hepatitis using Multi-Layer Perceptron and Colonial Competitive Algorithm, The J. of Mathematics and Computer Science, vol. 4, no. 2, pp. 237-245, 2012.
3. D. Calisir and E. Dogantekin, A new intelligent hepatitis diagnosis system: PCA-LSSVM, Expert Systems with Applications, vol. 38, pp. 10705-10708, 2011.
4. M. Neshat, M. Sargolzaei, A. N. Nadjaran, and A. Masoumi, Hepatitis Disease Diagnosis using Hybrid Case Based Reasoning and Particle Swarm Optimization, ISRN Artificial Intelligence, 2012, doi: 10.5402/2012/609718.
5. M. Mitra and R. K. Samanta, Hepatitis Disease Diagnosis Using Multiple Imputation and Neural Network with Rough Set Feature Reduction, Advances in Intelligent Systems and Computing 327, pp. 285-293, FICTA -2014, Springer International Publishing Switzerland 2015.
6. http://archieve.ics.uci.edu/ml/datasets/Hepatitis.
7. A. P. Dempster, N. M. Laird, D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm", Journal of the Royal Statistical Society. Series B, 1977, vol. 39, no. 1, pp. 1–38.
8. J. Honaker, G. King, “What to do About Missing Values in Time Series Cross-Section Data”, American J. of Political Science, vol. 54, no. 2, pp.561-581, 2010.
9. D. Calisir, and E. Dogantekin, A new intelligent hepatitis diagnosis system: PCA-LSSVM, Expert Systems with Applications, vol. 38, pp. 10705-10708, 2011.
10. E. Avci, A new optimum feature extraction and classification method for speaker recognition: GWPNN, Expert Systems with Applications, vol. 32, no2, pp. 485-498, 2007.
11. Z. Pawlak, Rough sets, Int. J. of Parallel Programming, vol. 11, no. 5, pp. 341-356, 1982.
12. D. O. Hebb, The Organization of Behavior, a Neuropsychological Theory, New York, John Wiley, 1949.
13. A. Roy, Artificial Neural Networks- A Science in Trouble, SIGKDD Explorations, vol. 1, issue 2, pp. 33-38, 2000.
14. D. E. Rumelhart, J. L. McClelland (eds.), Parallel Distributed Processing: Explorations in Microstructures of Cognition, vol. 1: Foundations, MIT Press, Cambridge, M.A., pp. 318-362, 1986.

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

16. S. Grossberg, Nonlinear Neural Networks: Principles, Mechanisms, and Architectures, Neural Networks, vol. 1, pp. 17 -61, 1988.

17. J. Moody and C. Darken, Fast Learning in Networks of Locally-Tuned Processing Units, Neural Computation, vol. 1, pp. 281-294, 1989.

18. L. Fu, H. Hsu, and J. C. Principe, Incremental Backpropagation Learning Networks, IEEE Trans. on Neural Networks, vol. 7, no.3, pp. 757-761, 1996.

19. 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.

20. K. Levenberg, A method for the solution of certain non-linear problems in least squares, Quarterly in Applied Mathematics, vol. 2, no. 2, pp. 164-168, 1944.

21. D. W. Marquardt, An algorithm for the least-squares estimation of nonlinear parameters, SIAM Journal of Applied Mathematics, vol. 11. No. 2, pp. 431-441, 1963.

22. Dr. Amitava Basu, MD, Pathologist, India (personal communication).

23. E. Hall, G. Frank, B. Holmes, P. Pfahringer, I. Reutemann, H. Witten, The Weka Data Mining Software: An Update, SIGKDD Explorations, vol. 11, no. 1, 2009.

24. K. Hornik, M. Stinchcombe, H. White, Multilayer feedforward networks are universal approximator, Neural Network, Vol.2, pp. 359-366, 1991.

25. D. Goa, On structures of supervised linear basis function feedforward three-layered neural networks, Chin. J. Comput., vol. 21, no. 1, pp. 80-86, 1998.

26. 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, pp. 865-873, 2010.

27. Alyuda NeuroIntelligence 2.2, http://www.alyuda.com.

28. A. P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recognition, vol. 30, no. 7, pp. 1145-1159, 1997.