Journal of Information Engineering and Applications
ISSN 2224-5782 (print) ISSN 2225-0506 (online)
Vol.3, No.4, 2013
www.iiste.org

Comparative Study of Artificial Neural Network based Classification for Liver Patient

Anil Kumar Tiwari[1*] Lokesh Kumar Sharma[2] G. Rama Krishna[3]

1. Disha College Raipur, Chhattisgarh,492001, India
2. National Institute of Occupational Health, Ahmedabad, 380016, India
3. Department of Computer Science and Engineering, K L University, Vijayawada, 520002, India

* E-mail of the corresponding author: anil1969_rpr@yahoo.com

Abstract

The extensive accessibility of new computational methods and tools for data analysis and predictive modeling requires medical informatics researchers and practitioners to steadily select the most appropriate strategy to cope with clinical prediction problems. Data mining offers methodological and technical solutions to deal with the analysis of medical data and construction of prediction models. Patients with Liver disease have been continuously increasing because of excessive consumption of alcohol, inhale of harmful gases, intake of contaminated food, pickles and drugs. Therefore, in this study, Liver patient data is considered and evaluated by univariate analysis and a feature selection method for predicator attributes determination. Further comparative study of artificial neural network based predictive models such as BP, RBF, SOM, SVM are provided.

Keywords: Medical Informatics, Classification, Liver Data, Artificial Neural Network

1. Introduction

Data mining has been increasingly used in the medical literature over the last few years. In general, the term has not been anchored to any precise definition but to some sort of common understanding of its meaning: the use of methods and tools to analyze large amounts of data. Its application to the analysis of medical data – notwithstanding high hopes – has until recently been relatively limited. This is particularly true of practical applications in clinical medicine which may benefit from specific data mining approaches that are able to perform predictive modelling, exploit the knowledge available in the clinical domain and explain proposed decisions once the models are used to support clinical decisions. The goal of predictive data mining in clinical medicine is to derive models that can use patient specific information to predict the outcome of interest and to thereby support clinical decision-making. Predictive data mining methods may be applied to the construction of decision models for procedures such as prognosis, diagnosis and treatment planning, which – once evaluated and verified – may be embedded within clinical information systems (Bellazzi & Zupan 2008).

The aim of this study is to provide a comparative framework among different neural network based predicative data mining techniques for diagnosing Liver disorder. The work is organized as follows. In the section 2, related works are reported. In the section 3 materials and methods are presented. The results of the experiments are presented in section 4. Final, this paper is concluded in section 5.

2. Related Works

A literature survey showed that there have been several studies on the survivability prediction problem using statistical approaches and artificial neural networks. However, a few studies related to medical diagnosis and survivability using data mining approaches have been reported.

Bellaachia and Guven (2006) presented an analysis of the prediction of survivability rate of breast cancer patients' using data mining techniques. The data used is the SEER Public-Use Data. The investigated three data mining techniques: the Naïve Bayes, the back-propagated neural network, and the C4.5 decision tree algorithms. Experiments were conducted using these algorithms. The achieved prediction performances are comparable to existing techniques. However, we found out that C4.5 algorithm has a much better performance than the other two techniques. Sug (2012) applied decision trees, C4.5 and CART to Liver Disorder Disease. Ramana et al. (2011) applied the classification algorithms considered here are Naïve Bayes, classifier, C4.5, Back propagation Neural Network algorithm, and Support Vector Machines for liver patients. In this study, they considered and men and women in single data unit. Also the children are considered in same data sample. The adult men, adult women and children data may be considered separately due to different physiological structures. Also, significant differences between means among women and men groups were measured. Therefore, in this work, the men and women are considered in different data set and more refined approaches RBF, SOM algorithms are applied.

3. Materials and Methods

3.1 Data Set

In this study, Indian Liver Patient Dataset (ILPD) (Frank & Asumcion 2010) was used. The data set was collected from north east of Andhra Pradesh, India (Ramana et al. 2012; Ramana et al. 2011). This data set comprises Age of the patient, Total Bilirubin, Direct Bilirubin Alkaline Phosphotase, Alamine Aminotransferase, Aspartate Aminotransferase, Total Protiens, Albumin, Albumin and Globulin Ratio, class labeled by the experts. Selector is a class label used to divide into groups (liver patient or not). In this work, the noisy data were ignored and only the age greater than 17 were considered. Men data set contains 316 liver patient and 105 non liver patient after ignoring the noisy and age less than 18. Similarly, women data set contains 86 liver patient and 47 non liver patient records. The detail descriptive statistics for Men and Women data set are given in the Table 1 and 2.

Table 1. Descriptive Statistics for Men N= 421

	Min	**Max**	**Mean**	**Skew**	**Kurtosis**
Age	18	90	46.7±15.1	0.08	-0.82
Total Bilirubin	0.4	75	3.7±6.7	4.84	35.89
Direct Bilirubin	0.1	19.7	1.7±3.0	3.01	10.15
Total Protein	75	2110	284.1±224.3	3.94	21.05
Albumin	10	2000	90.7±206.7	5.98	40.78
A/G Ratio	11	4929	125.9±330.4	9.57	120.18
SGPT	2.7	9.6	6.4±1.1	-.279	0.526
SGOT	0.9	5.5	3.1±0.8	0.02	-0.372
Alkphos	0.3	2.8	0.94±0.3	1.12	3.83

Table 2. Descriptive Statistics for Women N= 133

	Min	**Max**	**Mean**	**Skew**	**Kurtosis**
Age	18	85	44.8±14.7	0.37	-0.22
Total Bilirubin	0.5	27.7	2.2±4.5	4.1	17.03
Direct Bilirubin	0.1	12.8	0.96±2.2	3.98	15.88
Total Protein	63	1896	296.7±293.5	3.46	12.95
Albumin	10	509	50.7±72.6	4.09	19.40
A/G Ratio	10	623	63.7±100.1	3.85	16.31
SGPT	3.6	9.2	6.63±1.1	-0.313	-0.39
SGOT	1.0	5.5	3.25±0.8	-0.14	-0.14
Alkphos	0.3	1.8	0.94±0.3	0.69	1.22

3.2 Methods

Artificial Neural Network (ANN) is widely used algorithms in many applications and modeled based on biological neural systems. For classification, the model is trained in such a way that input and output nodes are connected with a set of weighted links based on input–output association of training data. More complex ANNs can have one or more hidden layers with hidden nodes in between input and output nodes to model more complex input–output relationships. Depending upon the application, either a feed-forward (nodes in one layer is connected only to nodes in next layer) or recurrent (nodes can be connected to nodes in the same layer, previous layer or next layer) network is built. The relationship between input and output is defined by an activation function which along with the weighted links decides the behaviour of the ANN. The activation function can be linear, sigmoid (logistic) and hyperbolic tangent function (Kotsiantis 2007). In this paper, back propagations (BP) learning, radial basis function networks (RBF), self organizing map (SOM), and support vector machine (SVM) methods are used to classify the Liver patient data.

Back Propagation is a multi layer preceptron neural network. The term back propagation means the backward propagation of an error signal through the network. After propagating a pattern through the network – feed forward, the output pattern is compared with a given target and the error of each output unit is calculated. This error is propagated backwards to the input layer – back propagation. Finally the errors of the units are used to modify the weights. The back propagation algorithm is based on Widrow-Hoff delta learning rule in which the weight adjustment is done through mean square error of the output response to the sample input (Sivanandam et al. 2006).

High-dimensional input patterns with prototype lattice structure are represented by SOM and can be visualized in two-dimensional lattice structure. Each unit in the lattice (neuron) and adjacent neurons are interconnected, which gives the clear topology of how the network fits itself to the input space. Input patterns are fully connected to all neurons via adaptable weights, and during the training process, neighboring input patterns are projected into the lattice, corresponding to adjacent neurons (Sivanandam et al. 2006). SOM is also used for supervised learning (Salah et al. 2009).

RBF is a special type of neural networks with several distinctive features. A RBF network consists of three layers, namely the input layer, the hidden layer, and the output layer. The input layer broadcasts the coordinates of the input vector to each of the units in the hidden layer. Each unit in the hidden layer then produces an activation based on the associated radial basis function. Finally, each unit in the output layer computes a linear combination of the activations of the hidden units. How a RBF network reacts to a given input stimulus is completely determined by the activation functions associated with the hidden units and the weights associated with the links between the hidden layer and the output layer (Oyang et al. 2005). There are several types of radial basis function which can be used for classification tasks such as Gaussian radial function, Normalized Gaussian radial function, Thin plate spline, Quadratic, Inverse quadratic (Fornberg & Piret 2008). In this work, Normalized Gaussian radial function was used.

SVMs, like other nonparametric classifiers such as Artificial Neural Networks, boast a robustness that has spearheaded its application into many areas. Like other supervised classifiers, training data is a prerequisite to define the decision boundaries within the feature space, based upon which classification decision rules are made. For SVMs, this decision boundary is a linear discriminate placed midway between the classes of interest. SVMs handle nonlinear datasets by use of a kernel function. Some examples of functions (also called kernels) used to this effect include: polynomial, Gaussian (more commonly referred to as radial basis functions) and sigmoid functions. Each function has parameters which have to be determined prior to classification and they are usually determined through a cross validation process. Operating in high dimension potentially renders the risk of over-fitting in the input space possible. SVMs control this through the principle of Structural Risk Minimization. The empirical risk of misclassification is controlled by maximizing the margin between the training data and the decision boundary (Anthony et al. 2008; Vapnik 1995).

4. Experimental Investigation and Results

Initially, the predicator attributes for the classification task was determined by computing significant differences between classes; in this regard, the univariate analysis of variance was applied. Also, correlation-based feature subset selection method was used to evaluate the predicator attributes (Hall et al. 2009; Hall 1999). Total Bilirubin, Direct Bilirubin, Total Protein, Albumin, A/G Ratio, SGOT, and Alkphos were found significant difference in men data set by univariate analysis. Total Bilrubin, Direct Bilirubin, Total Protein and A/G Ratio were found significant difference in women data set by univariate analysis. The correlation-based feature subset selection method selected Total Bilirubin, Direct Bilirubin, Total Protein, Albumin, A/G Ratio in case of men data and Total Bilirubin, Direct Bilirubin, Total Protein, A/G Ratio in case of women data. In this work, common attributes between univariate analysis and feature subset selection method were considered for classification task in both cases of men and women data set. The artificial networks classifiers i.e. BP, SOM, RBF, and SVM were trained and tested. The experiments were conducted in Weka 3.7 (Hall et al. 2009). The summary of results in case of men and women are shown on Table 3 and 4, respectively. Our experimental investigation yields a significant output in terms of the correctly classified success rate. The SVM provided high correctly classified success rate 99.76% and 97.7% in case of men and women data.

Table 3. Summary of classification result for Men N= 421

Parameters	BP	SOM	RBF	SVM
Accuracy %	74.8	75.05	75.8	99.76
Mean absolute error	0.33	0.24	0.30	0.002
Root mean squared error	0.40	0.49	0.39	0.05
Relative absolute error %	88.0	66.50	81.48	0.63
Root relative squared error %	92.29	115.42	90.30	11.26
Coverage of cases (0.95 level) %	99.29	75.05	99.52	99.76

Table 4. Summary of classification result for Women N= 113

Parameters	BP	SOM	RBF	SVM
Accuracy %	64.7	64.7	66.2	97.7
Mean absolute error	0.40	0.35	0.38	0.02
Root mean squared error	0.44	0.59	0.44	0.15
Relative absolute error %	88.4	77.21	83.1	4.9
Root relative squared error %	93.4	124.4	91.2	31.4
Coverage of cases (0.95 level) %	100	64.7	100	97.7

5. Conclusion

Predictive data mining is becoming an important instrument for researchers and clinical practitioners in medicine. Understanding the main issues underlying these methods and the application of agreed and standardized procedures is mandatory for their deployment and the dissemination of results. In this study, the Liver data was evaluated by univariate analysis and feature selection methods and predicator attributes are determined. Further, ANN based classifiers were applied on selected attributes. It shows the ANN classifiers may be used as patient predicator tool.

References

Anthony, G., Gregg, H. & Tshildzi, M. (2008), "An SVM Multiclassifier Approach to Land Cover Mapping", ASPRS 2008, Portland Oregon.

Bellaachia, A. & Guven, E. (2006), "Predicting Breast Cancer Survivability using Data Mining Techniques", Sixth SIAM Int. Conference on Data Mining (SDM'06).

Bellazzi, R. & Zupan, B. (2008), "Predictive data mining in clinical medicine: Current issues and guidelines", International Journal of Medical Informatics 77, 81–97.

Fornberg, B. & Piret, C. (2008), "On choosing a radial basis function and a shape parameter when solving a convective PDE on a sphere", Journal of Computational Physics, 227, 2758–2780.

Frank, A. & Asuncion, A. (2010), "UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]", Irvine, CA: University of California, School of Information and Computer Science.

Hall, M. A. (1999), "Correlation-based Feature Subset Selection for Machine Learning", Ph.D. Thesis, University of Waikato, Hamilton, New Zealand.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009), "The WEKA Data Mining Software: An Update", SIGKDD Explorations, 11(1), 10-18.

Keerthi, S.S., Shevade, S.K., Bhattacharyya, C. & Murthy, K.R.K. (2001), "Improvements to Platt's SMO Algorithm for SVM Classifier Design", Neural Computation. 13(3), 637-649.

Kotsiantis, S. B. (2007), "Supervised Machine Learning: A Review of Classification Techniques", Informatica, 31, 249-268.

Oyang , Y. J. et al. (2005), "Data Classification with Radial Basis Function Networks Based on a Novel Kernel Density Estimation Algorithm", IEEE Transactions on Neural Networks, 16(1), 225 – 236.

Ramana ,B. V., Prasadbabu, M. S. & Venkateswarlu, N. B. (2012), "A Critical Comparative Study of Liver Patients from USA and INDIA: An Exploratory Analysis", International Journal of Computer Science Issues, 9(3), 506-516.

Ramana, B. V., Prasadabu, M. S. & Venkateswarlu, N. B. (2011), "A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis", International Journal of Database Management Systems (IJDMS), 3(2), 101-114.

Salah M., Trinder, J. & Shaker, A. (2009), "Evaluation of the Self Organizing Map Classifier for Buildinng Detection from Lidar Data and Multispectral Arial Images", Spatial Science, 54(2), 15-34.

Sivanandam, S. N., Sumath, S. & Deepa, S. N. (2006), "Introduction to Neural Networks Using Matlab 6.0", Tata McGraw-Hill Education.

Sug, H. (2012), "Improving the Prediction Accuracy of Liver Disorder Disease with Oversampling", Applied Mathematics in Electrical and Computer Engineering, 331-335.

Vapnik, V. N. (1995), "The Nature of Statistical Learning Theory", Springer-Verlag New York.