Volume 4, Issue 2, February 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Liver Patient Classification using Intelligence Techniques

Jankisharan Pahareeya
Deptt. of Info-Tech
Rustmji Inst. of Tech.
BSF Academy, Tekanpur
India

Rajan Vohra
Deptt. of CSE
PDM College of Engg
Bahadurgarh, Haryana
India

Jagdish Makhijani
Deptt. of Info-Tech
Rustmji Inst. of Tech.
BSF Academy, Tekanpur
India

Sanjay Patsariya
Deptt. of Info-Tech
Rustmji Inst. of Tech.
BSF Academy, Tekanpur
India

Abstract- This paper presents computational intelligence techniques for Liver Patient Classification. Number of Liver Patients increasing day by day because excessive consumption of alcohol, inhale of harmful gases, intake of contaminated food and drugs. The efficacy of these techniques viz.Multiple Linear Regression, Support Vector Machine, Multilayer FeedForward Neural Network,J-48,Random Forest and Genetic Programming has been tested onTheILPD (Indian Liver Patient Dataset) Data Set. Since it is an unbalanced dataset with 72% liver patient and 28% non liver patient. We employed under sampling and oversampling for balancing it. Ten-fold cross validation is performed throughout the study.The results obtained from our experiments indicate that Random Forest oversampling with 200% outperformed all the other techniques.

Keyword- J-48, Multilayer FeedForward Neural Network, Random Forest, Multiple Linear Regression (MLR), Support Vector Machine (SVM) and Genetic programming (GP).

I. Introduction

Classification techniques are very popular in various automatic medical diagnoses. Problems with liver patients are not easily discovered in an early stage as it will be functioning normally even when it is partially damaged [1]. An early diagnosis of liver problems will increase patient's survival rate. Liver disease can be diagnosed by analyzing the levels of enzymes in the blood [2]. Moreover, now a day's mobile devices are extensively used for monitoring human's body conditions. Here also, automatic classification algorithms are needed. With the help of Automatic classification tools for liver diseases (probably mobile enabled or web enabled).So, the results of this study are very important for the development of automatic medical diagnosis system in future. So, that one can reduce the patient queue at the liver experts such as endocrinologists. In this project, selected classification algorithms were considered from Different category of classification algorithms. These algorithms are J48, MLP (Multilayer Perception), RF(Random Forest), MLR (Multiple linear Regression), Genetic programming(GP) and Support vector machine(SVM) have been considered for comparing their performance based on the *ILPD (Indian Liver Patient Dataset).*

The rest of this paper is organized as follows. Section 2 related work done in the field of software fault prediction. Section 3 overviews the data description and data preparation, Section 4 overviews of the techniques applied in this paper, section 5 presents the results and discussions. Finally, Section 6 concludes the paper.

II. Related work

Kun- Hong Liu and De-Shuang Huang [3] addressed the microarray dataset based cancer classification using rotation forest. Principalcomponent analysis (PCA) was applied to feature transformation in the original rotation forest. The experimental results shows that ICA improves the performance of rotation forest compared with original transformation methods.Juan J. Rodriguez et al. [4] proposed a method for generating classifier ensembles based on feature extraction. The idea of the rotation approach is to encourage simultaneously individual accuracy and diversity within the ensemble. Akin Ozcift and Arif Gultenb[5] constructed rotation forest (RF) ensemble classifiers of 30 machine learning algorithms to evaluate their classification performances using Parkinson's, diabetes and heart diseases from literature. Experiments demonstrate that RF, as a newly proposed ensemble algorithm, proves itself to be efficient to increase classifier accuracies significantly.BendiVenkataRamana et al. [6] compared popular Classification Algorithms for evaluating their classification performance in terms of Accuracy, Precision, Sensitivity and Specificity in classifying liver patients dataset.BendiVenkataRamana et al. [7] proposed Bayesian Classification for diagnosis of liver diseases. The Bayesian Classification is combined with Bagging and Boosting for better accuracy. This accuracy can be further improved with huge amount of data. BendiVenkataRamana et al. [8] proposed ANOVA and MANOVA for population comparison between ILPD data set and UCI data set. The results indicates that there exists more significant difference in the groups with all the possible attribute combinations except analysis on SGPT between non liver patients of UCI and INDIA data sets. BendiVenkataRamana et al. [9] proposed Bayesian classification for accurate liver disease diagnosis and its accuracy further improved using.Unfortunately the accuracy of these models is not satisfactory so there is always a scope for new classification models.

III. DATA DESCRIPTION DATA PREPARATION

This dataset we download from UCI machine Learning Repository (http://archive.ics.uci.edu/ml/). Entire ILPD (Indian Liver Patient Dataset) dataset contains information about 583 Indian liver patients. In which 416 are liver patient records and 167 non liver patient records. The data set was collected from north east of Andhra Pradesh, India. Selector is a class label used to divide into groups (liver patient or not). The first cleaning step was to remove the projects having null values. So after cleaning null records we have now 579 Indian liver patients. In which 414 are liver patients and 165 are non liver patient. Finally, we normalized the data set.

IV. OVERVIEW OF THE TECHNIQUES EMPLOYED

The following techniques are applied to predict the Liver Patient:J-48, MLP (Multi Layer Perceptron), Random Forest(RF), MLR (Multiple linear Regressions), *Support Vector Machine(SVM),* Genetic programming (GP).

A. J-48 :

C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan.[10] C4.5 is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier.C4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. The training data is a set $S = s_1, s_2, \dots$ of already classified samples. Each sample s_i consists of a p-dimensional vector $(x_{1,i}, x_{2,i}, \dots, x_{p,i})$, where x_j represent attributes or features of the sample, as well as the class in which s_i falls.

This algorithm has a few base cases.

- All the samples in the list belong to the same class. When this happens, it simply creates a leaf node for the decision tree saying to choose that class.
- None of the features provide any information gain. In this case, C4.5 creates a decision node higher up the tree using the expected value of the class.
- Instance of previously-unseen class encountered. Again, C4.5 creates a decision node higher up the tree using the expected value.

We used Weka tool for J-48, RF, MLP and SVM, implementation available at http://www.cs.waikato.ac.nz/~ml/weka/downloading.html

B. MLP (Multilayer Perceptron):

A multilayerperceptron (MLP) is a feedforward artificial neural network model that maps sets of input data onto a set of appropriate outputs. An MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Except for the input nodes, each node is a neuron (or processing element) with a nonlinear activation function. MLP utilizes a supervised learning technique called back propagation for training the network[11] .MLP is a modification of the standard linear perceptron and can distinguish data that are not linearly separable.[12]

The multilayer perceptron consists of three or more layers (an input and an output layer with one or more *hidden layers*) of nonlinearly-activating nodes. Each node in one layer connects with a certain weight w_{ij} to every node in the following layer. Some people do not include the input layer when counting the number of layers and there is disagreement about whether w_{ij} should be interpreted as the weight from i to j or the other way around.

C. Random Forest (RF):

Random forests are an ensemble learning method for classification (and regression) that operate by constructing a multitude of decision trees at training time and outputting the class that is themode of the classes output by individual trees. The algorithm for inducing a random forest was developed by Leo Breiman[13] and Adele Cutler, and "Random Forests" is their trademark.

D. MLR (Multiple linear Regression):

In statistics, linear regression is an approach to model the relationship between a scalar dependent variable *y* and one or more explanatory variables denoted *X*. The case of one explanatory variable is called *simple linear regression*. For more than one explanatory variable, it is called *multiple linear regression.*

Student version of the NewCom tool obtains at http://www.kedri.aut.ac.nz/areas-of-xpertise/data-mining-and-decision-support-systems/neucom#download.

E. Support Vector Machine (SVM):

The SVM is a powerful learning algorithm based on recent advances in statistical learning theory proposed by Vapnik [14, 15]. SVR is a learning system that uses a hypothesis space of linear functions in a high dimensional space, trained with a learning algorithm from optimization theory that implements a learning bias derived from statistical learning theory. SVR uses a linear model to implement non-linear class boundaries by mapping input vectors non-linearly into a high dimensional feature space using kernels. The training examples that are closest to the maximum margin hyperplane are called support vectors. All other training examples are irrelevant for defining the binary class boundaries. The support vectors are then used to construct an optimal linear separating hyperplane (in case of pattern recognition) or a linear regression function (in case of regression) in this feature space. The support vectors are conventionally determined by solving a quadratic programming (QP) problem.

F. Genetic programming (GP):

Genetic programming (GP) [16, 17] is an extension of genetic algorithms (GA). It is a search methodology belonging to the family of evolutionary computation (EC). GP mainly involve functions and terminals. GP randomly generates an initial population of solutions. Then, the initial population is manipulated using various genetic operators to produce new populations. These operators include reproduction, crossover, mutation, dropping condition, etc. The whole process of evolving from one population to the next population is called a generation. A high-level description of GP algorithm can be divided into a number of sequential steps [18,19]:

1. Create a random population of programs, or rules, using the symbolic expressions provided as the initialpopulation.
2. Evaluate each program or rule by assigning a fitness value according to a predefined fitness function that can measure the capability of the rule or program to solve the problem.
3. Use reproduction operator to copy existing programs into the new generation.
4. Generate the new population with crossover, mutation, or other operators from a randomly chosenset of parents.
5. Repeat steps 2 onwards for the new population until a predefined termination criterion has been satisfied, or a fixed number of generations has been completed.
6. The solution to the problem is the genetic program with the best fitness within all the generations.

In GP, crossover operation is achieved first by reproduction of two parent trees. Two crossover pointsare then randomly selected in the two offspring trees. Exchanging sub-trees, which are selected according to the crossover point in the parent trees, generates the final offspring trees. The obtained offspring trees are usually different from their parents in size and shape. Then, mutation operation is also considered in GP. A single parental tree is first reproduced. Then a mutation point is randomly selected from the reproduction, which can be either a leaf node or a sub-tree. Finally, the leaf node or the sub-tree is replaced by a new leaf node or sub-tree generated randomly. Fitness functions ensure that the evolution goes toward optimization by calculating the fitness value for each individual in the population. The fitness value evaluates the performance of each individual in the population.

GP is guided by the fitness function to search for the most efficient computer program to solve a given problem. A simple measure of fitness [18] is adopted for the binaryClassification problem which is given as follows.

Fitness (T) = no of samples classified correctly/ no of samples for training during evaluation

We used the GP implementation available at http://www.rmltech.com

V. Results and Discussion

In this study we used liverpatient data sets from *ILPD (Indian Liver Patient) Data Set*. It has 579 samples with 10 independent variables and one dependentvariable.*Since it is an unbalanced dataset with 72% liver patient and 28% non liver patient. We employed under sampling and oversampling for balancing it*. Under-sampling (25%) means that the majority class is reduced to 25% of its original size. Then, over-sampling is a technique in which the samples belonging to the minority class are replicated a few times and combined with the majority class samples. For example, over-sampling (100%) means that the majority class is replicated once.

We compared the performance of the Classification models on the basis of Accuracy, which defined as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

The results are presented in Tables from 2 to 6.

Table 2
Average Accuracy of 10 -fold cross validation for the original data

SN.	Method	Accuracy(Test)
1	J48	68.3938
2	MLP	69.4301
3	SVM	71.5026
4	MLR	71.9593
5	Random Forest	71.5026
6	**GP**	**84.75**

TABLE 3
AVERAGE ACCURACY OF 10 FOLD CROSS VALIDATION FOR THE OVER SAMPLING (100%)

SN.	**Method**	**Accuracy(Test)**
1	J48	76.177
2	MLP	67.9677
3	SVM	68.64
4	MLR	68.2697
5	**Random Forest**	**83.9677**
6	GP	80

TABLE 4
AVERAGE ACCURACY OF 10 FOLD CROSS VALIDATION FOR THE OVER SAMPLING (200%)

SN.	Method	Accuracy(Test)
1	**J48**	87.2387
2	**MLP**	72.3872
3	**SVM**	72.0272
4	**MLR**	71.0635
5	**Random Forest**	**89.1089**

TABLE 5
AVERAGE ACCURACY OF 10 FOLD CROSS VALIDATION FOR THE UNDER SAMPLING (25%)

SN.	**Method**	**Accuracy(Test)**
1	J48	60.3865
2	MLP	61.5942
3	SVM	62.4697
4	MLR	60.151
6	**GP**	**82.61**

TABLE 6
AVERAGE ACCURACY OF 10 FOLD CROSS VALIDATION FOR THE UNDER SAMPLING (50%)

SN.	**Method**	**Accuracy(Test)**
1	J48	76.3052
2	MLP	72.6908
3	SVM	72.6908
4	MLR	70.7166
5	Random Forest	71.0843
6	**GP**	**84.75**

We observed that Random Forest Over Sampling (200%) has outperformed all other techniques with 89.1089 Accuracy.

VI. CONCLUSIONS

We tested six intelligence techniques on the *ILPD (Indian Liver Patient) Data Set.* Throughout the study ten-fold cross validation is performed. The proposed Random Forest Over Sampling (200%) model outperformed all other techniques. We notice that GP for the original data and GP for the under sampling 50% stood second with an Accuracy value of 84.75. Hence we conclude thatthe Random Forest over sampling (200%) model is the relatively best predictor among all other techniques.

In further work, profiling of Liver Patients can be done, along with Prediction of Disease. In addition Distribution of the patients can be generated to identify sensitive geographical areas where preventive measures can be focused.

Also Demographic analysis can be done to identify segments of population needing greater attention along with locating suitable medical infrastructure and specialists for the same. The same paradigm can be done for other medical diseases most prevalent in a particular region. This study is most topical and of high utility for health care planners and service providers.

ACKNOWLEDGEMENT

The authors would like to thank the UCI Machine Learning Repository for providing the ILPD (Indian Liver Patient Dataset) Data Set.

REFERENCES

[1]. Rong-Ho Lin, “An intelligent model for liver disease diagnosis”, Artificial Intelligence in Medicine, Vol.47 (2009), PP. 53-62.

[2]. Eugene R. Schiff, Michael F. Sorrell, Willis C. Maddrey, Lippincott Williams and Wilkins, "Schiff's Diseases of the Liver", 10th Edition, 2007

[3]. Kun-Hong Liu and De-Shuang Huang: “Cancer classification using Rotation forest”. In Proceedings of the Computers in biology and medicine, Vol. 38(2008), PP. 601-610.

[4]. Juan J. RodriGuez, Member, Ludmila I. Kuncheva and Carlos J. Alonso, “Rotation Forest: A New Classifier Ensemble Method”. In Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 28(2006), PP. 1619-1630.

[5]. Akin Ozcift and ArifGulten, “Classifier Ensemble Construction With Rotation Forest To Improve Medical Diagnosis Performance Of Machine Learning Algorithms”, In Proceedings of the Computer Methods and Programs in Biomedicine,Vol.104 n.1(2011), PP. 443-451.

[6]. BendiVenkataRamana, Prof. M.Surendra Prasad Babu and Prof. N. B. Venkateswarlu, “ A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis”, International Journal of Database Management Systems (IJDMS), Vol.3,No.2(2011),PP.101-11.

[7]. BendiVenkataRamana, Prof.M.Surendra Prasad Babu and Prof. N. B. Venkateswarlu, “A Critical Evaluation of Bayesian Classifier For Liver Diagnosis Using Bagging and Boosting Methods”, International Journal of Engineering Science and Technology (IJEST), Vol.3 (2011), PP. 3422-3426.

[8]. BendiVenkataRamana, Prof. M. S. Prasad Babu and Prof. N. B. Venkateswarlu, “A Critical Comparative Study of Liver Patients From USA and INDIA: An Exploratory Analysis”, International Journal of Computer Science Issues 3, Vol.9 (2012).

[9]. BendiVenkataRamana, Prof. M. S. Prasad Babu and B. R. Sarathkumar, “New Automatic Diagnosis of Liver Status Using Bayesian Classification”, IEEE International Conference on Intelligent Network and Computing (ICINC), PP. 385-388(2010)

[10]. J. R. Quinlan, “C4.5: Programs for Machine Learning”, Morgan Kaufmann Publishers, 1993.

[11]. F.Rosenblatt, “Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms”, Spartan Books, Washington DC, 1961.

[12]. G. Cybenko, “Approximation by superpositions of a sigmoidal function”, *Mathematics of Control, Signals, and Systems*, Vol.2(1989), PP. 303-314.

[13]. Leo Breimen, "Random Forests", *Machine Learning* ,Vol. 45(2001), PP.5-32

[14]. V.N. Vapnik, “Statistical Learning Theory”, John Wiley,New York, 1998.

[15]. Jankisharan Pahariya, V. Ravi and M. Carr, “Software Cost Estimation Using Computational Intelligent Techniques”,International Conference on Computer Information System and Industrial Management Application, Coimbatore, Tamil Nadu, India(2009),PP.849-854.

[16]. R. Poli, W.B. Langdon and J.R. Koza, “A field guide to Genetic Programming”, publisher- Lulu.com, United Kingdom, 2008.

[17]. J. R. Koza, “Genetic Programming: On the programming of computers by means of natural selection”, Cambridge, MA: MIT press, 1992.

[18]. K. M. Faraoun and A. Boukelif, “Genetic programming approach for multi-category pattern classification applied to network intrusion detection”, International Journal of Computational Intelligence and Applications, Vol. 6(2006), PP. 77-99.

[19]. J.S.Pahariya, V. Ravi, M. Carr and M.Vasu,“ Computational Intelligence Hybrids Applied to Software Cost Estimation”, International Journal of Computer Information Systems and Industrial Management Applications, Vol.2 (2010), pp.104-112.