Volume 3, Issue 3, March 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Comparative Study of Different Data Mining Techniques Performance in knowledge Discovery from Medical Database

Olaiya Folorunsho
Department of Computer Science,
Federal University,
Oye-Ekiti, Nigeria

Abstract—Medical dataset is a vital ingredient used in predicting patient's health condition. In other to have the best prediction, there calls for a technique with high degree of accuracy. In this paper, we modeled data from diabetes patients and used it to predict the diabetes probability of any patient. The performances of both Artificial Neural Network and Decision Tree Algorithms on medical data were measured. The performance measures were based on time to model, kappa statistics, mean absolute error, mean-squared error and relative squared error. The classifier with the best performance measure was selected and used for the prediction. The results showed that Decision Tree Algorithms performed better than that of the Artificial Neural Network.

Keywords— Medical dataset, Prediction, Performance Measures, Artificial Neural Network and Decision Tree Algorithms.

1. Introduction

With the computerization in hospitals, a huge amount of data is collected. Although human decision-making is often optimal, it is poor when there are huge amounts of data to be classified. Medical data mining has great potential for exploring hidden patterns in the data sets of medical domain. These patterns can be used for clinical diagnosis [1].

Data Mining is a technology used to describe knowledge discovery and to search for significant relationships such as patterns, association, and changes among variables in databases. The discovery of those relationships can be examined by using statistical, mathematical, artificial intelligence and machine learning techniques to enable users to extract and identify greater information and subsequent knowledge than simple query and analysis approaches [3].

Neural network techniques have the potential to handle complex, nonlinear problems in a better way when compared to traditional techniques. However systems developed using neural network model suffer from certain drawbacks like local minima, model over fitting etc [2]. This work aims at a comparative statement on the performance of Artificial Neural Network Algorithms which are MLP and RBF to that of Decision Tree Algorithms which are RegTree and LADTree.

2. Brief Description of Both ANN and Decision Tree Algorithms

a. Multilayer Perceptrons (MLPs)

Multilayer perceptrons (MLPs) are feedforward neural networks trained with the standard backpropagation algorithm. They are supervised networks so they require a desired response to be trained. They learn how to transform input data into a desired response, so they are widely used for pattern classification. With one or two hidden layers, they can approximate virtually any input-output map. They have been shown to approximate the performance of optimal statistical classifiers in difficult problems. Most neural network applications involve MLPs.

Radial Basic Function (RBF) Model

Radial Basic Function is a feedforward network is the Radial-Basis Function (RBF) which has important universal approximation properties. A radial basis function network is an artificial neural network that uses radial basis functions as activation functions. It is a linear combination of radial basis functions. They are used in function approximation, time series prediction, and control [4].

REPTree Algorithm. It is a fast decision tree learner. It builds a decision/regression tree using information gain/variance and prunes it using reduced-error pruning (with back-fitting). The algorithm only sorts values for numeric attributes once. Missing values are dealt with by splitting the corresponding instances into pieces [5].

b. LADTree Algorithm

LADTree is a class for generating a multiclass alternating decision tree using loitboost strategy. LADTree produces a multi-class ADTree. It has the capability to have more than two class inputs. It performs additive logistic regression using the Logitoot Strategy.

3. Materials and Methods

3.1 Data Collection

This data used for this research was collected from Lagos State University Teaching Hospital (LASUTH) with Two hundred (200) datasets was collected. The obtained record has nine variables include the age of patient, no of exercise done per week by patient, plasma glucose level (mgdl), sink fold thickness (mm), body mass index (kg/m2), Diastolic blood pressure (MgHg), smoking status, Diabetes pedigree type and diabetes probability.

Table 1: Attribute of Medical Dataset

S/N	Variable Name	Description
1	Age	Age of patient
2	E/W	No of exercise done by patient
3	PGL	Plasma glucose level
4	SFT	Skin fold thickness
5	BMI	Body mass index
6	DBP	Diastolic blood pressure
7	SS	Smoking status
8	DPT	Diabetes pedigree type
9	DP	Diabetes probability

3.2 Data Preprocessing

An important step in the data mining process is data preprocessing. One of the challenges that face the knowledge discovery process in medical database is poor data quality. For this reason we try to prepare our data carefully to obtain accurate and correct results. First we choose the most related attributes to our mining task.

3.3 Data Mining Stages

The data mining stage was divided into three phases. At each phase all the algorithms were used to analyze the health datasets. The testing method adopted for this research was parentage split that train on a percentage of the dataset, cross validate on it and test on the remaining percentage. Sixty six percent (66%) of the health dataset which were randomly selected was used to train the dataset using all the classifiers. The validation was carried out using ten folds of the training sets. The models were now applied to unseen or new dataset which was made up of thirty four percent (34%) of randomly selected records of the datasets. Thereafter interesting patterns representing knowledge were identified.

4. Pattern Evaluation

This is the stage where strictly interesting patterns representing knowledge are identified based on given metrics.

4.1 Evaluation Metrics

In selecting the appropriate algorithms and parameters that best model the diabetes forecasting variable, the following performance metrics were used:

- **Time:** This is referred to as the time required to complete training or modeling of a dataset. It is represented in seconds
- **Kappa Statistic:** A measure of the degree of nonrandom agreement between observers or measurements of the same categorical variable.
- **Mean Absolute Error:** Mean absolute error is the average of the difference between predicted and the actual value in all test cases; it is the average prediction error.
- **Mean Squared Error:** Mean-squared error is one of the most commonly used measures of success for numeric prediction. This value is computed by taking the average of the squared differences between each computed value and its corresponding correct value. The mean-squared error is simply the square root of the mean-

squared-error. The mean-squared error gives the error value the same dimensionality as the actual and predicted values.

- **Root relative squared error:** Relative squared error is the total squared error made relative to what the error would have been if the prediction had been the average of the absolute value. As with the root mean-squared error, the square root of the relative squared error is taken to give it the same dimensions as the predicted value.
- **Relative Absolute Error:** Relative Absolute Error is the total absolute error made relative to what the error would have been if the prediction simply had been the average of the actual values.

5. Experimental Design

The Artificial Neural Networks and Decision Tree algorithms were used to analyse the health data. The ANN algorithms used were Multilayer Perceptron (MLP) and Radial Basis Function (RBF), and the Decision Tree Algorithms used are RegTree and LadTree. The ANN models were trained with 500 epochs to minimize the root mean square and mean absolute error. Different numbers of hidden neurons were experimented with and the models with highest classification accuracy for the correctly classified instances were recorded. For the Decision Tree models, each class was trained with entropy of fit measure, the prior class probabilities parameter was set to equal, the stopping option for pruning was misclassification error, the minimum n per node was set to 5, the fraction of objects was 0.05, surrogates was 5, 10 fold cross-validation was used, and generated comprehensive results.

6. Results and Discussion

RepTree algorithm was selected for the prediction because out of the four classifiers used to train the data, it had the best performance measures.

Run Information of the RepTree

```
=== Run information ===

Scheme:weka.classifiers.trees.REPTree
Relation:     ola data
Instances:    200
Attributes:   10
              GENDER
              AGE
              E/W
              PGI(mg/dl)
              SFT(mm)
              BMIkg/m2
              DBP(mmHg)
              SS
              DPT
              DP
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===


REPTree
============

DBP(mmHg) < 80.33
|   PGI(mg/dl) < 120.25 : low (48/5) [24/1]
|   PGI(mg/dl) >= 120.25 : medium (3/1) [2/0]
DBP(mmHg) >= 80.33
|   PGI(mg/dl) < 124
|   |   NOP < 70 : medium (33/9) [9/3]
|   |   NOP >= 70
|   |   |   DBP(mmHg) < 93.85 : medium (22/1) [17/0]
|   |   |   DBP(mmHg) >= 93.85
|   |   |   |   SS < 0.5 : medium (3/0) [3/1]
|   |   |   |   SS >= 0.5 : high (4/0) [2/0]
|   PGI(mg/dl) >= 124 : high (20/3) [10/1]

Size of the tree : 13
```


Time taken to build model: 0.05 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	165	82.5 %
Incorrectly Classified Instances	35	17.5 %
Kappa statistic	0.722	
Mean absolute error	0.1822	
Root mean squared error	0.3174	
Relative absolute error	43.0115 %	
Root relative squared error	68.9978 %	
Total Number of Instances	200	

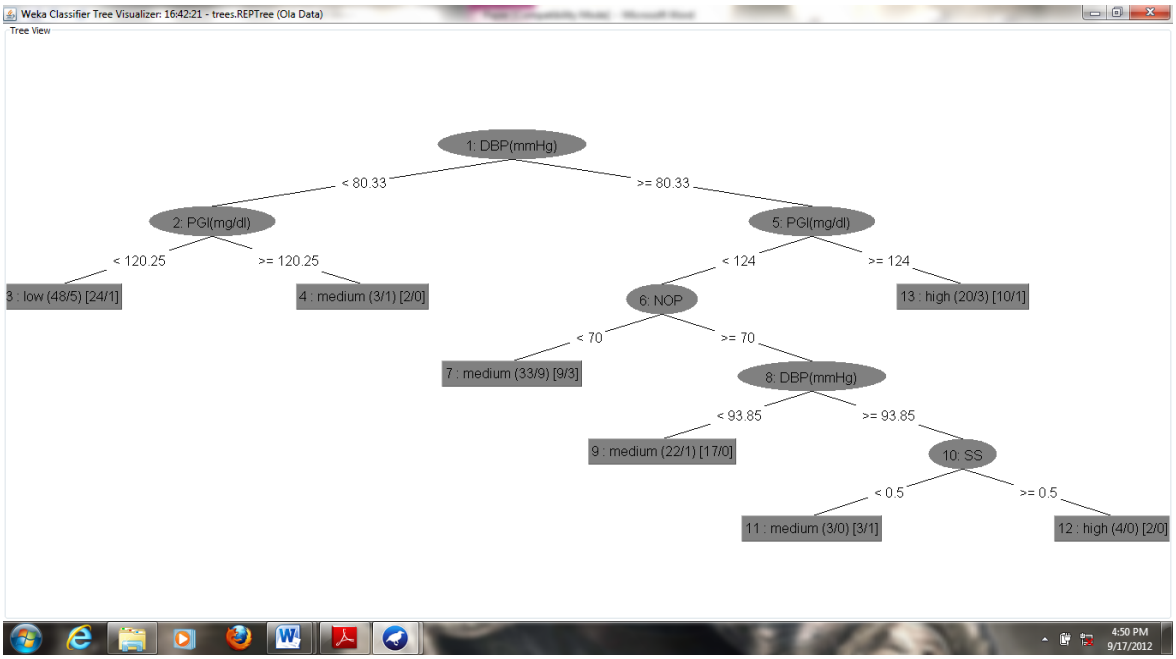

Figure 1: Visualization Tree for The prediction

Table 2: Performance of Both Decision Tree and Artificial Neural Network Algorithms

Performance Metrics				
	Artificial Neural Network		**Decision Tree**	
	MLP	**RBF**	**RepTree**	**LADTree**
Time	2.21	0.33	0.05	0.16
Kappa Statistics	0.5726	0.603	0.722	0.7157
MAE	1.883	0.2219	0.1822	0.1606
RMSE	0.3913	0.3625	0.3174	0.3206
RAE (%)	44.4585	52.3938	43.0115	37.9196
RRSE (%)	85.0425	78.7989	68.9978	69.6893

From the table 2 above, two types of Algorithms were used for both Artificial Neural Networks and Decision Tree Model. Multilayer Perceptron (MLP) and Radial Basis Function (RBF) were the two algorithms used for ANN while RegTree and LADTree algorithms were the Decision Trees model used. For the diabetes probability prediction, MLP algorithm used 2.21 secs to model, with kappa statistic of 0.5726, mean absolute error of 1.883 and root mean square error of 0.3913 while RBF algorithm was modeled within 0.33 secs, with kappa statistic of 0.603, mean absolute error of 1.2219 and root mean square

error of 0.3625. In the case of Decision Tree Performance analysis, RepTree algorithm used 0.04 sec to modeled, with kappa statistic of 0.722, mean absolute error of 0.1822 and root mean square error of 0.3174. While LADTree algorithm was modeled within 0.16 secs, with kappa statistic of 0.7157 mean absolute error of 0.1606 and root mean square error of 0.3206. Finally, from the result analysis by comparing the techniques, Decision Tree performs better than the Neural Networks based on the error report, number of correctly classified instances and accuracy rate generated.

7. CONCLUSION

This paper presents some examples of both Decision Tree and Artificial Neural Networks building process, of most common data mining techniques. The work revealed that, Decision Tree techniques outperformed Artificial Neural Networks with a lower error metrics and higher correlation coefficient. We have tried to highlight the way the stored data about diabetes record could be used in the predicting of the new patient. We can predict, with certain accuracy, the diabetes probability of any patient if we have data regarding some important aspects of the patient health record. For that we can use a decision tree built with Reptree or LADTree algorithm implemented in specialized software in data mining - Weka.

Reference

[1] Karegowda, A. G.; A.S. Manjunath; M.A Jayaram "Application of Genetic Algorith Optimized Neural Network Connection Weighs For Medical Diagnosis of Pima Indians Diabetes" International Journal on Soft Computing (IJSC), Vol. 2, No.2, May 2011

[2] Radhika, Y and Shashi, M., 2009 "Atmoshere Temperature Prediction using Support Vector Machines", International Journal of Computer Theory and Engineering, Vol 1, No 1, pp 55 – 57
Saurabh Pal, 2012, "Mining Educational Data to Reduce Dropout Rates of Engineering Student", International Journal of Information Engineering and Electronic Business. Published Online (http://www.mecs-press.org)

[3] Wikipedia, 2010, "Radial basis function network" *From Wikipedia - the free encyclopedia*, retrieved from http://en.wikipedia.org/wiki/ Radial_basis_function_network.htm in July 2010.

[4] Yasogha, P; M. Kannan "Analysis of a Population of Diabetics Patients Databases in Weka Tool", International Journal of Science & Engineering Research, Vol. 2, Issue 5, May 2011.

Bibliography

Gelfan S.G., Ravishanker C.S., Delp E.J., 1991, "An iterative Growing and Pruning Algorithm for Classification Tree Design, PAMI(13)", No. 2, February 1991, pp. 163-174.

Rahman S., Chowdhury B., 1988, "Simulation of Photovoltaic power systems and their performance prediction". IEEE Transactions on Energy Conversion 3,440-446 (1988)

Sherrod P. H., 2003, DTREG Predictive Modelling Software, Retrieved Feb. 7, 2011 from http://www.dtreg.com.

