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

Methodology of Adaptation of Data Mining Methods (ANN and Decision Trees) for Nervous System and Arrhythmia Diseases

Tawseef Ayoub Shaikh
Department of Computer Science and Engineering
GNDU Amritsar, Punjab, India

Abstract*——Data mining is a problem solving technique, which analyzes the data already stored in the data base. It is a process of discovering, classifying and finding patterns in data. The classification of data helps to make a decision in different types of problems. Machine Learning, Statistical and Neural Network algorithms [1] are applied for efficient data mining. The problem is to find algorithms suitable to apply in order to discover relationships between data attributes and make predictions that could be useful for decision support. While as lot of research has been or is going on common diseases whereas at the same time the diseases related to Nervous System and Mental Health have been ignored which are very common in the present Technological world. In this paper steps of data mining algorithms on segment dataset with 2100 instances and 20 attributes ,Primary Tumor with 340 instances and 18 attributes and Arrhythmia with 452 instances and 280 attributes have been applied . All the Datasets were taken from UCI Repository [2].

***Keywords*——Medical dataset, Performance Measures, Artificial Neural Network, Decision Tree**

I. INTRODUCTION

. Adjacent segment disease has been considered a late complication of spinal fusion. It's described as any degeneration that develops at mobile segments above or below a fused spinal segment [3]. The radiographic findings [4] of the present study using powerful machine learning models suggest that there adaptation using Data mining algorithms will help in early detection of this vital disease. Brain tumors are also not rare. Thousands of people are diagnosed every year with tumors of the brain and the rest of the nervous system. The diagnosis and treatment of the brain tumor depends on the type of tumor, tumor grade and where it started [5].

An arrhythmia is an abnormal heart rhythm. It may feel like fluttering or a brief pause. It may be so brief that it doesn't change your overall heart rate. Or it can cause the heart rate to be too slow or too fast. Some arrhythmias don't cause any symptoms. Others can make you feel lightheaded or dizzy. In the USA, it is estimated that there are nearly one million CHD patients, 15–20% with disease of severity to warrant surgical intervention. Arrhythmias complicate the care of many adults with CHD [6].This article will review the evaluation and management of these more common arrhythmia problems in adults with CHD using machine learning techniques.

II. DATA MINING PREDICTION MODELS

In this research WEKA (The Waikato Environment for Knowledge Analysis) for running several algorithms has been chosen. Three two types of classification models: Artificial neural networks (ANN) and decision trees were used . These models were selected for inclusions in this study due to their popularity in the recently published literature as well as their better than average performance in many preliminary comparative studies. What follows is a short description of these classification model types and their specific implementations for this research.

A. Artificial neural networks (Functions in Weka)

Artificial neural networks (ANNs) are commonly known as biologically inspired, highly sophisticated analytical techniques, capable of modeling extremely complex non-linear functions. Formally defined, ANNs are analytic techniques modeled after the processes of learning in the cognitive system and the neurological functions of the brain and capable of predicting new observations (on specific variables) from other observations (on the same or other variables) after executing a process of so-called learning from existing data [7].

i) Multi-layer perceptron (MLP)

Multi-layer perceptron (MLP) with back-propagation (a supervised learning algorithm) is known to be a powerful function approximator for prediction and classification problems. It is arguably the most commonly used and well-studied ANN architecture [8]. The MLP is essentially the collection of nonlinear neurons (a.k.a. perceptrons) organized and connected to each other in a feed forward multi-layer structure.

ii) Radial Basis Function (RBF)

The construction of the RBFNN involves an input layer, a hidden layer and an output layer with feed forward architecture. The input layer of this network is a set of n units, which accept the elements of an n-dimensional input feature vector. The input units are fully connected to the hidden layer with r hidden units. Connections between the input and the hidden layer have unit weights and, as a result, do not have to be trained. In this structure the hidden units are

referred to as the RBF units. The goal of the RBF units is to cluster the data and reduce its dimensionality with a nonlinear transformation and to map the input data to a new space.

The RBF units are also fully connected to the output layer. The output layer, which contains s units, implements a linear combination on this new space. radial basis function neural networks (RBFNNs) have been found to be very attractive for many engineering problems. An important property of the RBFNNs is that they form a unifying link among many different research fields such as function approximation, regularization, noisy interpolation and pattern recognition [9].

B. Decision Trees

Decision trees are powerful classification algorithms that are becoming increasingly more popular with the growth of data mining in the field of information systems. As the name implies, this technique recursively separates observations in branches to construct a tree for the purpose of improving the prediction accuracy.

i) J48

J48 is an implementation of C4.5 algorithm [10]. There are two methods in pruning support by J48, first one is known as sub tree replacement, it works by replacing nodes in decision tree with leaf. Basically by reducing the number of test with certain path. It works with the process of starting from leaves that overall formed tree and do a backward toward the root. The second type implemented in J48 is sub tree raising by moved nodes upwards toward the root of tree and also replacing other nodes on the same way [11].

ii) Random Forests

Random Forest developed by Leo Bremen is a group of un-pruned classification or regression trees made from the random selection of samples of the training data. Random features are selected in the induction process. Prediction is made by aggregating (majority vote for classification or averaging for regression) the predictions of the ensemble. Each tree is grown as described in [12] [13]:

- ✓ By Sampling *N* randomly, If the number of cases in the training set is *N* but with replacement, from the original data. This sample will be used as the training set for growing the tree.
- ✓ For *M* number of input variables, the variable *m* is selected such that $m<<M$ is specified at each node, *m* variables are selected at random out of the *M* and the best split on these *m* is used for splitting the node. During the forest growing, the value of *m* is held constant.
- ✓ Each tree is grown to the largest possible extent. No pruning is used. Random Forest generally exhibits a significant performance improvement as compared to single tree classifier such as C4.5. The generalization error rate that it yields compares favorably to Adaboost, however it is more robust to noise.

III. DATASETS

To review the performance of the three classifiers (MLP, RBF , J48, Random Forest) ,four datasets were used as shown in Table I.

i) 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 tried to prepare data carefully to obtain accurate and correct results. First we choose the most related attributes to the mining task.

ii) 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 is parentage split that train on a percentage of the dataset, cross validate on it and test on it 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.

The Datasets used are Segment Datasets, Primary Tumor, and Arrhythmia. Segment Datasets has come with 2100 instances and 20 attributes, The second dataset is Primary Tumor which consists of 340 instances and 18 attributes and the final dataset is of Arrhythmia with 452 instances and 280 as given below in table.

Table 1 Datasets and their types used

Datasets	Instances	Attributes
Segment Dataset	2100	20
Primary Tumor Dataset	340	18
Arrhythmia	452	280

IV. EXPERIMENTAL DSEIGN

The Artificial Neural Networks and Decision Tree algorithms were used to analyze the health data. The ANN algorithms used were Multilayer Perceptron , Radial Basis Function, the Decision Tree Algorithms used are J48,Randomforest). 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.

V. PERFORMANCE METRICS

In this paper, the performance measures which are used for comparison are: accuracy, sensitivity and specificity. A distinguished confusion matrix is obtained to calculate the three measures. Confusion matrix is a matrix representation of the classification results. the upper left cell denotes the number of samples classifies as true while they were true (i.e., true positives), and lower right cell denotes the number of samples classified as false while they were actually false (i.e., true false).

The other two cells (lower left cell and upper right cell) denote the number of samples misclassified. Specifically, the lower left cell denoting the number of samples classified as false while they actually were true (i.e., false negatives), and the upper right cell denoting the number of samples classified as true while they actually were false (i.e., false positives).Once the confusion matrixes were constructed, the accuracy, sensitivity and specificity are easily calculated as: sensitivity = TP/(TP + FN); specificity =TN/(TN + FP). Accuracy = (TP + TN)/(TP + FP + TN + FN); where TP, TN, FP and FN denotes true positives, true negatives, false positives and false negative. More Matrices include used are as:

- ✓ 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.
- ✓ 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.
- ✓ Precision: Percentage of retrieved documents that are relevant: precision=TP/ (TP+FP).
- ✓ ROC Curves: ROC curves are similar to lift charts. It stands for "Receive Operating Characteristics ".These are Used in signal detection to show tradeoff between hit rate and false alarm rate over noisy channel. It also Differences to lift chart: y axis shows percentage of true positives in sample rather than absolute number¨ *x* axis shows percentage of false positives in sample rather than sample size.
- ✓ Recall: Percentage of relevant documents that are retrieved: Recall=TP/(TP+FN).
- ✓ Fmeasure=(2 × recall × precision)/(recall+precision).

Table II Performance of Artificial Neural Network and Decision Tree on Segment Data

Performance Metrics	Artificial Neural Network (Functions)		Decision Trees	
Algorithms → / Metrics ↓	MLP	RBF	J48	Random Forest
Time	17.77	5.4	0.49	0.51
Kappa Statistics	.9556	.9323	.9556	.9672
MAE	.0149	.1006	.0126	0.018
RMSE	.0959	.1095	.1019	.0819
RAE%	6.0945	34.1314	5.1267	7.3444
RRSE%	27.3985	62.1143	96.7143	99.9048
Accuracy=TP+TN/ TP+FP+TN+FN	96.1905	89.93	96.1905	97.1905
Sensitivity =TP/TP+FN	97.81	77.25	98.54%	98.60%
Specificity=TN/TN+FP	95.27	88.36	96.23%	99.29
Precision	.962	.1018	.962	.972

Recall	.962	.8976	.962	.972
FMeasure=2*Precision*Recall/Precision+Recall	.962	.1829	.962	.972

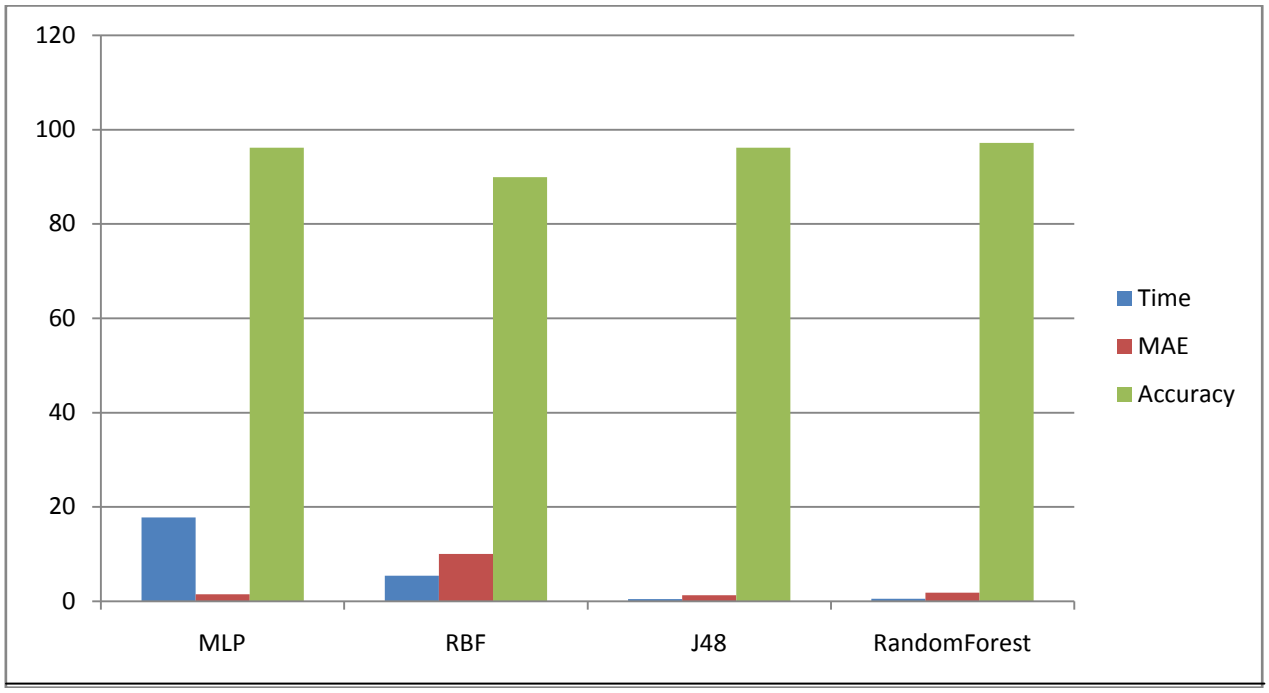

Figure I Performance Comparison of Time, MAE & Accuracy of ANN and Decision Trees on Segment Dataset

Table III Performance of Artificial Neural Network and Decision Tree on Primary Tumor

Performance Metrics	Artificial Neural Network (Functions)		Decision Trees	
Algorithms → / Metrics ↓	MLP	RBF	J48	Random Forest
Time	431.43	262.83	0.83	0.93
Kappa Statistics	.4793	.3177	.4601	.357
MAE	.0434	.3077	.0487	0.0642
RMSE	.185	.2021	.2002	.1794
RAE%	50.6691	62.1438	56.8448	74.9801
RRSE%	89.987	104.3613	97.3969	87.1593
Accuracy=TP+TN/ TP+FP+TN+FN	67.6991	57.4000	64.3805	64.6018
Sensitivity =TP/TP+FN	66.000	48.000	77.800	77.828
Specificity=TN/TN+FP	88.99	88.400	81.600	95.112
Precision	.642	.1286	.614	.600
Recall	.677	.6088	.644	.646
FMeasure=2*Precision*Recall/Precision+Recall	.650	.1008	.628	.579

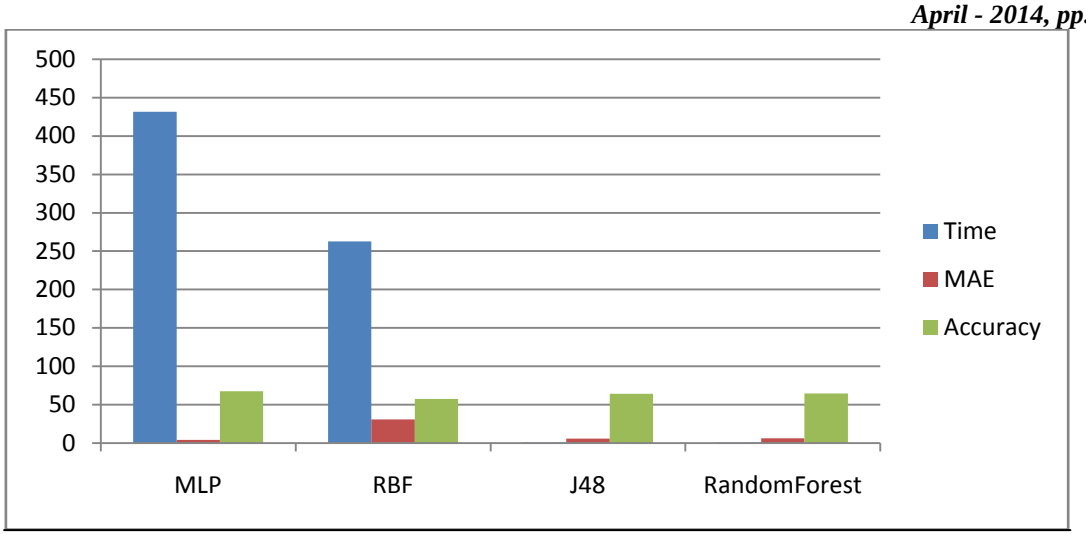

Figure II Performance Comparison of Time, MAE & Accuracy of ANN and Decision Trees on Primary Tumor Dataset

Table IV Performance of Artificial Neural Network and Decision Trees on Arrhythmia Dataset

Performance Metrics	Artificial Neural Network (Functions)		Decision Trees	
Algorithms → / Metrics ↓	MLP	RBF	J48	Random Forest
Time	19.79	5.27	0.07	0.1
Kappa Statistics	.3426	.3157	.3353	.3751
MAE	.0569	.0976	.0626	0.0614
RMSE	.2041	.1927	.1931	.1864
RAE%	69.921	71.81	76.5509	75.4801
RRSE%	101.4136	94.4156	95.9566	92.6507
Accuracy=TP+TN/ TP+FP+TN+FN	42.0588	40.0589	42.3529	45.00
Sensitivity =TP/TP+FN	50.00	32.63	32.500	50.00
Specificity=TN/TN+FP	42.00	43.77	40.00	60.00
Precision	.388	.0992	.338	.406
Recall	.421	.0900	.424	.450
FMeasure=2*Precision*Recall/Precision+Recall	.402	.9438	.371	.425

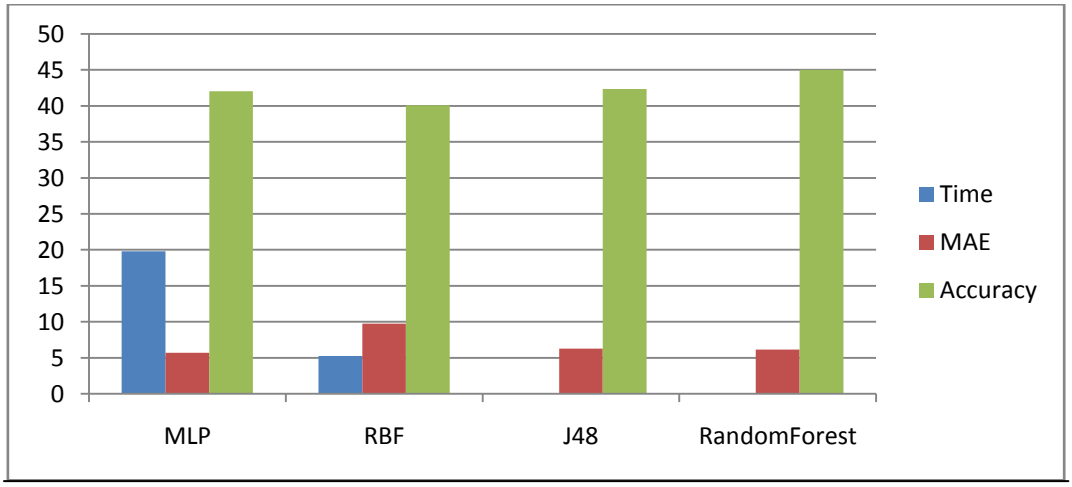

Figure III Performance Comparison of Time, MAE & Accuracy of ANN and Decision Trees on Arrhythmia Dataset

VI. CONCLUSION AND FUTURE WORK

On evaluating the different data mining algorithms on Nervous System and Arrhythmia Datasets of Segment, Primary Tumor and Arrhythmia datasets we came to the conclusion that all the two classifiers (ANN, Decision Trees) both have highest accuracy on the Numeric Dataset and their accuracy gets down below 45% when the dataset is Nominal as in case of Primary Tumor

Overall Random Forests from Decision trees and MLP from ANN performed best as given in above figures (I – III) but MLP overall takes more time than the Random Forests .

The classification accuracy of Nominal Datasets like Primary Tumor can increase by using other Data Mining Models like NavieBayes , SVM(Support Vector Machine). In the future accuracy of the above Mining Classification Models can further improve after applying genetic algorithm to reduce the actual data size to get the optimal subset of attribute. Also the hidden layer plays an important role for detecting the relevant features. The use of certain Supervised and Nonsupervised Filters, Sampling and Discretization also plays an additive effect on accuracy and reduces error rates to a certain level.

References

[1] I.H. Witten, E. Frank**.** Data Mining: Practical machine learning tools and techniques, 2nd Edition. Morgan Kaufmann, San Francisco, 2005.
[2] A. Asuncion, D.J. Newman. UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science,2007,http://www.ics.uci.edu/~mlearn/MLRepository.html.
[3] Adjacent Segment Degeneration in the Lumbar Spine Gary Ghiselli, Jeffrey C. Wang, Nitin N. Bhatia, Wellington K. Hsu and Edgar G. Dawson J. Bone Joint Surg. Am. 86:1497-1503, 2004.
[4] Whitecloud TS 3rd, Davis JM, Olive PM. Operative treatment of the degenerated segment adjacent to a lumbar fusion. Spine. 1994;19:531-6.
[5] 1995-2011,The Patient Education Institute ,Inc . www.X-Plain.com
[6] Arrhythmia's in adults with congenital heart disease John K Triedman Heart 2002; 87:383 389
[7] Haykin S. Neural networks: a comprehensive foundation. New Jersey: Prentice Hall; 1998.
[8] Hornik K, Stinchcombe M, White H. Universal approximation of an unknown mapping and its derivatives using multilayer feed forward network. Neural Network.
[9] W. Zhou, Veri6cation of the non parametric characteristics of back propagation neural networks for image classi6cation, IEEE Trans. Geo. Remote Sensing 37 (2) (1999) 771–779.
[10] Quinlan J. C4.5: programs for machine learning. San Mateo,CA: Morgan Kaufmann; 1993
[11] I. H. Witten, and E. Frank, "Data Mining Practical Machine Learning Tools and Techniques," Second Edition, Morgan Kaufmann Publisher, United States of America, 2005.
[12] Y. Zhao and Y. Zhang, "Comparison of Decision Tree Methods for Finding Active Objects," National Astronomical Observatories, Advances of Space Research, 2007..
[13] http://www.stat.berkeley.edu/~breiman/RandomForests/cc_h ome.htm#prox Symposium, volume 1, July, 2005.

