Predictive Data Mining for Medical Diagnosis: An Overview of Segment and Arrhythmia Diseases

Tawseef Ayoub Shaikh
Department of Computer Science and Engineering , Guru Nanak Dev University (GNDU) Amritsar, Punjab –India

Abstract

The successful application of data mining in highly visible fields like e-business, marketing and retail has led to its application in other industries and sectors. Among these sectors just discovering is healthcare. The healthcare environment is still information rich but knowledge poor. There is a wealth of data available within the healthcare systems. However, there is a lack of effective analysis tools to discover hidden relationships and trends in data. This research paper intends to provide a survey of current techniques of knowledge discovery in databases using data mining techniques that are in use in today's medical research particularly in Segment [1] and Arrhythmia [2] diseases. Number of experiment has been conducted to compare the performance of predictive data mining technique on the same datasets and the outcome reveals that Decision Tree [3] outperforms in case of Segment Dataset and Bayesian classification [4] over performed in case of Arrhythmia case.. We found that in case of Segment Dataset Random Forest got the highest accuracy of 97.1905% followed by J48 which got an accuracy of 96.1905% followed by BayesNet with an accuracy of 90.7143% and the least accuracy was get by NavieBayes with an accuracy 79.8095% among all other classification algorithms. Similarly, in case of Arrhythmia Datasets it is the BayesNet which got the highest classification accuracy of 69.9115% , followed by NaiveBayes with an accuracy of 62.3894% , followed by the decision Tree Algorithm Random forest with classification accuracy of 45.00% and the J48 got the lowest of all the accuracy among the other algorithms in this case with a minimum accuracy of 42.3529%.

Keywords NavieBayes, BayesNet, J48 , Random Forest , Segment Dataset , Arrhythmia Dataset

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. Several questions have arisen in the late years regarding this syndrome, we try to define

it, describe it and determine if it‘s caused by strain forces due to spinal arthrodesis or whether it‘s the natural history of the degenerative spinal process. Possibilities on conservative and surgical management are discussed. Spinal degeneration is a pre – determined genetic process and therefore the involvement of unfused levels is an expected result Adjacent segment disease in the lumbar and lumbo [5] sacral spine has been examined extensively in previous biomechanical and clinical studies, we are aware of no study that has specifically addressed the rate of degeneration of adjacent segments.

In addition, previous studies have not demonstrated an association between radiographic evidence of degeneration of adjacent segments [5] and the long-term clinical outcome of posterior lumbar fusion. The radiographic findings of the present study did show a significant progression of the arthritic grade of the adjacent segment. However, the clinical importance of this radiographic progression is undetermined. It is expected that arthritic degeneration of a motion segment will progress with time, regardless of whether or not the motion segment is adjacent to a fused segment. The radiographic findings of the present study backed by powerful machine learning models suggest that there adaptation using Data mining algorithms will help in early detection of this vital disease.

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. As surgical mortality has fallen, the number of adults living with major congenital heart defects has increased. Arrhythmias complicate the care of many adults with CHD [6] . Their prevalence and the difficulty of treatment have made arrhythmia a major focus of interest for physicians working in this area. The presence of longstanding CHD in an arrhythmia patient significantly alters the nature and potential severity of the arrhythmia complaint and the safety and feasibility of various treatments.

In addition to analysis of the targeted arrhythmia complaint, the physician must have complete and specific knowledge of the patient‘s cardiovascular anatomy and the consequences of that anatomy and subsequent surgical modifications on cardiovascular function. The arrhythmogenic substrate in adults with CHD is complex. All arrhythmias prevalent in the normal population may also occur in CHD, and some specific associations are observed—for example, Wolff-Parkinson-White syndrome and Ebstein‘s anomaly.However,more common are acquired arrhythmias that are rarely seen in normal young adult hearts, and that are associated with longstanding hypertrophy and fibrosis caused by cyanosis, chronic hemodynamic overload, and superimposed surgical scarring. These arrhythmias include re-entrant atrial and ventricular tachycardia’s, heart block, and sinus node dysfunction. This article will review the evaluation and management of these
More common arrhythmia problems in adults with CHD.

II. METHODOLOGY

Nowadays there are many available tools in data mining, which allow execution of several task in data mining such as data preprocessing, classification, regression, clustering, association rules, features selection and visualization [7]. All the above mention tasks are closed under different algorithms and are available an application or a tool. In this research WEKA (The

Waikato Environment for Knowledge Analysis) has been chosen for running several algorithms. Two different types of classification models: Bayesian Network and decision trees have been choosed as Classifiers. 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 my preliminary comparative studies. What follows is a short description of these classification model types and their specific implementations for this research.

A. 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. Popular decision tree algorithms include Quinlan‘s ID3 C4.5, C5 [3], and Breiman et al. ‘s CART . As the name implies, this technique recursively separates observations in branches to construct a tree for the purpose of improving the prediction accuracy. In doing so, they use mathematical algorithms (e.g., information gain, Gini index, and Chi-squared test) to identify a variable and corresponding threshold for the variable that splits the input observation into two or more sub groups. This step is repeated at each leaf node until the complete tree is constructed. The objective of the splitting algorithm is to find a variable-threshold pair that maximizes the homogeneity (order) of the resulting two or more subgroups of samples. The most commonly used mathematical algorithm for splitting includes Entropy based information gain (used in ID3, C4.5, J48), Gini index (used in CART), and Chi-squared test (used in CHAID) [8].

i) J48

J48 is an implementation of C4.5 algorithm. 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 [8].

ii) Random Forests

Random Forest developed by Leo Breiman [9] 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 [10]:

- 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 described IB4 and IB5, which handle irrelevant and novel data.

B. Bayesian Networks

ii) Naive Bayes

The Naïve Bayes [4] classifier provides a simple approach ,with clear semantics, representing and learning probabilistic knowledge. It is termed naïve because it relies on two important simplifying assumes that the predictive attributes are conditionally independent given the class, and it assumes that no hidden or latent attributes influence the prediction process.

i) Bayes Net

Bayes Net [4] learns Bayesian Networks under the pre assumption: nominal attributes(numeric one are pre-discredited) and no missing values (any such values are replaced globally).There are two different parts for estimating the conditional probability tables of the network. In this study we run BayesNet with the simple estimator and K2 search algorithm without using ADtree.K2 algorithm is a greedy search algorithm that works as follows. Suppose we know that total ordering of the nodes. Initially each node has no parents. the algorithm then incrementally adds the parent whose addition increases most of the score of the resulting structure. When no addition of a single parent can increase the score, it then stops by adding parents to the node. Since it is already known the ordering of the nodes beforehand, the search space under this constraint is much smaller than the entire space. And there is no need to check for cycles, since the total ordering guarantees that there is no cycle in the structure.

.

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 [11].

The Datasets used are Segment Datasets and Arrhythmia which were taken from open source UCI repository [12]. Segment Datasets has come with 2100 instances and 20 attributes and the 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
Arrhythmia	452	280

IV. 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).

V. RESULT & DISCUSSION

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

Every model was evaluated based on the above measures discussed above . The results were achieved using average value of 10 fold cross-validation for each algorithm. We found that in case of Segment Dataset Random Forest got the highest accuracy of 97.1905% followed by J48 which got an accuracy of 96.1905% followed by BayesNet with an accuracy of 90.7143% and the least accuracy was get by NavieBayes with an accuracy 79.8095% among all other classification algorithms.

Similarly, in case of Arrhythmia Datasets it is the BayesNet which got the highest classification accuracy of 69.9115%, followed by NaiveBayes with an accuracy of 62.3894%, followed by the decision Tree Algorithm Random forest with classification accuracy of 45.00% and the J48 got the lowest of all the accuracy among the other algorithms in this case with a minimum accuracy of 42.3529%. The detailed prediction results of the validation datasets are presented in form of confusion matrixes.

Table II Performance of Decision Tree and Bayesian Network on Segment Data

Performance Matrices	**Decision Trees(J48)**		**Bayesian Network**	
	J48	**Random Forest**	**NavieBayes**	**BayesNet**
Time	.39	.51	.09	.38
Kappa Statistics	.9556	.9672	.7644	.8917
MAE	.0126	.018	.058	.0279
RMSE	.1019	.0189	.2299	.1509
RAE(%)	5.1267%	7.3444%	23.6644%	11.3989%
RRSE(%)	29.1067%	99.9048%	65.6849%	43.1122%
Accuracy=(TP+TN)/ (TP+FP+TN+FN)	96.1905%	97.1905%	79.8095%	90.7143%
Sensitivity=TP/TP+FN	98.54%	98.60%	92.44%	93.21%
Specificity=TN/TN+FP	96.23%	99.29%	94.59%	92.33%
Precision	.962	.972	.819	.907
Recall	.962	.972	.798	.907
Fmeasure=2*Precision*Recall/Precision+Recall	.962	.972	.798	.907

Figure I Performance Comparison of Time, MAE & Accuracy of Decision Trees & Bayesian Network on Segment Dataset

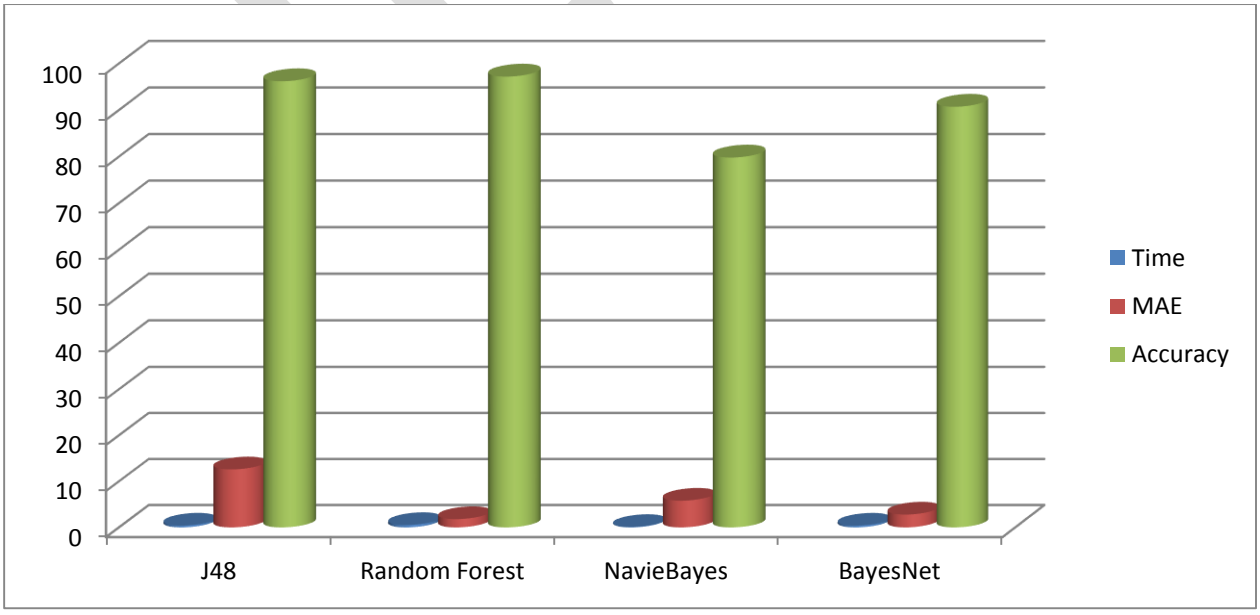
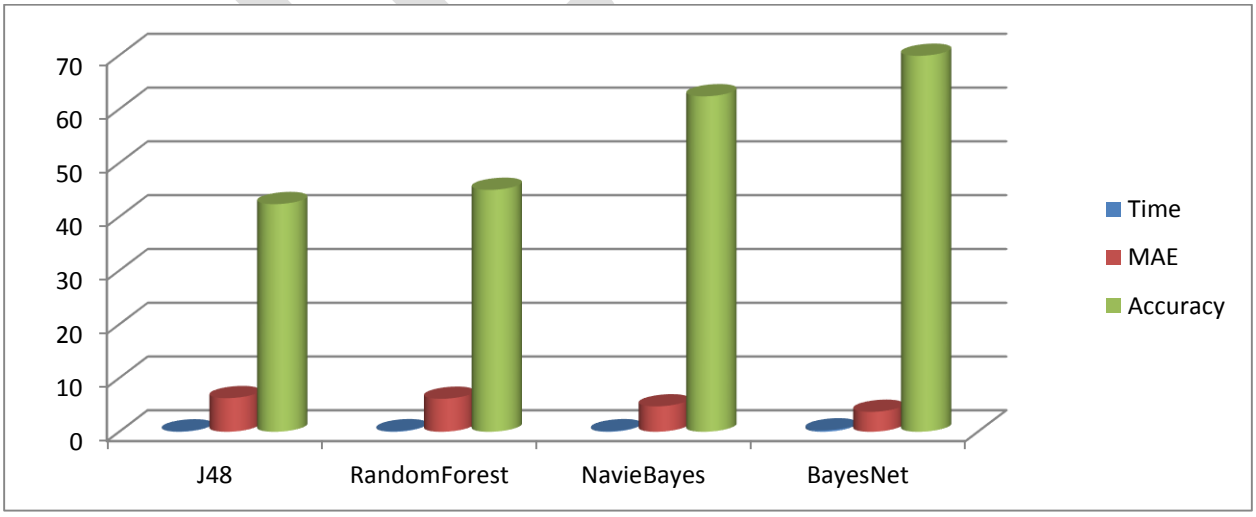


Table II Performance of Decision Tree and Bayesian Network on Arrhythmia Data

Performance Matrices	**Decision Trees(J48)**		**Bayesian Network**	
	J48	**Random Forest**	**NavieBayes**	**BayesNet**
Time	.07	.1	.1	.26
Kappa Statistics	.3353	.3751	.442	.538
MAE	.0626	.0614	.0473	.0371
RMSE	.1931	.1864	.2146	.1797
RAE(%)	76.5509%	75.4801%	55.244%	43.3401%
RRSE(%)	95.9566%	92.6507%	104.387%	87.4113%
Accuracy=(TP+TN)/ (TP+FP+TN+FN)	42.3529%	45.00%	62.3894%	69.9115%
Sensitivity=TP/TP+FN	32.500%	50.00%	76.55%	78.31%
Specificity=TN/TN+FP	40.00%	60.00%	79.65%	75.22%
Precision	.338	.406	.627	.664
Recall	.424	.450	.624	.674
Fmeasure=2*Precision* Recall/Precision+Recall	.371	.425	.623	.535

Figure II Performance Comparison of Time, MAE & Accuracy of Decision Trees & Bayesian Network on Arrhythmia Data



VI. Conclusion and Future Work

On evaluating the different data mining algorithms on Segment and Arrhythmia Datasets, the conclusion was drawn that all the BayesNet overall performs best on both the datasets. The Classification accuracy of Decision trees was highest in case of Segment dataset, whereas the classification accuracy of Bayesian Network overtook the Decision Trees in case of Arrhythmia dataset. Moreover, it is concluded that the Bayesian Networks perform best even on the dataset containing large number of instances (as in case of Arrhythmia Datasets).

Overall Random Forests from Decision trees and BayesNet from Bayesian Networks performed best as given above clearly mentioned in figures (I, II). 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 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] Whitecloud TS 3rd, Davis JM, Olive PM. Operative treatment of the degenerated segment adjacent to a lumbar fusion. Spine. 1994; 19:531-6.

[2] Arrhythmia's in adults with congenital heart disease John K Triedman Heart 2002;87:383 389

[3] J. Quinlan, "C4.5: Programs for Machine Learning," Morgan Kaufmann, San Mateo, 1993

[4] G.H.John and P.Langley, "Estimating Continuous Distributions in Bayesian Classifiers," Proceedings of the 11th Conference in University in Artificial Intelligance,San Francisco,1995,pp.338-345.

[5] 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.

[6] 1995-2011, The Patient Education Institute ,Inc . www.X-Plain.com

[7] I. H. Witten, and E. Frank, "Data Mining Practical Machine Learning Tools and Techniques," Second Edition, Morgan Kaufmann Publisher, United States of America, 2005.

[8] Y. Zhao and Y. Zhang, "Comparison of Decision Tree Methods for Finding Active Objects," National Astronomical Observatories, Advances of Space Research, 2007.[10] Lavrac N. Selected techniques for data mining in medicine. Artif Intell Med 1999;16:3 23.

[9] Breiman, L., Random Forests, Machine Learning 45(1), 5-32, 2001.

[10] http://www.stat.berkeley.edu/~breiman/RandomForests/cc_h ome.htm#prox Symposium, volume 1, July, 2005.
[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] 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

AUTHORS PROFILE:

NAME: TAWSEEF AYOUB SHAIK
S/O: MOHD AYOUB SHAIKH
STATE: JAMMU & KASHMIR (INDIA)
HAS GOT BTECH IN COMPUTER SCIENCE & ENGINEERING FROM ISLAMIC UNIVERSITY OF SCIENCE & TECHNOLOGY (IUST)
AWANTIPORA, PULWAMA KASHMIR-INDIA
THE AUTHOR IS PRESENTLY WORKING ON HIS THESIS IN THE FOURTH SEMESTER OF HIS MTECH IN SOFTWARE SYSTEMS IN DFPARTMENT OF COMPUTER SCIENCE, GURU NANAK DEV UNIVERSITY (GNDU) AMRITSAR PUNJAB-INDIA, 143005