Binary Categorization of Brain EEG data: A Case Study

1. S.Jayashree **2**.P.Keerthana **3.** S.Sanjana **4.** Shomona Gracia Jacob
1. B.E. CSE , SSNCE, Chennai. **2**. B.E. CSE , SSNCE, **3.**B.E. CSE , SSNCE,Chennai
4. Assistant Professor, SSNCE, Chennai.
Email ids: **1.** jayashree11jan@gmail.com **2**. keerthi.pkrn@gmail.com
3. sanjana.sahayaraj@gmail.com **4**. shomonagj@ssn.edu.in

ABSTRACT:

EEG Data plays an important role in exactly identifying the state of the human brain. The pattern of closure and opening of eyes is different as observed in a patient affected by coma and a normal person. While a normal person can voluntarily control the closing and opening of eyelids due to blood flow to that region, a coma patient cannot. Data mining is done to appreciate the change in the brain waves recorded during closing and opening of eyelids. Data mining involves extraction of important data which is previously unknown and hidden in large volumes of data. The study of most effective data mining algorithms can improve the treatment methodologies in the medical field. In this way, the attributes that play the most important role in binary categorization of EEG data can be identified. This enables the doctors to reduce spending their ample time trying to analyze massive data with many attributes. This paper reveals the performance of feature selection based classifiers on Brain EEG data. The findings of this research lead to the conclusion that the Correlation Based Feature Subset Evaluator is more optimal in detecting significant features to categorize brain data. Once the important attributes for the binary categorization of brain EEG data have been identified through data mining algorithms, more effective research can be done which leads to enhanced accuracy.

KEYWORDS: Binary Categorization, EEG data , Classification , Feature Selection.

1. INTRODUCTION

Data mining is being used as a computation process these days for discovering patterns in large data sets. This process which improves the performance involves methods at the intersection of artificial intelligence, machine learning statistics and database systems. Extracting previously unknown patterns involves cluster analysis, anomaly detection, association rule mining etc. The classification algorithms of WEKA aim at the predicted target being categorical. WEKA stands for Waikato Environment for Knowledge Learning. Data is being preprocessed and mined to obtain knowledge that might help humanity in many ways. The various data mining techniques that have been utilized in the past to discover novel patterns include association rules, classification, regression, summarization, clustering and visualization. Each attribute of the data set can be used to make a decision but sometimes it is important to extract the vital data using minimal resources available. This paper attempts to perform a comparative study of data mining techniques in discovering significant patterns in categorizing brain EEG data that can be effectively utilized to predict the state of a patient's brain based on the eyelid movements. The research work carried out thus far in this sphere of study is succinctly presented below.

2. LITERATURE SURVEY:

Surveying the work done so far by researchers , it is noted that a wide range of algorithms are being used to characterize brain data giving results of varying accuracy . CFS proves to be an effective method in [7] where a total of 1497 attributes were initially loaded as the training data

with 113 instances . No records were duplicated and there were no missing values. On ranking the attributes by the Gain Ratio criterion, a total of 134 attributes were assigned a gain ratio greater than zero. The CFS subset evaluator returned 39 features as the most optimal subset that was highly correlated to the target class but least correlated to each other. These features were then utilized for the Incremental feature Selection process. Few algorithms like Naïve Bayes, Bayes Net , Random forest, J48 are being used for accuracy when combined with Gain Ratio or Info Gain criterion for ranking . Deepika Kundra et al [1] classified the diseases using J48 algorithm of Weka tool. Data which is used for the experiment from hospital is pruned and tree generated by Weka. The data set consist of 100 records and 5 attributes. In which 75 records are used for training and 25 records are used for testing. The ROC curve is obtained from Weka (J48 algorithm) tool for dementia, Schizophrenia, ADHD, epilepsy and mood disorder. It Compares two operating characteristic that are true positive rate value and false positive rate value. Receiver operating characteristics curve (ROC) and area under curve (AUC) are calculated by Threshold Curve class in Weka.. This tool is used in many fields for diagnostic test evaluation. [8] Random Forest yielded the highest accuracy with the reduced feature subset. Random Forest is a powerful new approach to data exploration, data analysis and predictive modeling. It performs error detection, generation of strong predictive models, etc. The various feature selection methods will be evaluated in this paper to establish which one gives greater degree of accuracy for the binary categorization of brain EEG data.

3. MATERIALS AND METHODS:

The data set available at UCI (University of California, Irvine) Machine Learning Repository was utilized for this research. The EEG data set on the open or close state of the eye was used here. The various attributes correspond to the 14 channel readings on the EEG headset.**EEG Data 14 electrode positions:**AF3 , F7 , F3 , FC5 , T7 , P7 , O1 , O2 , P8 , T8 , FC6 ,F4 , F8 , AF4.

These can be broadly classified into cognitive suit that characterizes conscious thoughts, affective suit that describes emotions expressive suit for the facial expression and finally head rotation. The sensors - pads detect electrical activity on the surface of the brain that was used for binary categorization. The data set consists of 15 attributes including the eye open and close state. There are totally 575 records.

Data Mining Methods

The EEG data obtained for binary categorization was classified using several classifier algorithms after the important attributes were selected by the feature subset and feature ranking algorithms. The methodology adopted to perform the comparative study of data mining techniques in predicting Brain EEG data is depicted in Figure 1.

The datasets were collected and fed as input to the feature selection process. Correlation based feature subset selection method(CF subset) gave the optimal feature sets. It is an efficient feature selection algorithm, which gives high scores to subsets that includes features . These are the features that are highly correlated to the class attribute, but have low correlation to each other. CFS evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them [8-10]. Subsets of features that are highly correlated with the class while having low intercorrelation are preferred. The selected features were used to classify into two classes namely closed and open by using classification algorithms namely Bayes Net, Naive bayes , Random Tree, SMO, J48. This was

done to achieve the goal of being able to use the model to categorize the patient as Eyelid Open and Eyelid closed.

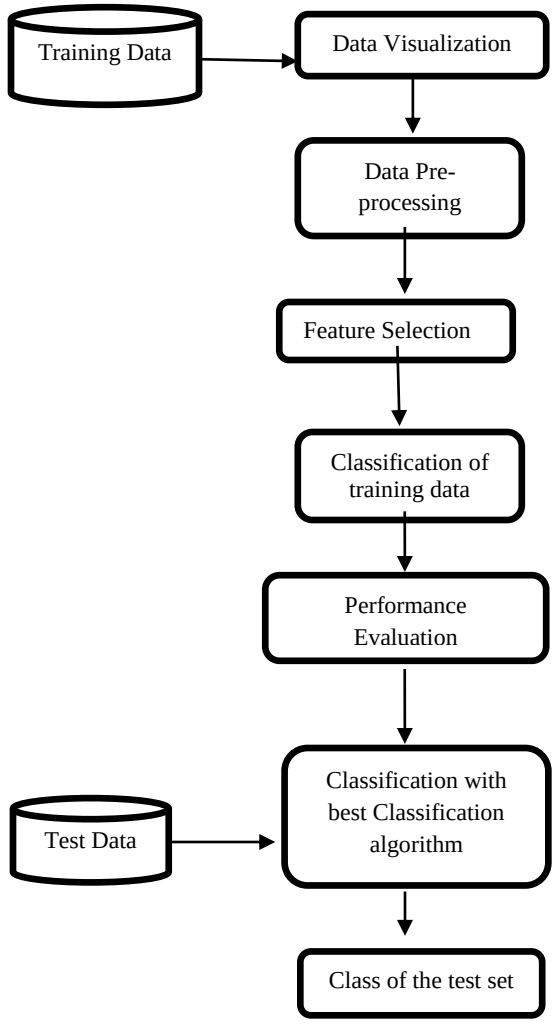


Figure 1: Performance Evaluation Methodology of Brain EEG Data

4. RESULTS AND DISCUSSION: The performance of the various classifier algorithms are studied here . This enables a performance evaluation of the proposed model. The various parameters used to interpret WEKA classification are :

a.Accuracy: The percentage of correctly classified instances is often called accuracy or sample accuracy. Accuracy= (TP+TN)/(TP+FP+FN+TN) . **b. Sensitivity*:*** Sensitivity, also called true positive rate, is estimated by calculating the percentage of correctly identified not-defective software modules and is formulated as: Sensitivity= TP/ (TP+FN) . **c. Specificity:** Specificity, also termed as true negative rate, is measured by calculating the percentage of correctly

recognized defective modules and is formulated as: Specificity= TN/ (TN+FP). **d. MCC**: It is a performance metric for binary classification, particularly when working with unbalanced classes.

(TP :True Positives, FP :False Positives,TN :True Negatives, FN :False Negatives)

Attribute Selection:CFS Subset Evaluation ; Search: Best First

Algorithm	Accuracy	Sensitivity	Specificity	MCC
BayesNet	97.7	0.975	0.977	0.949
NaiveBayes	95.1	0.938	0.951	0.889
SMO	98.6	0.991	0.986	0.969
Lazy.IBk	99.3	0.994	0.994	0.984
AdaBoost	99.1	0.993	0.991	0.98
RandomTree	96.7	0.965	0.967	0.926
J48	97.9	0.979	0.979	0.953

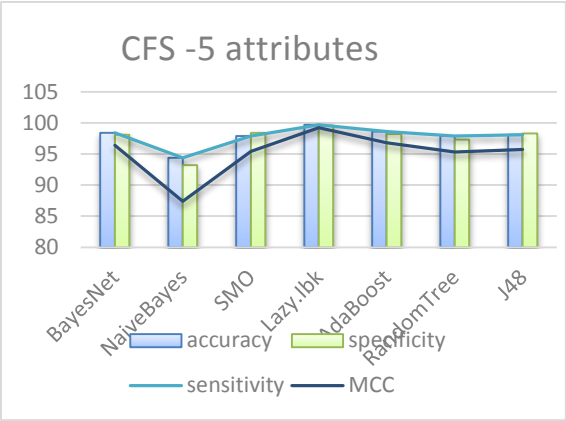


Fig 2. CFS subset evaluation – comparison of classification accuracy

Attribute Selection : Gain Ratio ; Search: Ranker Algorithm

Algorithm	Accuracy	Specificity	Sensitivity	MCC
BayesNet	97.2	0.954	0.972	0.937
NaiveBayes	92.0	0.871	0.920	0.816
SMO	98.3	0.992	0.983	0.962
Lazy.IBk	99.7	0.998	0.997	0.992
AdaBoost	98.1	0.985	0.981	0.957
RandomTree	97.7	0.975	0.977	0.949
J48	98.1	0.980	0.981	0.957

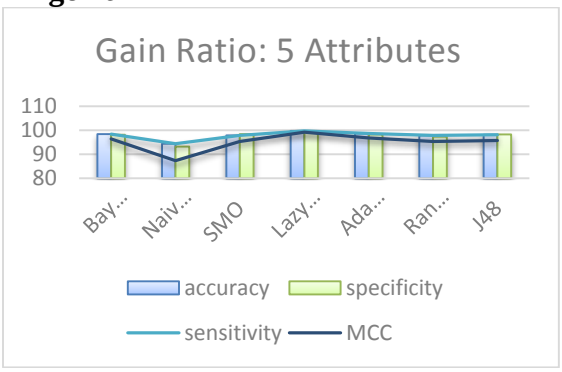
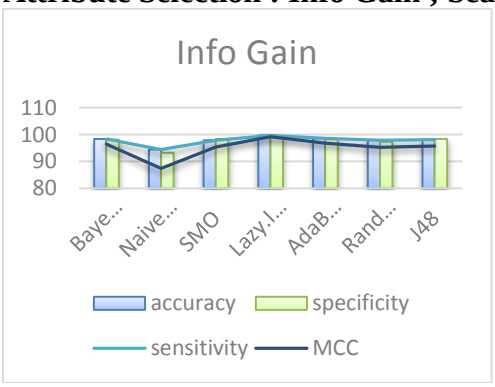


Fig 3. Gain Ratio – comparison of classification accuracy

Attribute Selection : Info Gain ; Search: Ranker Algorithm



Algorithm	Accuracy	Specificity	Sensitivity	MCC
BayesNet	98.4	0.981	0.984	0.964
NaiveBayes	94.4	0.932	0.944	0.874
SMO	97.9	0.984	0.979	0.954
Lazy.IBk	99.7	0.996	0.997	0.992
AdaBoost	98.6	0.982	0.986	0.968
RandomTree	97.9	0.973	0.979	0.953
J48	98.1	0.983	0.981	0.957

Fig 4. Comparison of accuracy of classification algorithms – Info Gain attribute selection

The CFS selects 5 attributes based on their correlation or importance to the data set and the maximum accuracy it can provide is 99.3%. The subset selected consisted of : F7 , FC5, P7 , O1 and O2. CFS evaluates the worth of a subset of attributes , by considering the individual predictive ability of each feature along with the degree of redundancy between them[5-7] . In EEG data , the potential obtained from each channel is recorded. If it is possible which of these attribute subsets taken produces an analysis with better accuracy , doctors can concentrate only on those in future. This eliminates unnecessary data mining because only what gives rise to an appropriate result will be analyzed.
In case of Info Gain and Gain Ratio, ranker algorithms are used. Attributes are ranked by their individual evaluations. The 5 attributes selected are : FC5 , P7 , F7 , O1 and FC6.

Out of all the methods studied , correlation based feature selection gives greater percentage of accuracy for nearly all the machine learning algorithms.

CONCLUSION:

The classifier which obtained highest accuracy with CFS was IBk . Other algorithms which attained a considerably good percentage of accuracy are SMO and AdaBoostM1 , while J48 and Random Tree can provide high level knowledge in the form of trees or rules. This study proposes a feature selection model with Instance Based k nearest neighbor algorithm (IBk) for binary classification of EEG data. Correlation based feature subset selection technique was used to select significant features which were helpful in the identification of coma patients . The study also revealed that there was an appreciable increase in accuracy in J48 , Random Tree , AdaBoost classifiers' performance that was developed using feature ranking which gave FC5 , P7 , F7 , O1 and FC6. This study has evaluated the predicting performance of proposed model for binary classification on EEG data. This study has also performed a comparative study on various machine learning algorithms with respect to the given data set . The experimental results show that the proposed model can give accurate results for the binary categorization of brain EEG data which can be used in a wide range of treatments.

References:

[1]Classification of EEG based Diseases using Data Mining – Deepika Kundra , Babita Pandey.

[2] Prediction of P53 Mutants (Multiple Sites) Transcriptional Activity Based on Structural (2D&3D) Properties – Shomona Gracia Jacob , R.Geetha Ramani.

[3] Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification - Tina R. Patil, Mrs. S. S. Sherekar .

[4] Random Model Trees: An effective and Scalable Regression Method - Bernhard Pfahringer.

[5] J. Kaur and Pallavi, "Data Mining Techniques for Software Defect Prediction", InternationalJournal of Software and Web Sciences (IJSWS), (2013), pp. 54-57.

[6] L. Guo, Y. Ma, B. Cukic and H. Singh, "Robust prediction of fault proneness by random forests", Proceedings of the 15th International Symposium on Software Reliability Engineering (ISSRE'04), (2004), pp. 417–428.

[7] Geetha Ramani R, Shomona Gracia Jacob., "Improved Classification of Lung Cancer Tumors Based on Structural and Physicochemical Properties of Proteins Using Data Mining Models", PLoS ONE (Impact Factor: 4.537) 8(3): e58772, 2013, ISSN: 1932-6203.

[8] J. Kaur and Pallavi, "Data Mining Techniques for Software Defect Prediction", International Journal of Software and Web Sciences (IJSWS), (2013), pp. 54-57.

[9] Geetha Ramani R, Shomona Gracia Jacob., "Improved Classification of Lung Cancer Tumors Based on Structural and Physicochemical Properties of Proteins Using Data Mining Models", PLoS ONE (Impact Factor: 4.537) 8(3): e58772, 2013, ISSN: 1932-6203.

[10] Geetha Ramani R., Shomona Gracia Jacob., "Prediction of P53 Mutants (Multiple Sites) Transcriptional Activity Based on Structural (2D&3D) Properties", PLoS ONE (Impact Factor:4.537) 8(2): e55401, 2013, ISSN:1932-6203.

[11] R. Geetha Ramani, Shomona Gracia Jacob., "Benchmarking Classification Models for Cancer Prediction from Gene Expression Data: A Novel Approach and New Findings", Studies in Informatics and Control Journal (Impact Factor: 0.587), 2013, ISSN: 1220-1766. In Press

[12] Shomona Gracia Jacob, Geetha Ramani R., "Design and Implementation of a Clinical Data Classifier: A Supervised Learning Approach", Research Journal of Biotechnology (Impact Factor: 0.143), Vol.8, No.2, pp. 16-26, 2013 ISSN: 0973-6263.

[13] Shomona Gracia Jacob, Dr.R.Geetha Ramani, "Data mining in Clinical Data Sets: A Review, International Journal of Applied Information Systems", Vol.4, No.6, pp.15-26, 2012. ISSN: 2249-0868.