Impact of Discretization on Classification of Data using Divide and Conquer Paradigm

Mridu Sahu, Shreya Sharma, Vyom Raj, N.K.Nagwani, Shrish Verma
National Institute of Technology Raipur
Chhattisgarh India

Abstract- Discretization of data has generally been proven to improve the classification accuracy in various data sets. Eye state classification of EEG data holds a lot of importance in the performance of prospective machines like Brain Computer Interfaces that work on EEG data. In this paper we have taken two sample data sets and seen the impact of discretization on classification does not always improve classification accuracy. Divide and Conquer approach has been applied on the large EEG data set to make its discretization easier.

Keywords—Discretization; EEG data; Divide and Conquer; Thoracic Surgery Data.

I. INTRODUCTION

Our paper attempts to combine three important facets of data mining while working on the two sample datasets- EEG data and Thoracic Surgery data. The main idea of this paper is to see the effect of discretization on the classification of datasets while using divide and conquer paradigm to make the dataset divide into segments that can be processed properly.

A. *Discretization*

Discretization can be defined as the process of converting real valued data to discrete valued data with the help of cut points [1]. It is a tool for making the tasks of machine learning easier. Even though a lot of new machine learning techniques are able to work on the raw real valued data[2], the task is less efficient. Discretization generally increases the efficiency of machine learning by providing a definite and discrete domain for the data set[3].

B. Classification

Another important aspect of data mining is classification. Classification is a process which attempts to separate the tuples of a data set by dividing them into groups or classes[4]. Classification makes the data more readable by giving it class labels. The tools which carry out classification are called classifiers[5]. Classifiers are generally divided into two types- supervised and unsupervised[6]. Supervised classifiers are the ones in which the classes are known beforehand. Unsupervised classifiers are the ones in which classes are not known previously. A lot of researches have shown that discretization significantly improves the performance of classifiers especially on biomedical data sets [7]. While this case might be true for a majority of data sets, we prove that it is not universally applicable.

C. *Divide and conquer*

Divide and Conquer is a methodology deployed to break large datasets into more manageable segments[8]. It basically involves dividing data into segments recursively up to the point where the segment size becomes practical to handle [9]. This creates a tree data structure where the factor of division into further segments is fixed, say n[10]. The task of proper discretization of EEG data is made very difficult due to the large number of instances, hence the Divide and Conquer approach has been implemented to divide data into more manageable chunks. Due to the manageable number of tuples, divide and conquer approach was not needed for this data set. It was directly fed into the discretization code and then to the Weka machine learning tool and the best classifier was evaluated.

D. *Overview of the paper*

The underlying sections deal with- 1. The literature review in this research area, 2. Methodology adopted in this particular research, 3. Classifiers (various classifiers used and the most suitable one for EEG and Thoracic surgery data), 4. Discretization (with main reference to Minimum Description Length Principle) and the final concluding section of results and discussion.

II. LITERATURE REVIEW

Various studies have been done in the field of discretization, classification and EEG data and many of them have aimed to see the correlated impact [8] [11]. The main aim of these researches have been increasing the classification accuracy of EEG data to make it a more practical data set that can be used in various devices, for example- BCI (Brain Computer Interfaces). The aim of our paper has been to check the performance evaluation of various classifiers with discretized and non discretized EEG data and Thoracic Surgery data.

Therefore the analysis of a few papers that have already appeared in this field becomes vital.

978-1-4673-9939-5/16/$31.00 ©2016 IEEE

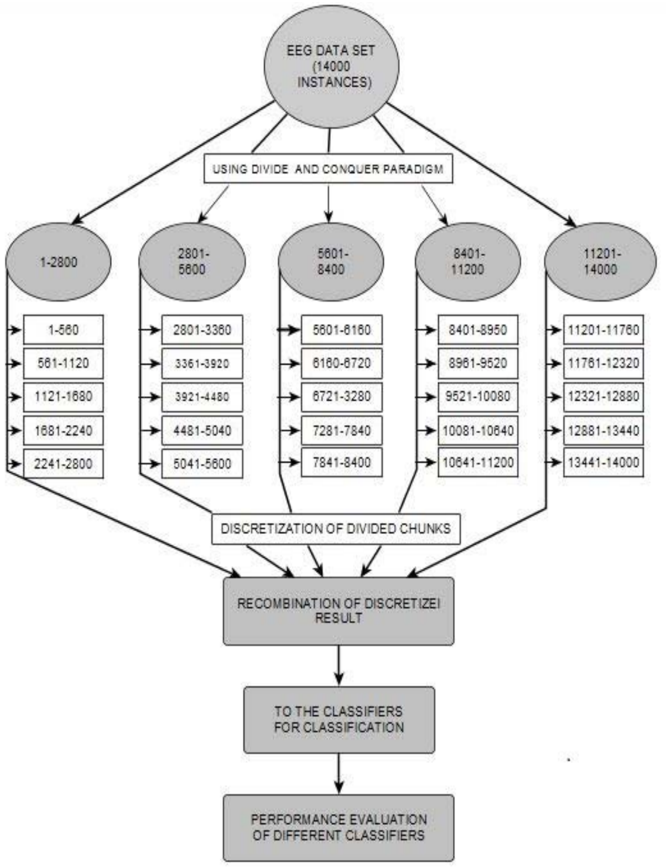


Fig. 1. Flow Diagram for EEG Dataset

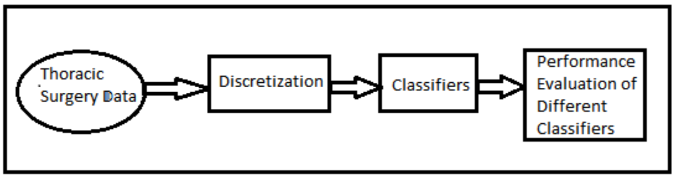


Fig. 2. Flow Diagram of Thoracic Surgery Dataset

In the paper literature [8], the IB1 classifier has proved to be the best instance based classifier giving an accuracy of 94.5087%. The experiments were carried out after data extraction and purification and dealt with just classification problem without the application of discretization as a pre filter. In another paper, Finding Non Dominant Electrodes Placed in Electroencephalography for Eye State Classification using Rule Mining [9], deals with the identification of non-dominant electrodes in an EEG system which can be eliminated while implementing the system practically and do not play much role in classification problems. The rankers search method with different evaluation algorithms has been used and it has been established that attribute removal considerably decreases time and space complexity in the building of a classification model.

Discretization and classification have also been studied as independent disciples themselves. Discretization is an important pre-processing tool in data mining. A lot of papers just deal with better discretization methods without referenced aim to a specific data set. In the paper- Discretization Methods with Back-tracking [12], the problem of finding an appropriate representative cut over a set of cuts has been discussed. Any discretization process has a set of possible cuts from which a subset is chosen for the process of discretization. This selection is done through s few basic steps. First a set of all possible cuts is taken, it is then divided into classes such that the cuts from the same class are indiscernible and then generally the median cut from every class is taken as the representative cut of the class. This method is not always correct while giving the best results and hence in the aforementioned paper, the obtained cut set is reconstructed after the completion of discretization process [12]. In yet another paper [13], Comparative Analysis of Supervised and Unsupervised Discretization Techniques; the two major classes of discretization techniques have been discussed. Unsupervised discretization methods are basically those classes of methods where the class labels are unknown. These methods were used traditionally like equal width methods, equal frequency methods etc. Supervised methods, on the other hand, come equipped with class labels and they use these labels while partitioning the continuous data. MDLP method is a well-known example of supervised discretization method. Another study on incremental discretization [14], Discretization from Data Streams: Applications to Histograms and Data Mining, deals with incremental discretization and compares it to the lines of batch discretization and the final output of incremental discretization is mapped to a histogram.

Classification is the process generally employed towards the end of data mining to put the processed data into broadly defining classes that enhances the better readability of data and can even help with enhanced readability and comprehensibility of data. Every classifier predicts the class label of the incoming data based on a previous training set and puts it into the closest appropriate class with this discretion. In the paper [15], Statistical Comparisons of Classifiers over Multiple Data Sets, an attempt has been made to compare various existing and even newly developed classifiers by comparing their performances on multiple data sets. Comparison of classifiers based on a single data set has been scrutinized for some time already because of the evident fact that some classifiers may be just suited for a particular data set and that does not guarantee it's superiority over any other classifier on the sole basis of that particular data set. This calls for randomly taken data sets that form the basis of the above stated work. In another paper dealing with the comparative study of classification algorithms [16], Bayes and Lazy classification algorithms have been compared. It is clearly established that Lazy classifiers have better performance evaluation than Bayesian classifiers. Lazy classifiers like K star mainly outperform Bayesian classifiers like Naïve Bayes in areas of correlated attributes and zero frequencies. The non-parametric nature of K star works well while dealing with such areas.

III. DATA SET DESCRIPTION

We have taken two data sets- EEG and thoracic surgery. EEG data refers to the data taken from the electroencephalography system which measures brain activities by means of electrodes. 16 electrodes named as F7, F3, F4, FC6, T8, P8,O2,CMS as eye open state and AF3, AF4, FC5, F8, T7, P7, O1 DRL as eye closed state are hooked on the patient's head and brain activity is measured [8]. We have taken an EEG sample dataset from the UCI repository. The dataset has 14 attributes and 14980 instances. After the application of this paradigm, the divided data chunks have been fed into the R programming code which discretizes them using MDLP method. The discretized results have been used as input in weka machine learning environment for classification using different classifiers and the best classifier has been recorded. Thoracic surgery data has been also been taken from the UCI repository. This data set is related to the classification of patients according to life expectancy after operations due to lung cancer [17]. The data has 17 attributes and 470 instances.

IV. DISCRETIZATION

Discretization refers to the conversion of real valued data into discrete values, generally having a finite well defined domain, which help them better serve the processes such as data mining. Any discretization process has a set of cuts over the domain of the dataset's attributes. The cut points form the most integral part of any discretization process in the sense that they are decisive of the discrete value of every real valued tuple. As a simple example- for a given data set let us say that attribute domain varies from real numbers in the range 0 to 10. Now the discretization cut points will basically determine, discrete values as follows- tuples ranging from 0 to 0.999 may be given discrete value 1, those ranging from 1 to 1.999 may be given value 2 and so on. This is obviously a very basic example and discretization methods generally do not apply such simple approaches.

The discretization methods are broadly divided into two types- supervised and unsupervised [20]. The supervised discretization methods come equipped with class labels.

A. *Minimum description length principle*

MDLP is the supervised discretization method that has been used in our study to discretize the EEG data set. For a dataset D, with an attribute A and let C be a cut point. Then we can write the information entropy of the partition put by C as I (A, C; D). Now this information entropy is defined as:-

$$I(A, C; D) = \frac{D1}{D} Ent\,|D1| + \frac{D2}{D} Ent\,|D2| \qquad (1)$$

Where Ent (Di) is the class entropy of Di. The Minimum Description Length Principle (MDLP) chooses a cut point Ca where the value of I (A, Ca; D) is minimum for all boundary values and then divides the dataset into two subparts at the cut point [18][28]. Subsequent process is followed to obtain successive cut points recursively [19].

V. CLASSIFICATION

In data mining, classification is the task of putting a new observation into a category based on the categories formed by training data set. Classification is basically defined in two steps. In the first step a classification model is made on the basis of the training set and class labels are determined. In the second step this classification model is used for determining the classes of unclassified instances on the main data set.

Various classifiers that have been used during the research work have been discussed here. The classifiers in the Weka explorer have been broadly divided into the following categories: - A. Bayes B. Lazy C. Meta D. Rules E. Trees [23].

A. *Bayes Classifiers*

They are the simplest class of classifiers which apply Bayes theorem for classification of datasets and keep naïve independence among data features. It is one of the oldest class of classifiers [27]. The main assumption behind every Bayes classifier is- the value of any tuple of an attribute in a dataset is independent of the value of all the other tuples of that attribute. Bayes classifiers are based on Bayes theorem and the statement of Bayes theorem is as follows:-

Let P (A) and P (B) be the probabilities of occurrence of events A and B respectively. Then the probability that event A will occur if event B has already occurred can be given as:-

$$P(A|B) = P\left(\frac{B}{A}\right)(A) \,/\, P(B) \qquad (2)$$

B. *Lazy Classifiers*

Lazy classifiers employ local approximation and are generally suited for datasets with fewer attributes and large number of instances [25]. Lazy classifiers are named so because they do not generalize data beyond training set until of manual intervention or query [29]. IB1 is a nearest neighbour lazy classifier which uses the Euclidean nearest neighbour distance to find the closest training set instance to the given data instance and puts the data instance in the same class of the training set instance [18]. Lazy K star is an instance based classifier which uses an entropy based similarity function. The class of the test instance is determined by the training instance most similar to it [18].

C. *Meta Classifiers*

Meta classifiers are based on Meta learning algorithms where automatic learning algorithms are used while also maintaining flexibility [30]. The main feature of a Meta classifier is that it offers flexibility in leaning, it has an inductive bias. This means that a Meta classifier will learn only when the bias matches the instance in data [21]. Meta decorate classifier uses specially constructed artificial training sets for classification of data [18].

D. *Rules Classifierrs*

Yet another category of classifiers is the rules classifier which works on rule mining and implementation [22]. Rules are first extracted from a decision tree. After the rules have been formulated they are used as conditions during classification of data.

Classification in EEG data we have dealt with is based on Eye State, where the eye state can be either 0 (for eye opened state) or 1 (for eye closed state). The main purpose of every classification algorithm applied on EEG data is the proper classification of eye states.

VI. METHODOLOGY

As it is evident from the Fig. 1 and 2, we have taken two data sets for our study- EEG and Thoracic Surgery data. Both these sets have been taken from UCI machine learning repository. For the understandings of this research, a brief introduction to data sets is needed. The EEG data set consists of 14 attributes and 14000 instances have been taken into account. A data set as large as that with 14000 attributes is difficult to discretize as a single unit as both the time and space complexity increases and results may not be very reliable. To curb this problem, divide and conquer approach has been adopted. The entire data set has been divided using the factor, n/5, creating a tree with two levels- the first level with chunk size 2800 and the second level with chunk size 560. This has created 30 distinct data sets in total which have been subjected to discretization using MDLP method. Each data set has been input to the R code for discretization in the csv format and a discretized result has been obtained. All the results from the 560 chunk files have then been combined to give a single result file which is stored in the arff format. This file has then been given as an input to the Weka explorer where its classification has been tested against various classifiers. The result obtained by each classifier has been tabulated in an excel sheet to compare these results and the best classifier has been found. The same procedure has been carried out for Thoracic surgery data, omitting the divide and conquer paradigm. This is because the Thoracic surgery data initially had only 478 instances and these many instances can be handled by the discretization tools. Pseudocode of the proposed work flow:-

```
1. Set INPUT = 'EEG Dataset'.
2. Use Divide and Conquer to divide the dataset.
3.Run the R code of discretization using MDLP on divided segments.
4. Combine the result obtained.
5. Use Weka explorer to classify the results with different classifiers.
6. Tabulate results for performance evaluation.
7. Take another INPUT1 = 'Thoracic Surgery Dataset'
8. Repeat steps 3 to 6.
9. End
```

VII. RESULT AND DISCUSSION

Raw, continuous and real valued data cannot readily be given for classification in most cases because it hampers the efficiency of classifiers and increases time and space complexity. For this purpose the EEG data and thoracic surgery data have been first discretized using MDLP. However, on classifying the discretized dataset with a variety of classifiers, it has been established that the best result is indeed obtained in the case of original EEG dataset and the classification accuracy, in fact, decreases after discretization for EEG data. Same is the trend observed in Thoracic surgery data. The following graph illustrates the classification accuracy with various classifiers on the original EEG data and discretized EEG data.

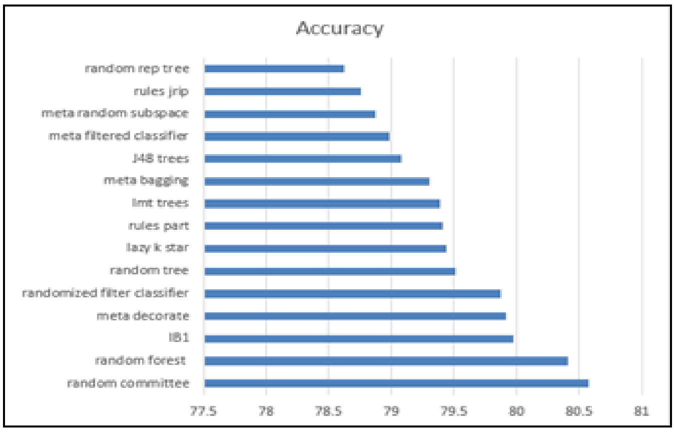


Fig. 3. Comparison of classifiers

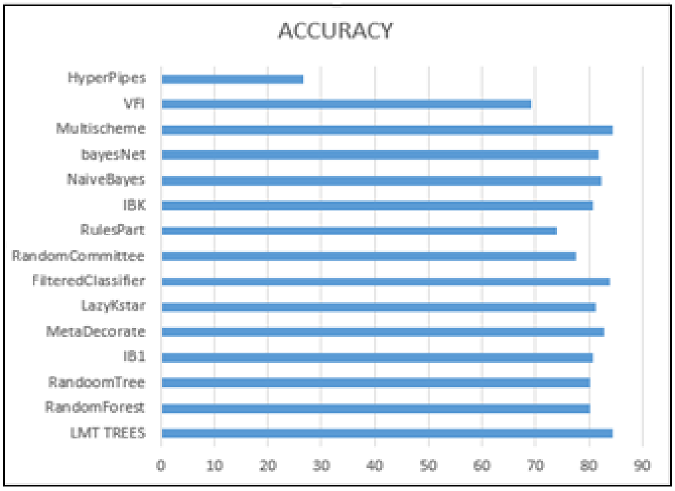


Fig.4. Comparison of classifiers on Thoracic Surgery Data

Fig. 3 clearly shows that the original dataset gets the classification accuracy of 95.03 percent after being classified with lazy k star. When discretized dataset is used, the classification accuracy has decreased and the best accuracy has been obtained in the case of random committee which is 80.57 percent, still less than the original accuracy of 95.03 percent. Similarly we can see in Fig. 4, that classification accuracy obtained in case of discretized thoracic surgery data is still less than the original dataset accuracy of 80.62%.

Before the comprehension of result and conclusions on it, a few terms are worth noting.

- Correctly classified instances' denotes the percentage of instances which were put in the correct class. A Fig. of 95.07 percent correctly classified instances mean that out of every 100 instances, approximately 95 will be put in the correct class.
- TP rate refers to the true positive rate, which denotes the 'True positive' classification or the measure of instances that were classified as positive and were actually positive. TPR= TP/ (TP+FN), where TP is True positive, FN is False Negative [24].
- FP rate follows the same analogy, it is the measure of instances that were classified as positive but were not actually positive. FPR = FP/ (FP+TN), where FP is false positive and TN is True Negative [26].

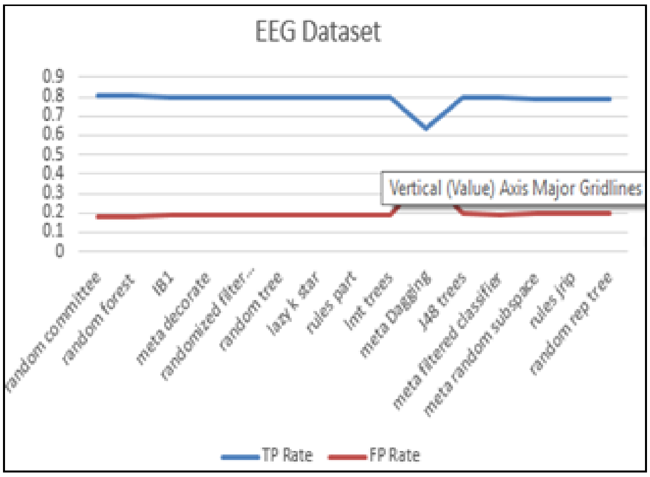


Fig. 5. TP and FP rates of classifiers on EEG

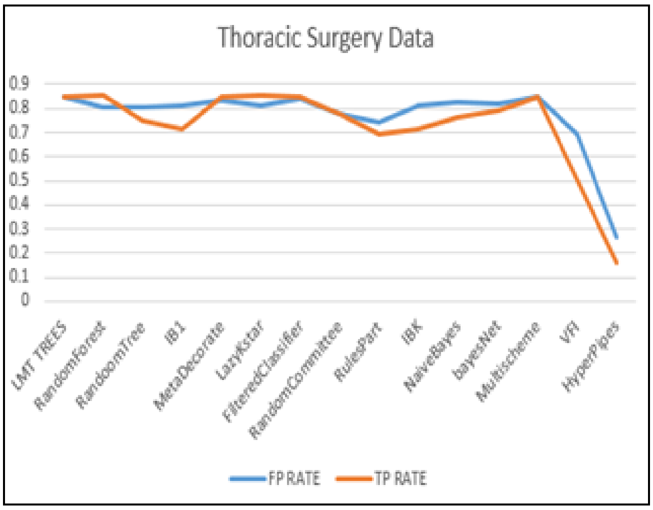


Fig. 6. TP and FP rates of classifiers on Thoracic Surgery data

TABLE I (Tabulation of performance of Classifiers on EEG data)

Classifier Algorithm	Accuracy	ROC Area	F measure	FP Rate
Randome Committee	80.57	0.896	0.802	0.179
Rendom Forest	80.41	0.895	0.8	0.18
IB1	79.98	0.808	0.796	0.184
Meta Decorate	79.98	0.887	0.795	0.185
Randomized Filter Classifier	79.87	0.885	0.795	0.185
Random Tree	79.51	0.875	0.791	0.189
Lazy Kstar	79.44	0.894	0.79	0.188
Rules PART	79.41	0.894	0.79	0.19
Tree LMT	79.41	0.882	0.79	0.19
Meta Bagging	79.3	0.883	0.789	0.191

The following tabulation shows the final list of the classifiers which yielded the best results in the classification of discretized EEG dataset. The accuracy of Random Committee is best in the case of discretized EEG dataset.

VIII. CONCLUSION

By this research experiment, it can clearly be established that while discretization of data helps in improving classification accuracy in most datasets, it is not universally true. There can always be some data sets like the two we have found (EEG and Thoracic Surgery Data) which may perform badly under classifiers after discretization. Whether there can be new discretization methods for such datasets is a scope for further research.

Acknowledgment

This research is supported by the National Institute of Technology, Raipur and many thanks to WEKA machine learning group as well as Rosler and Suendermann from Baden-Wuerttenberg cooperative state university (DHBW),Stuttgart, Germany for providing EEG corpus.

References

[1] Kotsiantis, Sotiris, and Dimitris Kanellopoulos. "Discretization techniques: A recent survey." GESTS International Transactions on Computer Science and Engineering 32, no. 1 (2006): 47-58.

[2] Colubi, Ana. "Statistical inference about the means of fuzzy random variables: Applications to the analysis of fuzzy- and real-valued data." Fuzzy Sets and Systems 160, no. 3 (2009): 344-356.

[3] Marzuki, Z., and F. Ahmad. "Data mining discretization methods and performances." lung 3, no. 32 (2012): 57.

[4] Tsai, Cheng-Jung, Chien-I. Lee, and Wei-Pang Yang. "A discretization algorithm based on class-attribute contingency coefficient." Information Sciences 178, no. 3 (2008): 714-731.

[5] Dash, Manoranjan, and Huan Liu. "Feature selection for classification."Intelligent data analysis 1, no. 3 (1997): 131-156.

[6] Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. "The WEKA data mining software: an update." ACM SIGKDD explorations newsletter 11, no. 1 (2009): 10-18.

[7] Lustgarten, Jonathan L., Vanathi Gopalakrishnan, Himanshu Grover, and Shyam Visweswaran. "Improving classification performance with discretization on biomedical datasets." In AMIA Annual Symposium Proceedings, vol. 2008, p. 445. American Medical Informatics Association,2008.

[8] Sahu Mridu, N. K. Nagwani, Shrish Verma, and Saransh Shirke. "An Incremental Feature Reordering (IFR) Algorithm to Classify Eye State Identification Using EEG." In Information Systems Design and Intelligent Applications, pp. 803-811. Springer India, 2015.Annual Symposium Proceedings, vol. 2008, p. 445. American Medical Informatics Association, 2008.

[9] Finding Non Dominant Electrodes Placed in Electroencephalography for Eye State Classification using Rule Mining. Mridu Sahu, N.K.Nagwani, ShrishVerma, Saransh Shirke.

[10] Friedman, Nir, Moises Goldszmidt, and Thomas J. Lee. "Bayesian Network Classification with Continuous Attributes: Getting the Best of Both Discretization and Parametric Fitting." In ICML, vol. 98, pp. 179-187. 1998.

[11] Hazarika, Neep, Jean Zhu Chen, Ah Chung Tsoi, and Alex Sergejew. "Classification of EEG signals using the wavelet transform." In Digital Signal Processing Proceedings, 1997. DSP 97, 1997 13th International Conference on, vol. 1, pp. 89-92. IEEE, 1997.

[12] Nguyen, Son H., and Hoa S. Nguyen. "Discretization methods with backtracking." In Proceedings of 5th European Congress on Intelligent Techniques and Soft Computing, vol. 201205. Heidelberg, Germany: Springer-Verlag, 1997.

[13] Dash, Rajashree, Rajib Lochan Paramguru, and Rasmita Dash. "Comparative analysis of supervised and unsupervised discretization techniques." International Journal of Advances in Science and Technology 2, no. 3 (2011): 29-37.

[14] Gama, Joao, and Carlos Pinto. "Discretization from data streams: applications to histograms and data mining." In Proceedings of the 2006 ACM symposium on applied computing, pp. 662-667. ACM, 2006.

[15] Demšar, Janez. "Statistical comparisons of classifiers over multiple data sets." The Journal of Machine Learning Research 7 (2006): 1-30.

[16] Vijayarani, S., and M. Muthulakshmi. "Comparative Analysis of Bayes and Lazy Classification Algorithms." International Journal of Advanced Research in Computer and Communication Engineering 2, no. 8 (2013): 3118-3124.

[17] UCI machine learning repository, Centre for machine learning and intelligent systems, Thoracic surgery data set.

[18] Weka Classifiers Summary Theofilis George-NektariosAthens University of Economics and BussinessIntracom-Telecom.

[19] An, Aijun, and Nick Cercone. "Discretization of continuous attributes for learning classification rules." In Methodologies for Knowledge Discovery and Data Mining, pp. 509-514. Springer Berlin Heidelberg, 1999.

[20] Le, Quoc V. "Building high-level features using large scale unsupervised learning." In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pp. 8595-8598. IEEE, 2013.

[21] Srinivas, Umamahesh, Vishal Monga, and Raghu G. Raj. "Meta-classifiers for exploiting feature dependencies in automatic target recognition." In Radar Conference (RADAR), 2011 IEEE, pp. 147-151. IEEE, 2011.

[22] Duin, Robert PW. "The combining classifier: to train or not to train?." In Pattern Recognition, 2002. Proceedings. 16th International Conference on, vol. 2, pp. 765-770. IEEE, 2002.

[23] Thepade, Sudeep D., and Madhura M. Kalbhor. "Novel data mining based image classification with Bayes, tree, rule, lazy and function classifiers using fractional row mean of cosine, sine and walsh column transformed images." In Communication, Information & Computing Technology (ICCICT), 2015 International Conference on, pp. 1-6. IEEE, 2015.

[24] Han, Hui, Wen-Yuan Wang, and Bing-Huan Mao. "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning." In Advances in intelligent computing, pp. 878-887. Springer Berlin Heidelberg, 2005.

[25] Cufoglu, Ayse, Mahi Lohi, and Kambiz Madani. "A comparative study of selected classifiers with classification accuracy in user profiling." In Computer Science and Information Engineering, 2009 WRI World Congress on, vol. 3, pp. 708-712. IEEE, 2009.

[26] Liu, Xu-Ying, Jianxin Wu, and Zhi-Hua Zhou. "Exploratory undersampling for class-imbalance learning." Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on 39, no. 2 (2009): 539-550.

[27] Jegou, Herve, Matthijs Douze, and Cordelia Schmid. "Product quantization for nearest neighbor search." Pattern Analysis and Machine Intelligence, IEEE Transactions on 33, no. 1 (2011): 117-128.

[28] Barron, Andrew, Jorma Rissanen, and Bin Yu. "The minimum description length principle in coding and

modeling." Information Theory, IEEE Transactions on 44, no. 6 (1998): 2743-2760.

[29] Veloso, Adriano, Wagner Meira, and Mohammed J. Zaki. "Lazy associative classification." In Data Mining, 2006. ICDM'06. Sixth International Conference on, pp. 645-654. IEEE, 2006.

[30] Sun, Yi, Mark Robinson, Rod Adams, Paul Kaye, Alistair G. Rust, and Neil Davey. "Using real-valued meta classifiers to integrate binding site predictions." In Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint Conference on, vol. 1, pp. 481-486. IEEE, 2005.

Impact of Ranked Ordered Feature List (ROFL) on Classification with Visual Data Mining Techniques

Mridu Sahu,Shreya Sharma,Vyom Raj, N.K.Nagwani, ShrishVerma
National Institute of Technology Raipur, Chhattisgarh
Raipur, India

***Abstract*—Classification of data is used in data analysis to group various instances in appropriate classes to enhance readability of data and study its characteristics easily. The main aim of every classification problem is the enhancement of classification accuracy. Ranked feature ordering helps in improving the classification accuracy by removing the least dominant features. Classification model uses only important features and eliminate least dominant, results obtained later can be better understood by plotting them in parallel plot using visual representation. Visual representation of classifier results helps in better comprehension and interpretation of results. Parallel plot is a one type of coordinate plot, that have ability to show any number of variables in one plane, the plane is either two dimensional plane or it is three dimensional plane. Parallel plot also shows relationships between variables. Proposed article used Electroencephalography (EEG) eye state classifier data for this account.**

Keywords: Classification; Ranked Feature ordering; visual representations; parallel plot.

I. Introduction

Classification is a one task performed by data mining tools [1], this task is useful for knowledge representation [2]. Classification model require feature subset selection [3]. Feature subset selection is a process to select important features towards decision variable. Filter out important one, needs ranking of these features using different scoring criteria [4]. After rank generation it is observed, if ordering is also done then possibility to improve accuracy of classification model. Data exploration needs better understanding of data using visual data mining techniques [5].Visual data mining maps data into two or three dimensional plane. Parallel plot is one of the visual techniques to represent the data in different coordinate positions. Proposed work uses EEG dataset taken from the UCI repository. The work outline is composed of two main parts. First we aim to improve the classification accuracy of the selected dataset by using feature ranking ordered list. The literature shows that “the m best features are not the best m features” [6]. If we take the inverse of this statement, we can say that the m worst features are not necessarily the worst m features [7]. This basically means that instead of selecting a subset of features (which is a NP hard problem), we can remove the most non dominant feature (MND) features, and the resulting subset can be called the set of most dominant features. This will clearly improve the classification accuracy as seen in this work. Next, we move to visualization of the obtained results. Even after the Improvement of classification accuracy it is not very easy for a layman, who may be the actual user of the interpreted results, to understand the result. This calls for visualization of these results. Parallel plot represent the results graphically and better understand them using human perception and the topologies of the plots. The underlying sections explain the main aspects of the paper in detail.

A. Classification

Classification is an important aspect of data mining which aims to enhance the readability of data. Classification refers to the process of dividing the instances of a dataset in groups called ‘classes’ based on certain parameters like correlation [8]. Classification helps in increasing the comprehensibility of data by giving it class labels. Every instance gets a class label which represents the class to which it belongs. Classifiers are data analysis tools which classify datasets. In this lazy classifier is used, literature shows [9], that it is a best classifier among different classifiers for this data set.

B. Feature Subset, Ranking and Ordering

The Feature subset selection from a given set of features is a non-polynomial time hard problem [10]. For this reason, it is desirable to obtain a way in which we can select features in a better manner. In the proposed work, the features have been selected in a ranked feature ordering manner. Ranked feature ordering, refers to the selection of features in the dominance order after obtaining results from the feature ordering algorithms. This ranked ordered feature subset is then fed into the classifier.

C. Visual Representation

Visual representation of classifier results helps in better comprehension and interpretation of results. Parallel plot is a concise way of representing the results. Parallel plot is an effective way to represent multidimensional data (such as the EEG dataset that has been used in this work) [11]. A parallel plot, in definition, is a two dimensional plot showing n dimensional data. For each dimension or attribute, we have a vertical axis, giving rise to an n number of vertical plots that are parallel to each other. The instances are then represented using connecting lines spanning across all dimensions.

D. Overview of the paper

The underlying sections deal with the 1. Literature Review in this area 2. Dataset Description 3. Methodology 4. Classification 5. Ranked Feature Ordering 6. Visual representation 7. Result and discussion 8. Conclusion

978-1-4673-9939-5/16/$31.00 ©2016 IEEE

II. Literature review

The field of classification and visual representation is a vibrant research arena in data mining and its applications. Numerous studies have been conducted in these fields. The proposed work deals with the improvement of classification accuracy in EEG dataset by using Ranked feature Ordering and representing the result in visual parallel, it is important that we take a look at the previous works of this field. In the paper, performance Evaluation of different classifier for eye state prediction using EEG Signal is used, the instance based (IB) classifier has proved to be the best classifier giving an accuracy of 94.5087%. The experiments were carried out after data extraction and purification and dealt with the classification problem with an aim to improve its accuracy. Statistical Comparisons of Classifiers over Multiple Data Sets, an attempt has been made to compare various existing and even newly developed classifiers by comparing their performances on multiple data sets. Comparison of classifiers based on a single data set has been scrutinized for some time already because of the evident fact that some classifiers may be just suited for a particular data set and that does not guarantee it's superiority over any other classifier on the sole basis of that particular data set. This calls for randomly taken data sets that form the basis of the above stated work. In another paper dealing with the comparative study of classification algorithms [12], Bayes and Lazy classification algorithms have been compared. It is clearly established that Lazy classifiers have better performance evaluation than Bayesian classifiers. Lazy classifiers like K star mainly outperform Bayesian classifiers like Naïve Bayes in areas of correlated attributes and zero frequencies. The non-parametric nature of K star works well while dealing with such areas. In reference to the feature ordering area, in the paper [13], feature ordering has been implemented with nine different algorithms. Lazy K star classifier is used to give best accuracy results. Paper [14], shifts our focus to the visual aspect of data mining. It deals with numerous data visualization techniques like parallel plots, star plots, survey plots, Andrew curves etc.. The choice of any visualization technique depends upon the end user needs and the nature of dataset being visualized. Some visualization is more suited to clustering and outliers, others are more suitable for representing correlations between data. The choice of an optimal visualizer will also depend upon the number of instances in the dataset.

III. Dataset description

In the proposed work, we have taken EEG dataset to study the impact of Ranked Ordered Feature List (ROFL) on classification and for the visualization of the obtained result. The dataset has been taken from the UCI repository. EEG data refers to the data taken from the electroencephalography system which measures brain activities by means of electrodes. 16 electrodes named as F7, F3, F4, FC6, T8, P8,O2,CMS as eye open state and AF3, AF4, FC5, F8, T7, P7, O1, DRL as eye closed state are hooked on the patient's head and brain activity is measured [15]. The EEG dataset we have taken in this work has 14980 instances and 15 attributes. Out of these 15 attributes, 14 attributes are the electrodes or features and the last attribute taken just give the state of the eye, open or closed (0 or 1). EEG data has many widespread applications today. It finds usage in the diagnosis of sleep apnea in patients [16]. Most of the modern Brain Computer Interfaces like Neurosky mindset and Focusband are based on EEG data. This makes the analysis of this dataset all the more important. A decreased feature subset of EEG data with increased efficiency will greatly revolutionize the influence of commercial brain computer interfaces making them more affordable.

IV. Methodology

The main aim of the proposed work is to study the impact of ranked feature ordering in classification accuracy and visualize the end result for better readability. For this purpose we have taken the EEG dataset. This dataset is then fed into the Weka machine learning tool and then classified using Lazy K star classifier [17]. The obtained accuracy is noted. In the next step, the original dataset is taken again and the outliers present in the dataset are removed. After this, the feature ordered set is created from the result of seven feature reordering algorithms. This feature ordered set is then studied to find the ranked ordered feature set (evaluated manually on the basis of ranking in the individual feature ordered lists). A new datasheet is built on the basis of the ranked ordered feature list. This ranked feature ordered dataset is then fed into the Weka K star classifier and then classification accuracy is noted. In the final step for the better comprehensibility of these classification results, they are visualized in parallel plots and are studied for further interpretation. Fig. 1 explains this methodology concisely. Pseudo code for the proposed workflow is as follows:-

```
1. Set Input as EEG Dataset data set.
2. Apply Classification Algorithm
3. Apply feature ranking using six algorithms-
   3.1 Ranker and Correlation Attribute Evaluator
   3.2 Ranker and GainRatio Attribute Evaluator
   3.3 Ranker and Relief Attribute Evaluator
   3.4 Ranker and InfoGain Attribute Evaluator
   3.5 Ranker and OneR Attribute Evaluator
   3.6 Ranker and Symmetrical Uncertain Attribute
       Evaluator
4. Create a ranked ordered feature list clubbing the six
   algorithms' results.
5. Apply Classification on ordered list.
6. Compare performance of classifier.
7. Visualize the obtained results in parallel plot
```

V. CLASSIFICATION

In the field of data mining, classification simply means putting instances into their respective groups.

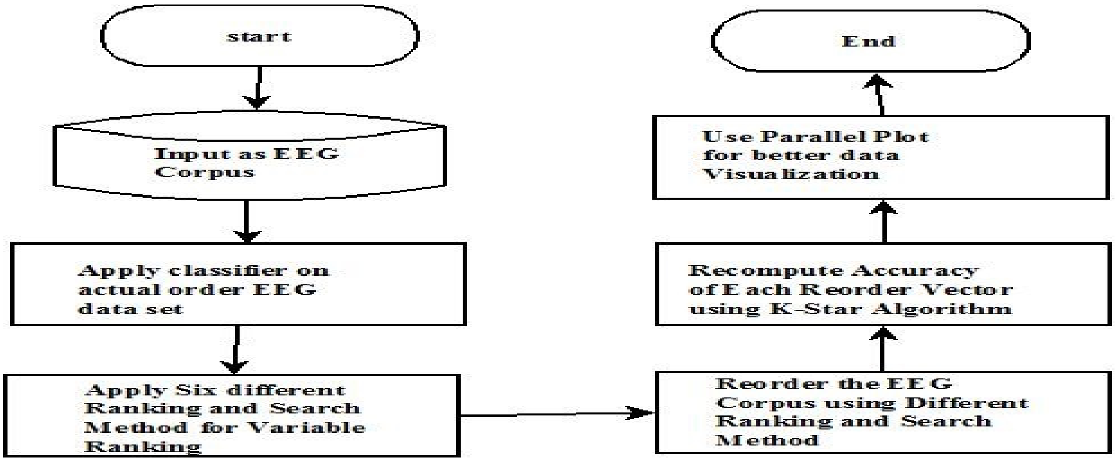

start
Input as EEG Corpus
Apply classifier on actual order EEG data set
Apply Six different Ranking and Search Method for Variable Ranking
Reorder the EEG Corpus using Different Ranking and Search Method
Recompute Accuracy of Each Reorder Vector using K-Star Algorithm
Use Parallel Plot for better data Visualization
End

Fig.1. Flow Chart of proposed article

Classification can be defined as the process in which instances in a dataset are assigned appropriate class groups. Every classifier achieves this goal in two basic steps- first a training set is used to define classes and class labels and then the dataset is put into classes based on these class labels [18]. Every instance is taken by the classifier and checked with the nearest instance in the training set. After this nearest instance is found, the dataset instance is put into the same class as the training set instance. Lazy classifiers are better suited for EEG dataset as they work best with datasets having fewer attributes and large number of instances. Lazy K star is an instance based classifier that uses entropy based similarity function. This classifier has been used in this work to obtain the highest classification accuracy. In the next step, each feature has been ranked on the basis of its relative position in every ordered list and then a single ranked ordered feature set has been obtained. This set is hypothesized to increase the classification accuracy. The Ranked ordered list of attributes that was found for this EEG dataset is as follows:
O1, AF3, F7, AF4, P8, P7, F8, FC6, F4, T8, T7, FC5, O2, F3.

VI. Ranked Featured Ordering

Feature ordering refers to the ordering of algorithms in order of their dominance. Every feature ordering algorithm adopts a different strategy for ordering the features. Greedy strategy is generally the most widely adopted one. In this work we have used 6 different algorithms to obtain 6 feature ordered lists. This is presented in Table 1.

VII. Visual representation

Visual representation is a vital facet of data mining in terms of readability. Most of the data mining tools like discretization and classification give textual or numeric results post processing. Often such results are difficult to comprehend and the end users are not able to detect patterns or comprehend any useful results from them. This is the point where visual representation becomes important. In this work we have used parallel plot as the way of visual representation. The main reason behind the choice of parallel plot is the fact that it is well suited for multidimensional data and EEG dataset falls into that particular category. Apart from dimensionality, parallel plot also provides flexibility in the choice of dimensions and their reduction if needed [19]. A parallel plot, in definition, is a two dimensional plot showing n dimensional data. For each dimension or attribute, we have a vertical axis, giving rise to an n number of vertical plots that are parallel to each other. The instances are then represented using connecting lines spanning across all dimensions. Color coding techniques can also be applied to these plots to enhance readability and grasp hidden patterns [20]. In this work, parallel plot visualization has been used to depict classification result and interpret correlations accordingly.

VIII. Result and discussion

From Fig. 2 we can see that the classification accuracy of original EEG dataset reaches its peak value on using the lazy K star classifier, 95.01 %. After the completion of this work, having used Ranked Ordered feature subset, the classification accuracy was again calculated using lazy K star classifier and the accuracy clearly increased and 95.09 % has now been obtained. This work has proved that the classification accuracy for EEG dataset is enhanced with the use of ranked ordered feature subset. The results of classification obtained after ranked ordered feature has been visualized in parallel plot. Fig. 3 shows the parallel plot of the reordered dataset with 14980 instances. The parameters named v1, v2… v14 represent the attribute names in ranked order set stated earlier. The plot looks very dense and cluttered. The main reason behind such an appearance is the large number of instances that have been plotted in this visualization. As we have taken EEG dataset, 14980 instances have been plotted in both these plots. This means 14980 lines in total which makes the plot cluttered. Clustering of lines brings relatively eases the plot [19-20], too many lines that seem to have bundled up together clearly have high correlation and this fact can be exploited to replace them with a single thicker line. If we still want to study individual lines, without clustering, we can take lesser number of instances as shown in Fig. 4. In this Fig. we have just taken 50 instances and 3 top features denoted by v1, v2, v3.

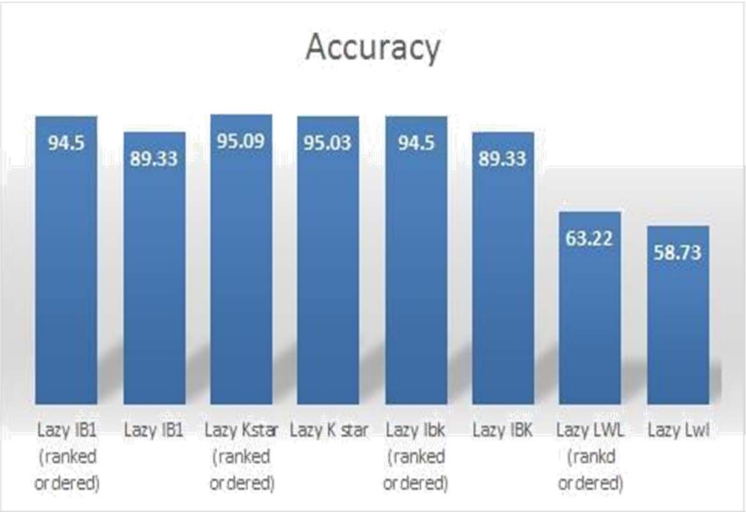


Fig.2. Classification results in ranked ordered and original Dataset

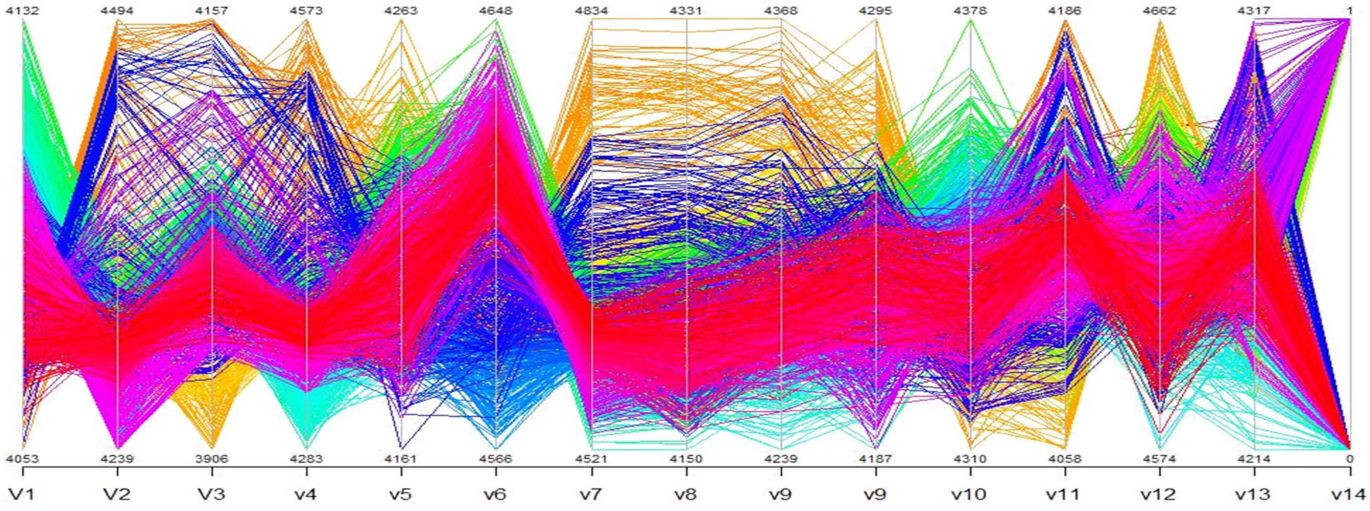


Fig. 3. Parallel plot of result with 14980 instances

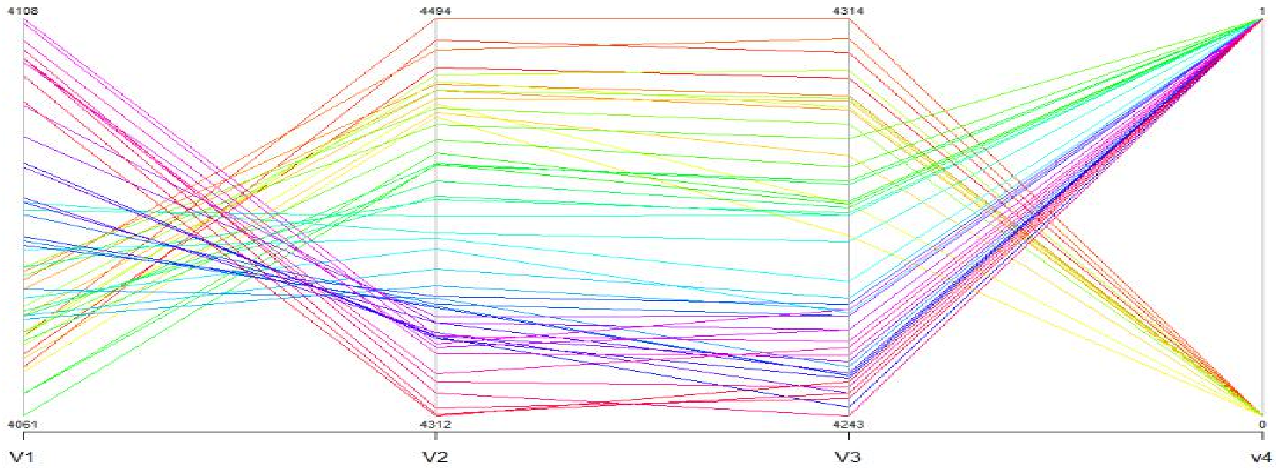


Fig. 4. Parallel Plot with 50 Instances

Table 1: Feature Ordering with K Star Classifier

S.No	Feature Ordering Algorithm	Feature Order Vector	K Star Accuracy
1	Ranker and Correlation	AF4,F7,F8,F4,T8,AF3,FC6,P7,P8,FC5,F3,O2,O1,T7	95.01
2	Ranker and Correlation	P8,AF3,O1,FC6,P7,AF4,F8,T8,T7,F4,FC5,O2,F3,F7	95.01
3	Ranker and Info Gain	O1,P7,AF3,AF4,F8,F4,P8,FC6,T8,O2,T7,FC5,F7,F3	95.01
4	Ranker and OneR	O1,P7,AF3,AF4,F8,P8,F3,T7,T8,FC5,O2,F4,FC6,F7	95.01
5	Ranker and Relief	F7,O1,P7,F8,FC6,AF4,AF3,T7,FC5,T8,F3,O2,F4,P8	95.01
6	Ranker and Symmetrical Uncertain	AF3,O1,P7,AF4,P8,F8,FC6,F4,T8,T7,FC5,O2,F7,F3	95.01

are actual three top ranked ordered features- O1, AF3, F7. The fourth feature is the eye state attribute which can have only two values 0 or 1. This plot looks less cluttered and hence can we can detect patterns more easily. For example, we can clearly see that when v3 (F7) crosses a middle threshold value; the eye state abruptly changes from 1 to 0. Many more useful patterns can be found through intensive study of this plot.

IX. Conclusion

From this research, we can conclude that ranked ordered feature subsets give higher classification accuracy as observed in case of EEG dataset taken. We also note that the visual representation of the obtained results, helped in interpreting data more easily than any other textual format. This is because the human mind is capable of perceiving patterns and recognizing them more easily in a visual representation.

Acknowledgement

This research is supported by the National Institute of Technology, Raipur and thanks to WEKA machine learning group as well as Rosler and Suendermann from Baden-Wuerttenberg cooperative state university (DHBW), Stuttgart, Germany for providing EEG corpus.

References

[1] Aggarwal, Charu C., ed. *Data classification: algorithms and applications*. CRC Press, 2014.

[2] Freitas, Alex A. *Data mining and knowledge discovery with evolutionary algorithms*. Springer Science & Business Media, 2013.

[3] Devaraj, Senthilkumar, and S. Paulraj. "An Efficient Feature Subset Selection Algorithm for Classification of Multidimensional Dataset." *The Scientific World Journal* 2015 (2015).

[4] Kavsaoğlu, A. Reşit, Kemal Polat, and M. Recep Bozkurt. "A novel feature ranking algorithm for biometric recognition with PPG signals." *Computers in biology and medicine* 49 (2014): 1-14.

[5] Otasek, David, Chiara Pastrello, Andreas Holzinger, and Igor Jurisica. "Visual data mining: Effective exploration of the biological universe." In Interactive Knowledge Discovery and Data Mining in Biomedical Informatics, pp. 19-33. Springer Berlin Heidelberg, 2014.

[6] Chandrashekar, Girish, and Ferat Sahin. "A survey on feature selection methods." *Computers & Electrical Engineering* 40, no. 1 (2014): 16-28.

[7] John, George H., Ron Kohavi, and Karl Pfleger. "Irrelevant features and the subset selection problem." In *Machine learning: proceedings of the eleventh international conference*, pp. 121-129. 1994.

[8] Mirkin, Boris. *Mathematical classification and clustering: From how to what and why*. Springer Berlin Heidelberg, 1998.

[9] Li, Jinyan, Guozhu Dong, Kotagiri Ramamohanarao, and Limsoon Wong. "Deeps: A new instance-based lazy discovery and classification system."*Machine Learning* 54, no. 2 (2004): 99-124.

[10] Zhong, Ning, Juzhen Dong, and Setsuo Ohsuga. "Using rough sets with heuristics for feature selection." *Journal of intelligent information systems*16, no. 3 (2001): 199-214.

[11] Andrienko, Gennady, and Natalia Andrienko. "Constructing parallel coordinates plot for problem solving." In *1st International Symposium on Smart Graphics*, pp. 9-14. 2001.

[12] Sahu, Mridu, N. K. Nagwani, Shrish Verma, and Saransh Shirke. "Performance Evaluation of Different Classifier for Eye State Prediction Using EEG Signal."

[13] Sahu, Mridu, N. K. Nagwani, Shrish Verma, and Saransh Shirke. "An incremental feature reordering (IFR) algorithm to classify eye state identification using EEG." In *Information Systems Design and Intelligent Applications*, pp. 803-811. Springer India, 2015.S.

[14] Keim, Daniel A. "Information visualization and visual data mining."*Visualization and Computer Graphics, IEEE Transactions on* 8, no. 1 (2002): 1-8.

[15] Wang, Ting, Sheng-Uei Guan, Ka Lok Man, and T. O. Ting. "EEG eye state identification using incremental attribute learning with time-series classification." *Mathematical Problems in Engineering* 2014 (2014).

[16] Lee, Jong-Min, Dae-Jin Kim, In-Young Kim, Kwang-Suk Park, and Sun I. Kim. "Detrended fluctuation analysis of EEG in sleep apnea using MIT/BIH polysomnography data." *Computers in biology and medicine* 32, no. 1 (2002): 37-47.

[17] Vijayarani, S., and M. Muthulakshmi. "Comparative analysis of bayes and lazy classification algorithms." *International Journal of Advanced Research in Computer and Communication Engineering* 2, no. 8 (2013): 3118-3124.

[18] Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. "The WEKA data mining software: an update." *ACM SIGKDD explorations newsletter* 11, no. 1 (2009): 10-18.

[19] Raidou, Renata Georgia, Martin Eisemann, Marcel Breeuwer, Elmar Eisemann, and Anna Vilanova. "Orientation-Enhanced Parallel Coordinate Plots." *Visualization and Computer Graphics, IEEE Transactions on* 22, no. 1 (2016): 589-598.

[20] Alon, Noga, Raphael Yuster, and Uri Zwick. "Color-coding." *Journal of the ACM (JACM)* 42, no. 4 (1995): 844-856.

978-1-4673-9939-5/16/$31.00 ©2016 IEEE

Applying auto regression techniques on ALS patients' EEG dataset with P300 speller

Mridu Sahu, Vyom Raj, Shreya Sharma
Department of Information Technology
National Institute of Technolgy Raipur
{mrisahu.it@nitrr.ac.in,vyomraj13@gmail.com,shreya.sharma16@gmail.com}

Abstract. This paper deals with the application of auto regression techniques to find the best fitting curves for the EEG data of ALS patients with P300 speller. ALS or amyotrophic lateral sclerosis is a degenerative neuron disease bringing gradual impairment of motor neurons leading to total loss of voluntary limb movement in sometime. A P300 speller is a 6X6 matrix of English alphabets in which each column and each row is highlighted periodically and the patient has to concentrate on the correct alphabet to evoke P300 event related potential. Auto regression is a curve fitting technique for sampled data. The best fit obtained in this study for the ALS patients' EEG channels which can used to predict incomplete or subsequent EEG data to enhance communication through P300 speller.

1 Introduction

1.1 Amyotrophic Lateral Sclerosis

ALS or Amyotrophic lateral sclerosis is basically a motor neuron disease. It involves the death or loss of functioning of neurons that control the voluntary movement muscles of the body, often leaving the patient devoid of any visible movements. It is a degenerative neuron disease bringing gradual impairment of motor neurons ranging from the cerebral cortex to the spinal cord [1]. In majority of cases, the cause of disease is not known but in a few of them, inheritance relationship can be seen. Generally along with loss of motor abilities cognitive impairment is also noticed in patients. Patients suffering from ALS may become quadriplegic and can completely lose the ability to communicate (even gestural communication subsides due to the total loss of limb functioning).

Due to these factors, important concerns arise about the ALS patients. Having lost their ability to generate any physical movements and total loss of communication due to no limb movements, ALS patients need special care in a lot of ways. Usually ALS patients need a lot of support in virtually all aspects of their life- rehabilitation support to maintain any residual motor function, dietary supports, respiratory supports and even augmentative communication devices [2]. In this work we attempt to study the communication aspect of ALS patients through Brain Computer Interfaces. Loss of motor abilities means that often the patients cannot do the most basic of daily tasks. Brain computer interfaces can help generating desired movements. The idea is simple- use EEG data to gauge the intention of the patient and bring alive that intention in the form of a physical movement by deciphering the EEG data. But the execution part of this seemingly plain approach is a challenging task in itself. In this study, we scrutinize the dataset obtained by a study on ALS patients with P300 speller and explore auto regression results that may fit suitably for the data of a particular patient.

1.2 Brain Computer Interfaces

Brain Computer Interfaces or BCIs are devices that actually bring about neurofeedback by giving the means to both recording EEG data and observing and manipulating the subject accordingly. A brain computer interface generally consists of small flat electrodes which are attached to the scalp with a little electrode paste that enables to record the EEG activity of the brain. The placement of EEG electrodes is generally based on the international 10-20 system, which is used to describe their position. In this system, the entire scalp is divided into five parts-frontal, temporal, central, parietal, and occipital denoted by the alphabets F, T, C, P and O respectively. In each of these parts the electrodes in the right part of the brain are denoted by even numbers and the electrodes on the left are denoted by odd numbers. For example- the electrode 'O2' in this system denotes the right electrode of the occipital region. The important point to note is that while BCIs generally have an EEG recording component, they are not limited to just the recording. Infact their main function starts after the EEG recording have been taken. The main function is to use the EEG recording of a person (BCIs can be based on EEG recordings or other measure of brain activity) as an alternative to their muscular movements and manifest their intent through these recordings [3]. A very simple example of a BCI is a word processing system where instead of typing words to convey a message, the BCI interprets the brain signals of a person and present it in the form of words written on a screen.

The main advantage of using BCI devices for communicative abilities of ALS patients is that they can be used in the moderate to severe stages of a patient's case because technically no ocular motor ability is required. Brain computer interfaces may also take another approach apart from EEG, that is, ERP approach. ERP is event related potential. ERP is a form of brain activity that is generated in response to internal or external events in a person's brain. P300 is a component of ERP [4]. A P300 speller has been used to obtain the EEG dataset which we have used in this study.

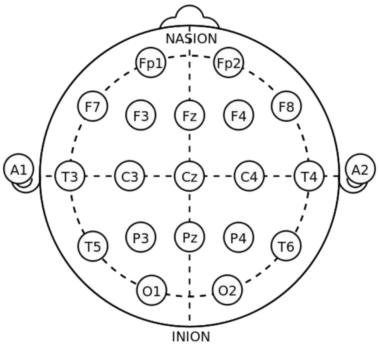
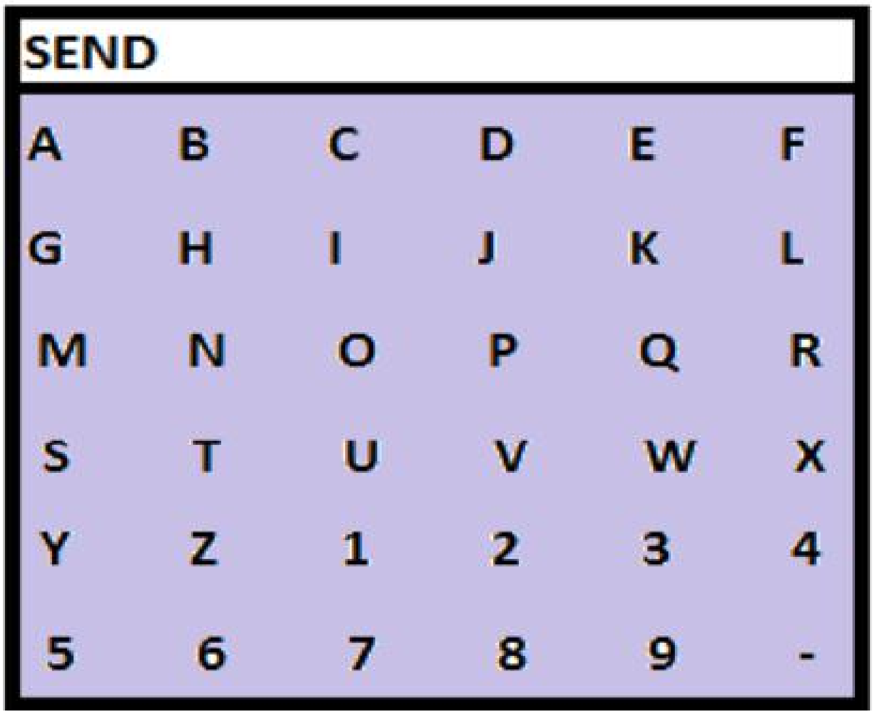


Fig 1. EEG electrodes and their naming system

1.3 P300 Speller

It has been seen that the P300 Event related potential, generated by EEG of external stimuli, can be used as a fairly reliable means serving as the basis of brain computer interfaces [5]. A lot of studies, starting from the first ever use of the P300 ERP, demonstrate that even a limited amount of ERP data recordings can provide effective communication [6] [7] [8] [9].

The P300 speller, first manifested by Farewell and Donchin, is a simple 6X6 matrix of English alphabet characters [9]. Communication ensues when the user focuses his/her attention on one of the 36 cells, each denoting a character. Each row and each column are intensified for 100 milliseconds and a P300 response is elicited when the row or the column containing the desired character is intensified. The rare presentation of the target stimuli in the random sequence of stimuli constitutes an Oddball Paradigm [10] and will elicit a P300 response to the target stimuli. This forms the basis of a P300 speller. As the underlying principle of a P300 speller is solely based on the P300 ERP response stimuli, little or no role is played by the transience of eye gaze.



SEND

A	B	C	D	E	F
G	H	I	J	K	L
M	N	O	P	Q	R
S	T	U	V	W	X
Y	Z	1	2	3	4
5	6	7	8	9	-

Fig 2. P300 speller matrix

2 Literature review

In paper [11], the main reason for adopting auto regression modeling data over other static parametric modeling techniques has been discussed. Auto regression modeling has been considered in particular with reference to physiological human body signals such as EEG. It is discussed that the non stationary nature of EEG signal over long periods of time make their analysis through non adaptive models unsuitable. Furthermore, if it is still desired to analyze EEG signals through non adaptive models, then it is suggested that they should be divided into small intervals wherein the signal remains stationary and then non adaptive AR modeling techniques like Levinson-Durbin and Burg algorithms can be used for time series analysis. In paper [12], the authors have investigated the statistical properties of the autoregressive (AR) distance between ARIMA processes. In the approach, the statistical properties of AR distances which highlight the dissimilarities between two given ARIMA processes have been studied. Paper [13] is a testimony to the effectiveness of AR modelling techniques. The authors have taken epilepsy EEG time series data and devised a coercively adjusted auto regression model for analysing the chosen data. Through their work, it is seen that auto regression modelling techniques can even be tailored to adapt modelling of data which is non linear or non periodic. In paper [14], adaptive AR modelling has been discussed as the data in consideration is nonstationary and is kept as such. The authors adopt multivariate time series analysis with Kalman filtering in the analysis of multichannel EEG signals. It has been discussed that such time series and frequency series analysis models can be solved by using fast forurier transform (FFT). But FFT is an interval oriented method and demands stationary values in the segments concerned. For this reason, parametric modelling techniques are preferred and in this paper, adaptive AR is the parametric modelling technique adopted. Paper [15], is yet another article dealing with the modelling of non stationary signals like EEG using AR techniques. Here ARMA, auto regressive moving average model has been used as an enhancement over the conventional AR model for achieving accurate modelling on non stationary signals. In paper [16], linear, non linear and feature selection methods for the classification of EEG signal have been discussed and compared. The paper underscores the fact that the general assumption stating that non linear classifiers outperform linear classifiers in case of EEG data can often be erroneous when the data in consideration is noisy. Linear discriminant analysis (as linear classifier) and neural networks and support vector machines (as non linear classifiers)were tested on the same set of EEG data and it was established that non linear classifiers produced only slightly better results than their linear counterparts at the cost of much increased complexity.

3 Communicative abilities of ALS patients via P300 speller

As noted earlier, ALS patients lose their voluntary muscle movement ability, sooner or later during their disease. Augmentative communication brought about by P300 speller, enhances and opens a way for ALS patients to communicate with the outer world. It has already been demonstrated by Farewell and Donchin [9], that a P300 evoked string of characters on a P300 speller can be used as a substitute to typing in adults. In a study performed by Donchin et all [7], the use of P300 speller was tested on ALS patients. The main takeaway from such studies is the fact that even though communication via P300 spellers may not be as effective in ALS patients as normal people, due to a variety of reasons like involuntary eye ball movement resulting in difficulty to concentrate on one character or the inability of weak ALS patients to focus for a long time, nonetheless it can serve as a decent medium for communication for ALS patients with a scope of further improvement. In [17], P300 Speller has been tested on many disabled subjects including ALS patients. Classifiers best suitable for stroke and ALS patients using P300 speller have been tested. The work flow adopted in this paper has been shown below:-

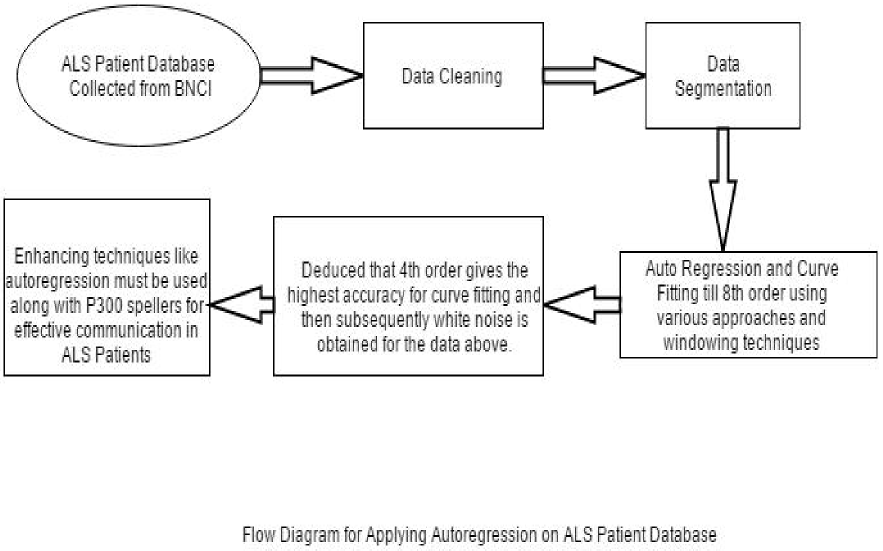

ALS Patient Database Collected from BNCI
Data Cleaning
Data Segmentation
Auto Regression and Curve Fitting till 8th order using various approaches and windowing techniques
Deduced that 4th order gives the highest accuracy for curve fitting and then subsequently white noise is obtained for the data above.
Enhancing techniques like autoregression must be used along with P300 spellers for effective communication in ALS Patients
Flow Diagram for Applying Autoregression on ALS Patient Database

Fig 3. Workflow Description.

4 Dataset Description

In this study we have taken the dataset from the BNCI horizon. This dataset was recorded in the study [18] and submitted to BNCI Horizon under goodwill. The dataset is a record of P300 evoked potentials of ALS patients in an experiment with P300 speller. The aim of the experiment was that the ALS patients must select the correct character in the 6X^ matrix of P300 speller from the selection epoch choices provided. Eight volunteers were used in the study all of whom were new to BCI training and had ALS diagnosis. Eight channels were used for recording EEG based on the 10-10 standard (Fz, Cz, Pz, Oz, P3, P4, PO7 and PO8). The EEG signal was digitized at 256 hz and bandpass filtered between 0.1 and 30 hz. Participants were made to spell seven words of five characters each on the P300 matrix speller. Rows and columns on the interface were randomly intensified for 125ms, with an inter stimulus interval (ISI) of 125 ms, yielding a 250 ms lag between the appearance of two stimuli (stimulus onset asynchrony, SOA). For each character selection (trial) all rows and columns were intensified 10 times (stimuli repetitions) thus each single item on the interface was intensified 20 times [18]. After taking the reading the recordings for each trial fell into one of the three categories- 0 (no activity) 1 (non target stimulus) 2 (target stimulus).

5 Auto Regression

Any parametric modeling technique is a process of taking a sampled data and finding the best fitting curve for it. The auto regression modeling technique can be formulated in frequency domain as a spectral matching problem or in time domain as a linear prediction problem [19][20]. In the time domain approach, we assume that the value of a sample is the weighted average of the previous ‘n’ values that occurred in the sample provided that the total numbers of samples in the sequence is very greater than ‘n’. Simply put, an output variable in an auto regressive model depends on the previous inputs and a random variable for that sampled data. This can be expressed mathematically as:-

$$X(t) = \sum_{i=1}^{N} a_i x_{t-i} + E_t \rightarrow eq.(1)$$

Where a_i= auto regression coefficients

x_t= series under investigation

E_t= Gaussian white noise

N= order (length) of filter

Gaussian white noise is absolutely essential to the application of auto regression on any dataset for curve fitting. In our study, we have checked and verified the presence of white noise and then auto regression techniques were applied.

5.1 Approaches of AR

These are the rules used for calculating the least square values in AR:-

Burg- It is a lattice based method which uses the lattice filter equations to calculate harmonic mean. This method aims to minimize forward and backward errors while simultaneously satisfying Levison- Durbin recursion [21]. In our work, we have used Burg’s approach because it accurately estimates the reflection coefficients and it resolves closely spaced sinusoidal signals (provided the noise is low). It is also computationally efficient for lower order auto regression models and hence made an ideal choice for our work, because AR optimized at 4th order of curve in this article.

Forward-backward (fb)- In this approach, sum of least square values is minimized for a forward model and then for a corresponding time reversed model. Fb approach is different from Burg’s approach in the sense that Burg uses unconstrained leats square algorithm to calculate AR coefficients whereas Fb first forms, forward and backward linear prediction estimates and then their corresponding forward and backward errors [22].

Geometric lattice (gl)- This approach is very similar to Burg’s approach, except that it uses geometric mean for the squared prediction errors instead of Burg’s harmonic means.

Least square (ls)- This approach simply attempts to minimize the standard sum of squared forward prediction errors.

Yule- Walker (yw)- This approach uses the Yule walker equations. In paper [23], the author explicitly discuss the shortcomings of Yule walker approach in noisy data channels and suggest the use of Burg’s method over Yule- Walker. Evidently, the use of Yule- Walker method should be avoided when a poorly conditioned co-variance matrix is obtained because a relatively small co-variance estimate bias can produce a large deviation in estimated parameters.

5.2 Windows in AR

The window concerned in auto regression deals with the past and future values relative to the data being considered in the moment. It pertains to the data outside the current interval. Different autoregression techniques can have one of the four (mentioned below) takes on the kind of windowing technique to be used:-

No- windowing- No windowing or now, as the name suggests is not concerned with past or present data's information. This is the default windowing technique in all approaches of autoregression (except Yule- Walkers method)

Post-windowing- This is concerned with the values coming after the current values in consideration and it replaces the end values (missing) with zeroes.

Pre and post windowing- This is concerned with both past and future values relative to the current values in consideration. It replaces both the past and future missing values with zeroes.

Pre windowing- This is concerned with the values that came before the current values in consideration. The past missing values are replaced with zeroes

6 Application of auto regression model to EEG dataset of ALS patients recorded with P300 speller

In this work, we have taken the EEG dataset of ALS patients recorded with P300 speller from the BNCI horizon and applied auto regression on the dataset of each of the 8 patients and on each of the 8 channels in MATLAB. The best fitting curves have been found for all of them and the results of auto regression for two patients (A01 and A02) have been shown below:-

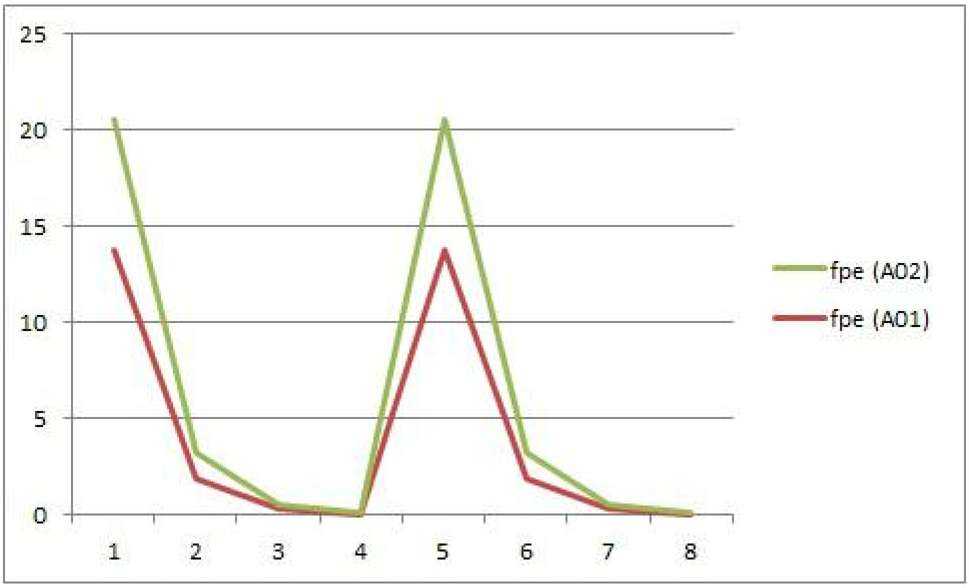
Subject	Signal	Approach	Window	Order	FPE	MSE	Accuracy	Z^0	Z^-1	Z^-2	Z^-3	Z^-4
A02	Fz	burg	now	1	6.75	6.75	76.67	1	-0.97			
A02	Fz	burg	now	2	1.32	1.32	88.66	1	-1.84	0.89		
A02	Fz	burg	now	3	0.19	0.19	96.04	1	-2.67	2.6	-0.92	
A02	Fz	burg	now	4	0.04	0.04	98.19	1	-3.49	4.91	-3.29	0.88
A02	Fz	burg	prw	1	6.75	6.75	76.67	1	-0.97			
A02	Fz	burg	prw	2	1.32	1.32	88.66	1	-1.84	0.89		
A02	Fz	burg	prw	3	0.19	0.19	96.04	1	-2.67	2.6	-0.92	
A02	Fz	burg	prw	4	0.04	0.04	98.19	1	-3.46	4.91	-3.29	0.88

Table 1. Results of auto regression for A02

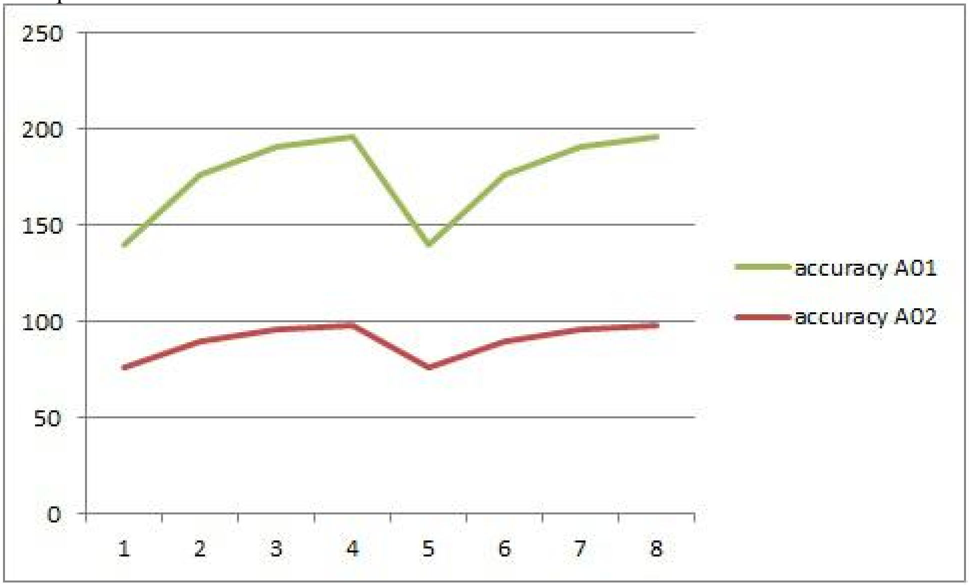
Table 2 Results of auto regression for A01

Subject	Signal	Approach	Window	Order	FPE	MSE	Accuracy	Z^0	Z^-1	Z^-2	Z^-3	Z^-4
A01	Fz	burg	now	1	13.75	13.75	63.65	1	-0.93			
A01	Fz	burg	now	2	1.899	1.899	86.49	1	-1.79	0.92		
A01	Fz	burg	now	3	0.313	0.313	94.51	1	-2.64	2.57	-0.91	
A01	Fz	burg	now	4	0.061	0.061	97.60	1	-3.46	4.88	-3.29	0.89
A01	Fz	burg	prw	1	13.75	13.75	63.65	1	-0.93			
A01	Fz	burg	prw	2	1.899	1.899	86.49	1	-1.79	0.92		
A01	Fz	burg	prw	3	0.313	0.313	94.51	1	-2.64	2.57	-0.91	
A01	Fz	burg	prw	4	0.061	0.061	97.60	1	-3.46	4.88	-3.29	0.89

In table 1, A01 is the concerned patient and the channel under consideration is Fz and in table 2, A02 is the concerned patient and the channel under consideration is Fz. We see four windows in the result sheet- now, prw, pow and ppw. For each window we have four fitting curves with the degree of coefficients increasing linearly from one coefficient to four (z^0 ,z^-1,z^-2 and z^-3). The parameter ‘MSE’ stands for mean squared error and is the main criteria for choosing the best fitting curve. Based on the mean squared error, the best fitting curve’s coefficients have been highlighted in both the tables. The significance of such fits in their usefulness by having the capacity to predict further values that can be expected in samples based on the best fit.



Graph 1- FPE vs Order



Graph 2- Accuracy vs Order

7 Result and Conclusion

In this work we have derived the best fitting curves of the EEG of eight ALS patients using P300 speller by using auto regression. From the literature review it is evident that Autoregression is the best parametric modeling technique for noisy data with small static intervals. As the EEG data of ALS patients satisfied this criteria, we deployed auto regression modeling techniques in our work. It is seen that we have chosen the 4th degree curve for our dataset. The reason for this choice is the high obtained accuracy in the 4th order coefficients curve (97.6 % for A01 and 98.19 in A02). The increasing accuracy in proportion to the increasing degree of the autoregression curve can be seen in the graph 2. We chose the 4th degree curve as the best fit because it delivered high accuracy with just four coefficients. Choosing higher orders like 5, 6 or 7 would provide a small accuracy increase (as it has already touched 97 % in 4th degree in A01 and 98% in 4th degree with A02) at the cost of more coefficients and increased complexity. That is why the fourth degree curve fit is the most optimal solution.

The further scope of this research lied in the fact that ALS patients fully or partially devoid of communicative abilities can effectively use P300 spellers for spelling out words and characters to enable communication with the people around them and enhancing techniques like autoregression models will make this communication more effective. Even though, its extensive and exhaustive use may be limited in scope for ALS patients now, applications of techniques such as auto regression for the generation of best fitting curves will further enhance the effectiveness of communication channels for ALS patients by providing ability to generalize and even predict future EEG recordings.

Acknowledgments. We thank BNCI horizon for providing us the ALS dataset which made this work possible.

References

[1] [1] Cipresso, Pietro, et al. "The use of P300-based BCIs in amyotrophic lateral sclerosis: from augmentative and alternative communication to cognitive assessment." Brain and behavior 2.4 (2012): 479-498.

[2] Mitsumoto, Hiroshi, and Judith G. Rabkin. "Palliative care for patients with amyotrophic lateral sclerosis:“prepare for the worst and hope for the best”."Jama 298.2 (2007): 207-216.

[3] Kübler, Andrea, et al. "Patients with ALS can use sensorimotor rhythms to operate a brain-computer interface." Neurology 64.10 (2005): 1775-1777.

[4] Donchin, Emanuel, Kevin M. Spencer, and Ranjith Wijesinghe. "The mental prosthesis: assessing the speed of a P300-based brain-computer interface."IEEE transactions on rehabilitation engineering 8.2 (2000): 174-179.

[5] Krusienski, Dean J., et al. "A comparison of classification techniques for the P300 Speller." Journal of neural engineering 3.4 (2006): 299.

[6] Bostanov V. BCI competition 2003-data sets Ib and IIb: feature extraction from event-related brain potentials with the continuous wavelet transform and the t-value scalogram. IEEE Trans Biomed Eng 2004; 51: 1057-61.

[7] Sellers EW, Donchin E. (in press). A P300-based brain-computer interface: Initial tests by ALS patients. Clinical Neurophysiology.

[8] Serby H, Yom-Tov E, Inbar GF. An improved P300-based brain-computer interface. IEEE Trans Neural Syst Rehabil Eng 2005; 13: 89-98.

[9] Farwell LA, Donchin E. Talking off the top of your head: Toward a mental prosthesis utilizing event-related brain potentials. Electroenceph clin Neurophysiol 1988; 70: 510-23.

[10] Fabiani M, Gratton G, Karis D, Donchin E. Definition, identification, and reliability of measurement of the P300 component of the event-related brain potential. Advances in Psychophysiology 1987;2:1– 78.

[11] Pardey, James, Stephen Roberts, and Lionel Tarassenko. "A review of parametric modelling techniques for EEG analysis." *Medical engineering & physics* 18.1 (1996): 2-11.

[12] Corduas, Marcella, and Domenico Piccolo. "Time series clustering and classification by the autoregressive metric." *Computational Statistics & Data Analysis* 52.4 (2008): 1860-1872.

[13] Kim, Sun-Hee, Christos Faloutsos, and Hyung-Jeong Yang. "Coercively adjusted auto regression model for forecasting in epilepsy EEG." *Computational and mathematical methods in medicine* 2013 (2013).

[14] Arnold, Matthias, et al. "Adaptive AR modeling of nonstationary time series by means of Kalman filtering." *IEEE Transactions on Biomedical Engineering* 45.5 (1998): 553-562.

[15] Grenier, Yves. "Time-dependent ARMA modeling of nonstationary signals." *IEEE Transactions on Acoustics, Speech, and Signal Processing* 31.4 (1983): 899-911.

[16] Garrett, Deon, et al. "Comparison of linear, nonlinear, and feature selection methods for EEG signal classification." *IEEE Transactions on neural systems and rehabilitation engineering* 11.2 (2003): 141-144.

[17] Manyakov, Nikolay V., et al. "Comparison of classification methods for P300 brain-computer interface on disabled subjects." Computational intelligence and neuroscience 2011 (2011): 2.

[18]. Reuderink, Boris, Mannes Poel, and Anton Nijholt. "The impact of loss of control on movement BCIs." IEEE transactions on neural systems and rehabilitation engineering 19.6 (2011): 628-637.

[19] Pardey, James, Stephen Roberts, and Lionel Tarassenko. "A review of parametric modelling techniques for EEG analysis." Medical engineering & physics 18.1 (1996): 2-11.

[20] Makhoul, John. "Linear prediction: A tutorial review." Proceedings of the IEEE 63.4 (1975): 561-580.

[21] Faust, Oliver, et al. "Analysis of EEG signals during epileptic and alcoholic states using AR modeling techniques." *IRBM* 29.1 (2008): 44-52.

[22] Kazlauskas, Kazys, and Rimantas Pupeikis. "A Forward–Backward Approach for Instantaneous Frequency Estimation of Frequency Modulated Signals in Noisy Environment." *Informatica* 23.1 (2012): 65-76.

[23] De Hoon, M. J. L., et al. "Why Yule-Walker should not be used for autoregressive modelling." *Annals of nuclear energy* 23.15 (1996): 1219-1228.