Investigating the performance improvement by sampling techniques in EEG data

[1]Mrs.V.Baby Deepa
Asst.Professor
M.Kumarasamy College of Engineering,
Karur- 639 113

[2]Dr. P.Thangaraj.
Dean, School of CT
Kongu Engineering College
Perundurai.Erode -638 052

[3]Dr.S.Chitra
Vice Principal
M.Kumarasamy College of Engineering
Karur- 639 113

***Abstract* - In this paper the performance of over-sampling methods such as SMOTE (Synthetic Minority Over-sampling Technique) and PCA (Principal Component Analysis) which are used for pre-processing are applied for the Brain computer interface dataset. The pre-processed data is used for classification by SMO and Naïve Bayes. In the EEG recordings, the transient events are detected while predicting the conditions of Central Nervous System and are classified as epileptic spikes, muscle activity, eye blinking activity and sharp alpha activity. The Pre-processing technique SMOTE is an over-sampling method which combines the informed over-sampling of minority class with random under-sampling of the majority class. Principal Component Analysis (PCA) is an exploratory data analysis technique. It involves a mathematical procedure which transforms a number of possibly correlated variables into smaller number of uncorrelated variables called Principal Components. It is mostly used as a tool in data analysis and for making predictive models. Based on the experimental results derived through SMOTE and PCA when they are applied to SMO and Naïve Bayes, it is concluded that PCA can be a better option since its performance improvement is better than that of SMOTE.**

Index terms – Synthetic Minority Over-Sampling Technique (SMOTE), Principal Component Analysis (PCA), Electro-encephalogram (EEG), Brain Computer Interface (BCI), Pre-processing.

I. INTRODUCTION

One of the fields which has developed over the few decades with the intent of exposing the internal brain states to the external world is Brian Computer Interface (BCI). It is a system that permits transformation of brain states to actions and overlooks the natural muscle pathways. A BCI system works by recording the brain signals and applying some machine learning algorithms to classify the brain state and performing a computer controlled action. Recording of brain signals is called Electro-encephalography. The EEG dataset [1] obtained from BCI is used for classification. The classification is done to make the dataset a balanced one. A dataset is imbalanced if the classes are not approximately equally represented. Here we apply SMO and Naïve Bayes for classification along with SMOTE and PCA individually to assess which sampling method improves the performance. In this paper the EEG data is explained in section II. Section III explains how pre-processing is done using the pre-processing techniques. Section IV contains the classification techniques - Naïve Bayes and SMO. Section V exhibits the experimental results.

II. EEG DATA

In EEG (Electro-encephalogram) [2, 3] signals, there would be a cluster of features. It is vital to extract the useful features from them. Identifying and extracting good features from the signals is a crucial step in the design of BCI [6, 7] (Brain Computer Interface). It is to be noted that if the features extracted from EEG [4, 5, 6] are not relevant and the neuro-physiological signals employed are not well described, then the classification algorithm which will use such features will have trouble in identifying the class of these features, i.e., the mental state of the user. Consequently, the correct recognition rates of mental states will be very low, which will make the use of the interface that is not convenient or even impossible for the user.

III. PREPROCESSING

Sometimes it is not impossible to use raw signals as the input of the classification algorithm. It is recommended to select and extract good features so as to maximize the performance of the system by making the task of the subsequent classification algorithm easier. According to some researchers, it is said that the choice of a good pre-processing and feature extraction method has more impact on the final performance rather than the selection of a good classification algorithm. Hence we employ two Pre-Processing methods SMOTE and PCA.

A. *Synthectic Minority Over Sampling Technique (SMOTE)*

The SMOTE [11] is an over-sampling method which combines the informed over-sampling of minority class with random under-sampling of the majority class. The number of synthetic samples generated by SMOTE is fixed in advance, thus not allowing any flexibility in the re-balancing rate.

(i) SMOTE Algorithm

```
Algorithm SMOTE(T, N, k)

Input: Number of minority class samples T; Amount of
SMOTE N%; Number of nearest neighbours k
Output: (N/100) * T synthetic minority class samples
(If N is less than 100%, randomize the minority class
samples as only a random percent of them will be
SMOTEd.)
  If N < 100 then randomize the T minority class samples
       T = (N/100) T
       N = 100
   end if
  N = (int) (N/100) (The amount of SMOTE is assumed to
be in integral multiples of 100.)
  k = Number of nearest neighbours
      nattr = Number of attributes
      Sample[ ][ ]: array for original minority class samples
       newindex: keeps a count of number of synthetic
samples
       generated, initialized to 0
       Synthetic[ ][ ]: array for synthetic samples
(Compute k nearest neighbours for each minority class
sample only.)
       for i ← 1 to T
            Compute k nearest neighbours for i, and save the
indices in the nnarray
            EEG(N, i, nnarray)
         endfor
EEG(N, i, nnarray) ( Function to generate the synthetic
samples.)
      while N 6= 0
         Choose a random number between 1 and k, call it
nn. This step chooses one of the k nearest neighbours of i.
        for attr ← 1 to nattr
       Compute: dif = Sample[nnarray[nn]][attr] −
                        Sample[i][attr]
            Compute: gap = random number between 0 and 1
              Synthetic[newindex][attr] = Sample[i][attr] +
gap dif
        endfor
         newindex++
          N = N − 1
      endwhile
        return (End of EEG)
End of Pseudo-Code.
```

B. *Principal Component Analysis (PCA)*

Principal Component Analysis is a linear transformation from a high dimensional data space to principal component feature space. It involves a mathematical procedure that transforms a number of possibly correlated variables into a smaller number of uncorrelated variables called Principal Components [8]. It is mostly used as a tool in exploratory data analysis and for making predictive models. PCA is useful for summarizing variables whose relationships are approximately linear or at least monotonic. It involves the calculation of the eigenvalue decomposition of a data covariance matrix or singular value decomposition of a data matrix, usually after mean centering the data for each attribute.

PCA [9] is theoretically the optimal linear scheme, in terms of least mean square error, for compressing a set of high dimensional vectors into a set of lower dimensional vectors and then reconstructing the original set. It is a non-parametric analysis and the answer is unique and independent of any hypothesis about data probability distribution. However, the latter two properties are regarded as weakness as well as strength, in that being non-parametric, no prior knowledge can be incorporated and that PCA compressions often incur loss of information.

PCA [10] is mathematically defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on. PCA is theoretically the optimum transform for given data in least square terms.

IV. CLASSIFICATION

In this section the two classification techniques which have been used for classifying the EEG data are explained. The classification techniques are Naïve Bayes classifier and SMO. While extracting the feature, classification plays a vital role and hence a good classification technique must be deployed.

A. NAÏVE BAYES CLASSIFIER

The Naïve Bayes classifier is [13, 14] known for its high efficiency and generalization ability. It has the advantage of good classification accuracy and is used widely in several domains. The Naïve Bayes classifier contains structure and parameters. It has star like structure as shown in the figure 1.

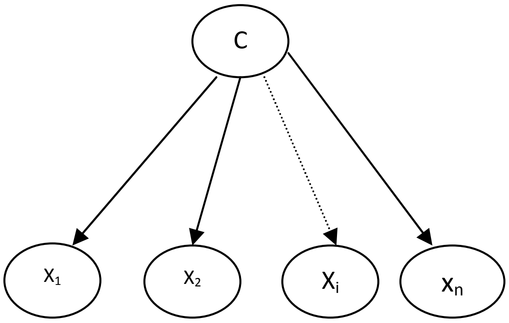


Figure 1: The structure Naïve bayes Classifier

The classifier has a conditional model over a dependent variable C over a dependent class variable *C* with a small number of outcomes or *classes*, conditional on several feature variables F_1 through F_n. Using Bayes theorem the classifier model is formulated as

$$p(C \mid F_1,\ldots, Fn) = \frac{p(C)\, p(F_1,\ldots, Fn \mid C)}{p(F_1,\ldots, Fn)}.$$

The Naïve Bayes classifier estimation includes parameter estimation of class and which is otherwise called prior probability estimation and conditional probability or density estimation. Since the classifier has high efficiency the EEG data have been classified using this technique. After the classification SMOTE and PCA are applied with this classifier to yield efficiency in classification. Once the SMOTE and PCA are applied with the Naïve Bayes separately the performance in the classification is observed.

B. SMO (Sequential Minimal Optimization)

SMO [13] is one of the well known classification techniques that are used in data mining. With this classifier also SMOTE and PCA are applied to see the performance in the classification. A notable feature of SMO is that it is very easy to implement, much faster and has better scaling properties.

V. EXPERIMENTAL RESULT

Naïve Bayes

Percentage of Correctly Classified Instances 43.45 %

Percentage of Incorrectly Classified Instances 56.54 %

SMO

Percentage of Correctly Classified Instances 52.97 %

Percentage of Incorrectly Classified Instances 47.02 %

Naïve Bayes classification with SMOTE sampling

Percentage of Correctly Classified Instances 60.71 %

Percentage of Incorrectly Classified Instances 39.28 %

SMO Classification with SMOTE sampling

Percentage of Correctly Classified Instances 64.28 %

Percentage of Incorrectly Classified Instances 35.71 %

Naïve Bayes Classification with Principal components sampling

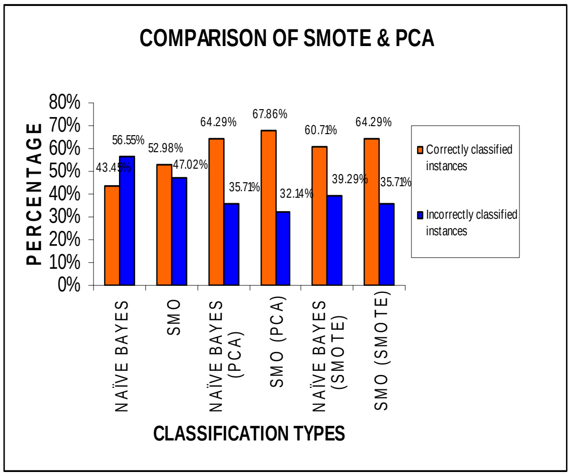
Percentage of Correctly Classified Instances 64.28%

Percentage of Incorrectly Classified Instances 35.71 %

SMO Classification with Principal components sampling

Percentage of Correctly Classified Instances 67.85%

Percentage of Incorrectly Classified Instances 32.14 %



VI. CONCLUSION

Based on the experimental results derived through SMOTE and PCA it is concluded that the performance improvement of PCA is better than SMOTE. It can be concluded that classification of data can be improved significantly to identify the rare events from the datasets by using PCA rather than the SMOTE for BCI dataset. However more investigation needs to be carried out for other related BCI dataset.

VII. REFERENCES

[1] A feature-based approach to combine functional MRI, structural MRI and EEG brain imaging data ,Calhoun, V.; Adali, T.; Liu, J.;Engineering in Medicine and Biology Society, 2006. EMBS '06. 28th Annual International Conference of the IEEE Digital Object Identifier: 10.1109/IEMBS.2006.259810 Publication Year: 2006, Page(s): 3672 - 3675

[2] On data reduction in EEG monitoring: Comparison between ambulatory and non-ambulatory recordings, Casson, Alexander J.;Rodriguez-Villegas, Esther; Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE Digital Object Identifier: 10.1109/IEMBS.2008.4650553 Publication Year: 2008, Page(s): 5885 – 5888

[3] Data Mining an EEG Dataset With an Emphasis on Dimensionality ReductionJahankhani, P.; Revett, K.; Kodogiannis,V.; Computational Intelligence and Data Mining, 2007. CIDM 2007. IEEE Symposiumon on Digital Object Identifier: 10.1109/CIDM.2007.368903Publication Year: 2007 Page(s): 405 - 412

[4] EEG data classification with several mental tasks,Choi Kyoung ho; Sasaki, M.;Systems, Man and Cybernetics, 2002 IEEE International Conference on Volume: 6 Publication Year: 2002

[5] Classification of Epileptic EEG Data Using Multidimensional Scaling, Direito, B.; Dourado, A.; Vieira, M.; Sales, F.; Bioinformatics and Biomedical Engineering, 2008. ICBBE 2008. The 2nd International Conference on Digital Object Identifier: 10.1109/ICBBE.2008.134 Publication Year: 2008 , Page(s): 551 - 555

[6] EEG based BCI-towards a better control. Brain-computer interface research at aalborg university, Nielsen, K.D.; Cabrera, A.F.; Omar Feixdo Nascimento;Neural Systems and Rehabilitation Engineering, IEEE Transactionson Volume: 14,Issue: 2 Digital Object Identifier: 10.1109/TNSRE.2006.875529 Publication Year: 2006 , Page(s): 202 - 204

[7] Preprocessing and Meta-Classification for Brain-Computer Interfaces Hammon, P.S.; de Sa, V.R.; Biomedical Engineering, IEEE Transactions on Volume: 54 , Issue: 3 Digital Object Identifier: 10.1109/TBME.2006.888833 Publication Year: 2007 , Page(s): 518 - 525

[8] Feature detection in motor cortical spikes by principal component analysis, Jing Hu; Si, J.; Olson, B.P.; Jiping He; Neural Systems and Rehabilitation Engineering, IEEE Transactions on Volume: 13 , Issue: 3 Digital Object Identifier: 10.1109/TNSRE.2005.847389 Publication Year: 2005 , Page(s): 256 - 262

[9] A technique to increase performance of a distributive tactile sensing system through an application of the principal component analysis (PCA), Tongpadungrod, P.; Systems, Man and Cybernetics, 2003. IEEE International Conference on Volume: 5 Publication Year: 2003, Page(s): 4210 - 4215 vol.5

[10] A Weighted Principal Component Analysis and Its Application to Gene Expression Data, Pinto da Costa, J; Alonso, H; Roque, L; Computational Biology and Bioinformatics, IEEE/ACM Transactions on Volume: PP , Issue: 99 Digital Object Identifier: 10.1109/TCBB.2009.61 Publication Year: 2009, Page(s): 1 - 1

[11] Classification of Imbalanced Data by Using the SMOTE Algorithm and Locally Linear Embedding, Juanjuan Wang; Mantao Xu; Hui Wang; Jiwu Zhang; Signal Processing, 2006 8th International Conference on Volume: 3 Digital Object Identifier: 10.1109/ICOSP.2006.345752 Publication Year: 2006

[12] SMO algorithm for least squares SVM, Keerthi, S.S.; Shevade, S.K.; Neural Networks, 2003. Proceedings of the International Joint Conference on Volume: 3 Digital Object Identifier: 10.1109/IJCNN.2003.1223730 Publication Year: 2003, Page(s): 2088 - 2093 vol.3

[13] Investigating the Performance of Naive- Bayes Classifiers and K-Nearest Neighbor Classifiers, Islam, M.J.; Wu, Q.M.J.; Ahmadi, M.; Sid-Ahmed, M.A.; Convergence Information Technology, 2007. International Conference on Digital Object Identifier: 10.1109/ICCIT.2007.148 Publication Year: 2007, Page(s): 1541 - 1546

[14] Learning Naive Bayes Classifiers with Incomplete Data, Cuiping Leng; Shuangcheng Wang; Hui Wang; Artificial Intelligence and Computational Intelligence, 2009. AICI '09. International Conference on Volume: 4 Digital Object Identifier: 10.1109/AICI.2009.402 Publication Year: 2009, Page(s): 350 - 353

AUTHORS

[1]V.Baby Deepa, received her Bachelors and Masters degree in Computer Science from Barathidasan university, Trichy and received her M.Phil. degree as well from the same university. She has 11 years of teaching experience.

Besides being an Assistant professor in the faculty of software Engineering, she is serving as the head for the same faculty in M.Kumarasamy College of Engineering, Karur. She has presented papers on various topics in several National Conferences. She is a research scholar of Anna University Chennai and her research area is Fuzzy and Data Mining.

[2]Dr.P.Thangaraj**,** received the Bachelor of Science in Mathematics from Madras University in 1981 and his Master of Science Degree in Mathematics from the Madras University in 1983.He completed his M.Phil degree in the year 1993 from Bharathiar University. He completed his research work on Fuzzy Metric Spaces and awarded Ph.D degree by Bharathiar University.He completed the post graduation in Computer Applications at ICNOU in 2005.

He completed his Master of Engineering degree in Computer Science in the year 2007 at Vinayaka Missions University. Currently he is a professor and Dean, School of Computer Technology and Application, Kongu Engineering College, Anna University. His current area of research interest is in Fuzzy Metric Spaces and Data Mining.

[3]Dr. S. Chitra, is is the Vice Principal and head of the department of M.E computer science and engineering in M. Kumarasamy College Of Engineering, Karur. She has 17 years experience in teaching. She completed her BE, ME and Ph.D in Computer Science and Engineering. Her research area is Software Reliability in Software Engineering. She presented more than 17 papers including national, international conference and journals.