Advances in Parallel Computing Algorithms, Tools and Paradigms
D.J. Hemanth et al. (Eds.)

doi:10.3233/APC220022

Comparison of SVM and KNN Classifiers on an EEG Signal with a Simple Dataset

Gouri M S[a,1] and Dr. K S VijulaGrace[a]
[a]*Department of Electronics and Communication Engineering, Noorul Islam University Thuckalay, Tamil Nadu, India*

Abstract. Brain-Computer Interface (BCI), deals with controlling of different assistance devices by utilising brain waves. The application of BCI is not simply limited to medical applications, and therefore its research has gained significant attention. It was noticed that huge amount of research papers had been published based on BCI in the last decade through which new challenges are constantly discovered. BCI uses many medical techniques such as EEG, ECG, ultrasound scans etc. Here in this paper we many deal with EEG and a detail comparsion of two commonly used classifiers used in classification and regressions are been done and output were obtained.

Keywords. Brain Computer Interface, EEG , SVM ,Hyperplane, KNN

1. Introduction

EEG ,non-invasive technology for studying brain activity through records of cerebral waves obtained by the placement of electrodes along the scalp. EEG[1] signal is a complex signal. It represents the brain's electrical activity. The EEGis separable to a series of sinusoids like other signals. The majority of the processed EEG parameters arepower spectralanalyses based on representing the sinusoids' amplitude as a function of frequency.EEG analysis shows abnormalities in recorded signal waves in case of the presence of disorders. Each scalp area produces waves that allow obtaining the state of each cerebral section. The physicians can identify the characteristics of these abnormalities to estimate the disease and thus simplifies the clinical diagnosis. One of the main difficulties in studying EEG signals is represented by signal dimensions and noise presence[2]. This leads to a huge disadvantage because of the lack of signal understanding could bring an incorrect diagnosis. Therefore proper, efficient and accurate methods to support signals reading, pre-processing and storing, allowing appropriate diagnosis and therapy designing have become a necessity. The presence of a high number of features automatic data reduction and classification of EEG signals is often mandatory to support physicians in diagnosis definition. In the last few years, considerable results have been produced in the analysis of EEG signals and the extraction of useful information for brain studies [3]. As per the literature, there are several classification algorithms that are used to analyze EEG signals. The main

[1]Gouri M S,Noorul Islam Centre for Higher Education, Thuckalay; E-mail: gaurivysakh16@gmail.com

advantage of using classification algorithms is that it causes error reduction in diagnosis definition, thus supporting physicians in large-scale data analysis.

In this paper, SVM and KNN algorithms are been compared with a simple dataset and an EEG Dataset and accuracy and outputs are been compared. A basic Dataset comparing the salary details with respect to age has been classified. This output had been taken as reference output. Also, another dataset is an EEG Dataset of 10 students while they watch a MOOC video. The videos include both confusing and non-confusing ones.

2. Related Work

EEG [1]is a technique used forbrain signal acquisition from brain scalp using electrodes. EEG has many advantages[4]such as it is painless, has no much side-effective and providesa more accurate interpretation of signals. These signals have a frequency range from 0 Hz – to 100 Hzs[5].The main aim is to classify the EEG data in a more simple and effective manner. The following techniques have been used for achieving the goal. The pre-processing technique of these signals is performed by allowing the raw EEG signalsto pass through a band-pass filter. The significant features were then been extracted signals from the received EEG through this pre-processed EEG. In the time domain, EEG is used for the extraction of mobility, activity and complexity which are some of the statistical parameters that had to be measured. There are many algorithms [3]that are used for EEG-based projects mainly on diagnosis and in BCIs.

A Brain-Computer Interface, a communication systemthat does not require any peripheral muscular activity [6]. BCI systems allowpassingof commands fromone object to another (electronic device)through brain activation [7]. These interfaces can be ponderedas being one of the most effective waysfor people struggling with motor disabilities [8] to get communicated.

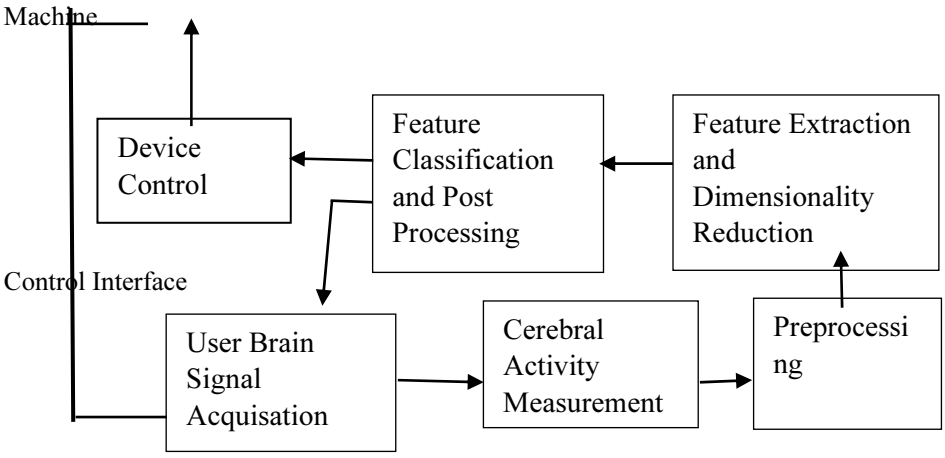

Figure . 1. Basic BCI

The BCI is been controlled by various brain activity patterns obtained from the user, that when inputted to the system get identified and converted into commands. In most of the BCIs that areavailable to date, this identification depends on a classification algorithm [3]. This algorithm is meant for automatic estimation of the data class as denoted by a feature vector [9]. The EEG collects the brain signal and measures the cerebral activity according to the delta (0.5 to 4Hz), theta(4 to 7Hz), alpha(8 to 12Hz),sigma (12 to 16Hz) and beta(13 to 30Hz) [10].As per the literature survey, it was found that SVM [5]and KNN[6] algorithms are usually used for the classification of EEG signals.

3. Support Vector Machine

SVM[4] is a machine learning-based method and is widely usedas kernel for classification tools. Recently, in EEG classification, the SVM[5] algorithm has commonly been used to identification of enormous brain diseases. Inorder to process these signals, spectral analysis has been performed but the most commonly found FFT method is constructed on simple sinusoid functions and is not appropriate for complex signals such as EEG ones. Usually,feature derivation is performed through temporal frequency analysis or by a continuous wavelet transform that is adjoiningto non-stationary signals(eg : EEG). This model was also used for many disorder predictions, where an autoregressive model (AR) was used for features extraction,preprocessing of data, andclassification of signals. Usually, linear discriminant analysis (LDA)[12] used features reduction.

Support Vector Machine , asupervised learning algorithmsfor classification as well as regression problems. The main utility of this algorithm isfor classification problems in machine learning. This algorithm aimsat creating the best decision boundary. This separates n-dimensional space into classes whichassists oneineasily putting the new data point in the correct category/class in the future.This best decision boundary so created is called a hyperplane.This algorithm thenpicks the extreme points/vectors that promotethe creation of the hyperplane. These extreme points are called support vectors, and so this algorithm is termed a Support Vector Machine.The dimensions of the hyperplaneare severely affected by the features in the dataset, that is for 2 features, the hyperplane will be a straight line and for 3 features, then will be a 2-dimension hyperplane.
SVM can be of two types:

3.1. Linear SVM:

Linear SVM is used when a straight line alonecan classify a dataset into twoclasses. Such data is linearly separable and therefore,the classifier so-called as Linear SVM classifier.Assume considering a dataset thatconsists oftwo tags (green and blue), and two features x1 and x2.The classifier available can classify the pair(x1, x2) of coordinates in either green or blue. Since this is a 2-D space, a straight line is enough forthe separation of theseclasses. But, the chances of having multiple lines for separating these classes is also a possibility. Therefore, the SVM algorithm serves to

search for the best line or decision boundary, and the so-called boundary or region is called a hyperplane.

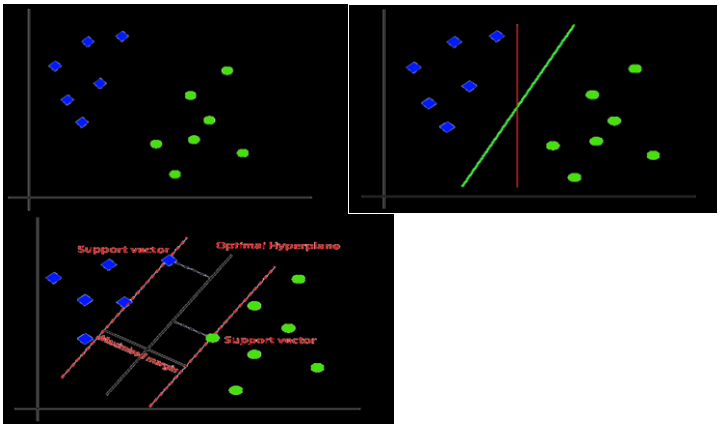

Figure 3.1.Linear SVM in a single straight line, multiple lines and hyperplane in 2D space

3.2. Non-linear SVM:

In areas where classification of the dataset cannot be donethrough a straight line nonlinear SVM shows its talent. The so-called data in the datasetis referredto as non-linear data and so classifier as Non-linear SVM classifier.If the data is non-linear, then we cannot draw a single straight line.

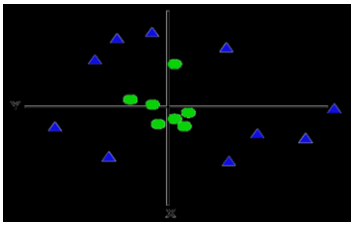

Figure 3.2. Nonlinear SVM

Therefore, to separate these data points,one more dimension has to be added. Linear data hastwo dimensions x and y, meanwhile, non-linear datawill add a third new dimension z that can be calculated as:

$$z = x^2 + y^2 \quad (1)$$

3.3. Python Implementation of Support Vector Machine

- Data Pre-processing step
- Fitting SVM classifier to a training set
- Predicting the test set result
- Creating the confusion matrix
- Visualizing training set resultwithtest set result

3.4. k-Nearest NeighborClassifier

k- Nearest NeighborClassifier[6], a non-parametric technicality and simplest classification algorithm.It is used for the identification of data points that are shattered into several classes for the classification prediction of a new sample point.The KNN algorithm believes that similar charactersremain in close proximity. KNN collects the idea of similarity (sometimes called distance, proximityetc) using the distance formula. Predictions are then madeby direct usage of the training dataset.After the entire training set search for the K most similar instance (The neighbors), predictions are created for a new instance (x) by bringing to the point the output variable for that K instance. It then determinesthose K instances in the training dataset that are the most similar to the new input and the distance measuring iscarried out.For the real-valued input variables,Euclidean distance[13] is most popularly used. The basic equation for Euclidean distance is shown below :

$$(x,xi)=sqrt(sum(xj-xij)^2) \quad (2)$$

where x is new point , xi is an existing point across all input attributes j.

The various distance measures available includethehamming distance for calculatingthe distance between binary vectors,themanhattandistance(city block distance)forthe distance between real vectors by addition of their absolute difference andtheMinkowskidistance which is a hybrid of Euclidean and Manhattan distance.

These k-Nearest Neighborsare broken down into 3 parts[14]:

- Calculate Euclidean Distance.
- Get Nearest Neighbors.
- Make Predictions

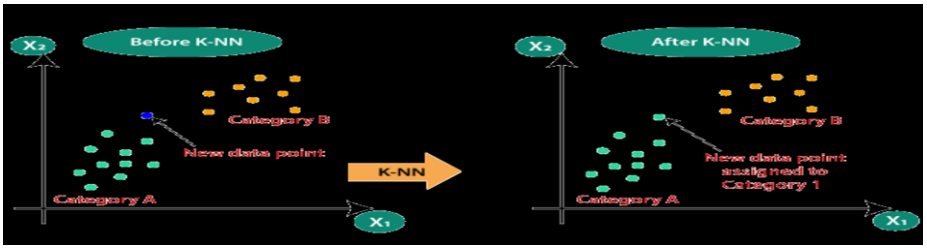

Figure 3.3. Before and After KNN

K-NN [12]is lazy learning algorithm that classifies new caseson a similarity measure basis(i.e., distance functions) and a given data point according to the majority in its neighbors. The KNN algorithm finalizesthe execution in two steps:by discovering the number of nearest neighbors and thencategorizing the data point into a particular class based on the first step output. Using theEuclidean distance, the nearest neighborhas been found. It picks the nearest k samples from the training set, and then earnsthe majority polls of the respected class where k should be an odd number inorder to avoid ambiguity.

4 Experimental Results

This paper, a comparison on the performance of SVM and KNN on the EEG dataset [15]with a simple dataset. Accuracy and outputs were then compared. A basic dataset comparing the salary details with respect to age had been classified. This output had been taken as reference output. Also, another dataset was an EEG Dataset of 10 students while they watched the MOOC video. The videos included both confusing and non-confusing ones. There were10 in each category. Each video has a length of 2 minutes and gotsplit in the middle of the topic to make the videos more confusing. In each video, the first 30 seconds and last 10 seconds were removedand collected the EEG data during the middle 1 minute.

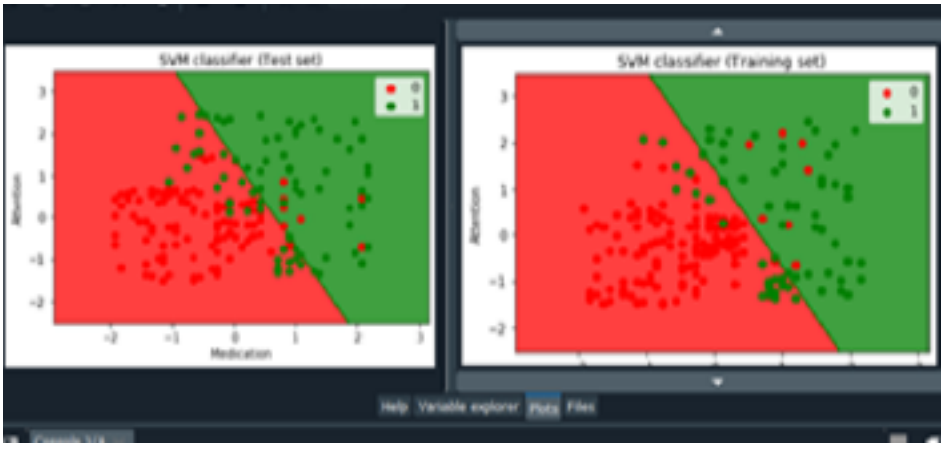

Figure . 4.1. SVMTrain and Test Dataset Visualisation on EEGDataset

Accuracy obtained after performing SVM is 84.2 %

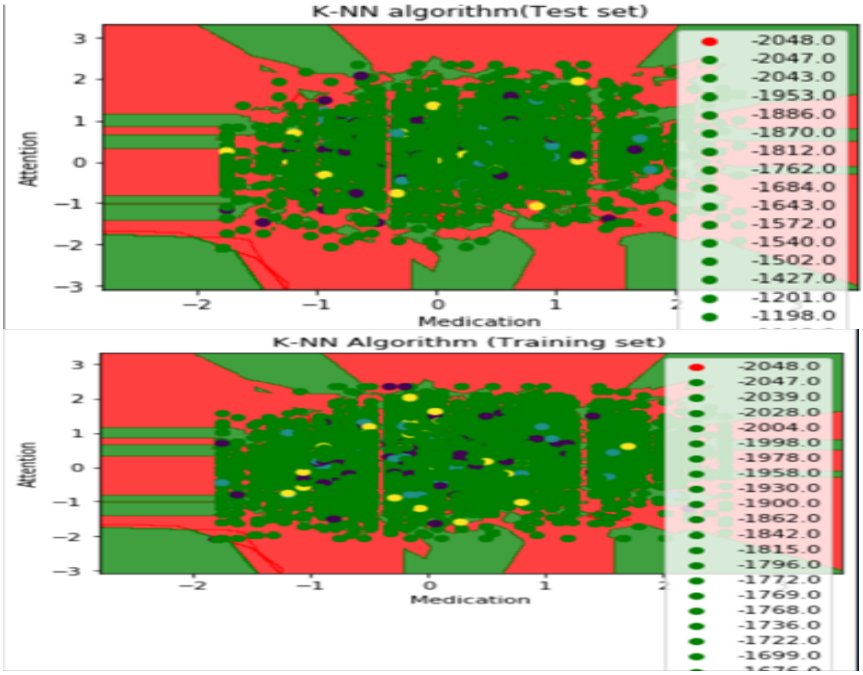

Figure . 4.2. KNN Training and Test Dataset Visualisation on EEG Dataset

Accuracy obtained after performing KNN is 48.7%%

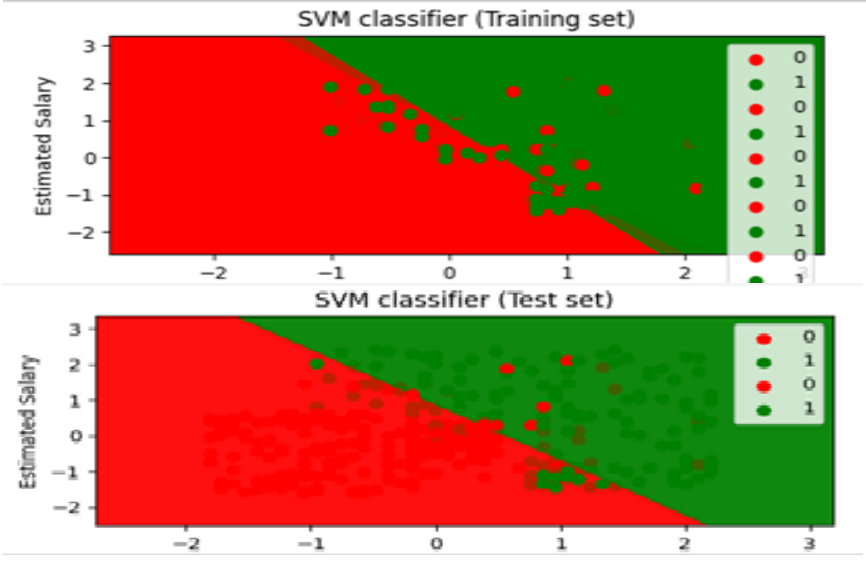

Figure 4.3. SVM Training and Test Dataset Visualisation in a Simple dataset

Accuracy obtained through SVM is 90 %

5. Conclusion

The experimental output shows that SVM provides better accuracy than KNN in EEG-based dataset which is 84.2%whereas KNN provides more accuracy than SVM in a normal simple dataset which is 93%. From the plots obtained it was also found that the output of the test data is a bit too confusing since EEG signal data is very much sensitive to noise and also due to its large margin. This is considered as a disadvantage for SVM when used in EEG-based data. This can be reduced through proper filtering of the EEG signal data.

References

[1] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5766087/#:~:text=The%20EEG%20is%20a%20complex,as%20a%20function%20of%20frequency
[2] Valeria Sacca´, Maurizio Campolo, Domenico Mirarchi, Antonio Gambardella, PierangeloVeltri, and Carlo Francesco On the classification of eeg signal by using an svm based algorythmChapter in Smart Innovation August 2018
[3] Subashini, A., L. SaiRamesh, and G. Raghuraman. "Identification and Classification of Heart Beat by Analyzing ECG Signal using Naive Bayes." In 2019 Third International Conference on Inventive Systems and Control (ICISC), pp. 691-694. IEEE, 2019.
[4] Analysis of electroencephalogram (eeg) signals and its categorization astudyProcedia Engineering38 2012 pp. 2525 – 2536
[5] MamunurRashid ,NorizamSulaiman , Anwar P P Abdul Majeed , Rabiu Muazu Musa , Ahmad Fakhri Ab Nasir , BiftaSama Bari and SabiraKhatunCurrent status, challenges, and possible solutions of eeg-based Brain-computer interface: a comprehensive review Frontier in robotics June 2020 vol. 14 art. 25
[6] JR Wolpaw, N Birbaumer, D J McFarland, G Pfurtscheller, and T M Vaughan Braincomputer interfaces for communication and control Clinical Neurophysiology 2002113(6)pp. 767–791
[7] T M Vaughan, W J Heetderks, L J Trejo, W Z Rymer, M Weinrich, M M Moore, A Kubler, ¨ B H Dobkin, N Birbaumer, E Donchin, E W Wolpaw and J R WolpawBrain-computer interface technology: a review of the second international meeting IEEE Transactions on Neural Systems and Rehabilitation Engeneering2003(2)pp. 94–109
[8] A Kubler, BKotchoubey, J Kaiser, JR Wolpaw and N Birbaumer Brain-computer communication: unlocking the locked Psychologybulletin 2001 127(3)pp.358–375
[9] R O Duda P E Hart and D GStork Pattern recognitionWiley inter science2001second edition
[10] https://www.ncbi.nlm.nih.gov/books/NBK539805
[11] Umer I Awan , U H Rajput , Ghazaal Syed , Rimsha Iqbal , IfraSabat , M Mansoor Effective classification of eeg signals using k-nearest neighbor algorithm IEEE 2016 978-1-5090-5300-1/16 pp. 120
[12] Siuly , Yan Li and Peng Wen Classification of eeg signals using sampling techniques and least square support vector machines P. Wen et al. (Eds.): RSKT 2009, LNCS 5589, pp. 375–382
[13] Altman An introduction to kernel and nearest-neighbor nonparametric regressionthe American Statistician 46, 175–185
[14] https://books.google.co.in/books?id=6H8IEAAAQBAJ&printsec=frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false
[15] https://www.kaggle.com/birdy654/eeg-brainwave-dataset-feeling-emotions