Epileptic Seizure Prediction Using Entropy Based Features of EEG

Mohammad Mohammad Beigi[a]

[a]Student, Department, Sharif University of Technology

This manuscript was compiled on February 7, 2024

This study addresses the unpredictability of epileptic seizures, affecting 1% of the global population. Leveraging MIT Physionet's EEG data from 24 epileptic patients, we propose an algorithm employing spectral entropy for seizure prediction. By calculating power spectral density and utilizing its frequency components as probability density functions, we extract features for prediction. Two classifiers, Support Vector Machine (SVM) and K-nearest neighbor (KNN).

Keywords: Shannon entropy | Epileptic seizure prediction | Support vector machine classifier | k-nearest neighbor classifier

In this research endeavor, our primary objective is to develop a robust model for detecting seizures in EEG signals. To achieve this, we utilized samples obtained from five subjects. Our approach involved the extraction of 9 seconds of EEG records preceding seizure occurrences. Given the prevalence of seizure samples, a stratified division was performed, isolating recordings without seizures into 9-second segments to serve as the negative class. Subsequently, we meticulously curated a dataset with balanced representation for both positive and negative classes.

Materials and Methods

Dataset. The CHB-MIT Scalp EEG Database is a collection of electroencephalogram (EEG) recordings acquired from patients with epilepsy. It's a dataset provided by the Children's Hospital Boston (CHB) and the Massachusetts Institute of Technology (MIT). The purpose of this database is to facilitate research in epilepsy and seizure prediction.

The dataset includes EEG recordings from different subjects, mainly children, during both interictal (non-seizure) and ictal (seizure) periods. Researchers use this data to study patterns in EEG signals and develop algorithms for seizure detection and prediction.

It's a valuable resource for the scientific community working on understanding and managing epilepsy, as well as for the development and evaluation of computational methods related to EEG analysis.

Shannon Entropy

In information theory, Shannon entropy quantifies the amount of uncertainty or surprise associated with a random variable. It's calculated using the formula $H(X) = -\sum_{i=1}^{n} P(x_i)\log_2(P(x_i))$, where X is the random variable, $P(x_i)$ is the probability of each possible outcome x_i, and the logarithm is base 2. Essentially, it measures the average amount of "surprise" in a probability distribution, with higher entropy indicating more unpredictability.

K-Nearest Neighbor

K-nearest neighbors (KNN) is a straightforward and versatile machine learning algorithm commonly used for both classification and regression tasks. The underlying principle of KNN is based on the idea that similar data points in a feature space tend to have similar outcomes. In the context of classification, when given a new data point, KNN identifies its k-nearest neighbors based on feature similarity and assigns the majority class among those neighbors to the new point. The choice of "k" refers to the number of neighbors considered, and it impacts the algorithm's sensitivity to local variations in the data.

The algorithm's simplicity is both a strength and a limitation. On one hand, KNN is easy to understand and implement, making it an accessible choice for beginners in machine learning. On the other hand, its computational efficiency can be a drawback for large datasets, as the algorithm requires storing and comparing distances to all data points. Additionally, the algorithm's performance can be sensitive to irrelevant or redundant features. Properly selecting the value of "k" and preprocessing the data to handle such issues are crucial steps in optimizing the effectiveness of KNN in a given problem.

In summary, KNN is a proximity-based algorithm that relies on the similarity of data points in a feature space to make predictions. Its simplicity and ease of use make it a popular choice, but careful consideration of parameters and data preprocessing is necessary for optimal performance.

Support Vector Machine

Support Vector Machines (SVM) is a powerful supervised machine learning algorithm used for both classification and regression tasks. It works by finding the optimal hyperplane that best separates different classes in the feature space. The "support vectors" are the data points that lie closest to the decision boundary or hyperplane, influencing its position and orientation. SVM aims to maximize the margin between classes, which is the distance between the support vectors and the decision boundary.

One of the strengths of SVM is its effectiveness in high-dimensional spaces, making it well-suited for tasks with many features. SVM can handle linear and non-linear classification through the use of different kernel functions, which transform the input data into a higher-dimensional space where a hyperplane can effectively separate the classes. This flexibility allows SVM to capture complex relationships within the data.

However, SVM's performance can be sensitive to the choice of hyperparameters, and it may not perform well on very large datasets. Despite these considerations, SVM remains a popular and widely used algorithm in various applications, such as image classification, text categorization, and bioinformatics. Its ability to handle diverse datasets and capture intricate patterns makes it a valuable tool in the machine learning toolkit.

Data Processing and Feature Extraction

To facilitate classification, we extracted pertinent features from the EEG data. The chosen features encompassed Shannon entropy, average, minimum, maximum, and standard deviation. These features were computed for each epoch, and a comprehensive feature vector was constructed, encapsulating the distinctive aspects of each recording. This method ensured that our classification models, namely the k-nearest-neighbor and support-vector-machine, received rich input vectors reflective of the underlying characteristics of the EEG signals.

Model Training and Evaluation

The classification models underwent rigorous training on the subject-specific datasets, utilizing the curated feature vectors. Post-training, the models were meticulously evaluated to gauge their effectiveness in discriminating between seizure and non-seizure instances. Performance metrics such as accuracy, precision, recall, and F1-score were employed to provide a comprehensive assessment of the model's efficacy.

Results

Subject	Classifier	Metric	Value
1	KNN	Accuracy	75
		Sensitivity	50
		Specificity	100
	SVM	Accuracy	75
		Sensitivity	50
		Specificity	100
		s_ws_start_points	[2296, 1467, 1732, 1015, 1720, 1862]
3	KNN	Accuracy	100
		Sensitivity	100
		Specificity	100
	SVM	Accuracy	75
		Sensitivity	100
		Specificity	50
		s_ws_start_points	[731, 2162, 1982, 2592, 1725]
8	KNN	Accuracy	100
		Sensitivity	100
		Specificity	100
	SVM	Accuracy	100
		Sensitivity	100
		Specificity	100
		s_ws_start_points	[2670, 2856, 2988, 2417, 2083]
14	KNN	Accuracy	100
		Sensitivity	100
		Specificity	100
	SVM	Accuracy	100
		Sensitivity	100
		Specificity	100
		s_ws_start_points	[1986, 1911, 1838, 3239, 1039, 2833]
16	KNN	Accuracy	75
		Sensitivity	100
		Specificity	50
	SVM	Accuracy	75
		Sensitivity	100
		Specificity	50
		s_ws_start_points	[2290, 1120, 1854, 1214]

Table 1. Summary of Classification Metrics and Start Points for Different Subjects

Overall, the performance varies across subjects, with some achieving perfect accuracy and sensitivity. Subjects 8, 14, and the KNN classifier for Subject 3 stand out with consistently high performance across all metrics. In contrast, Subject 1 and Subject 16 show lower accuracy and sensitivity for both classifiers. The choice of classifier also seems to impact performance, with SVM generally achieving higher metrics compared to KNN.

The outcomes of our experimentations are meticulously tabulated, delineating the classification results for each subject separately. These tables elucidate the models' performance metrics, offering a granular insight into the effectiveness of the k-nearest-neighbor and support-vector-machine classifiers in discerning seizure occurrences within EEG signals.

Individualized Subject Analysis

Recognizing the subject-specific nature of EEG signals, we opted for an approach wherein each subject's recordings were classified independently. This subject-centric strategy allowed us to tailor our models to the unique characteristics of each individual, enhancing the sensitivity and specificity of our seizure detection system.

Conclusion

Our analysis underscores the subject-specific nuances in seizure detection, with varying degrees of performance across subjects. We delve into the implications of our findings, discussing the strengths and limitations of our chosen features and classifiers. Additionally, we explore potential avenues for further refinement and improvement in the model's predictive capabilities.

In conclusion, our research endeavors to contribute to the field of seizure detection by employing a subject-centric approach with carefully curated datasets and feature extraction techniques. The individualized analysis of subjects, coupled with the application of sophisticated classification models, showcases promising results. However, ongoing work is essential to enhance the model's robustness and generalize its applicability across diverse EEG datasets.