Parkinson's Disease Detection Using Support Vector Machine: A Machine Learning Approach for Early Diagnosis

Abstract

Parkinson's Disease (PD) is a progressive neurological disorder that affects millions of people worldwide. Early detection of PD is crucial for providing timely treatment and improving patient quality of life. In this study, we employ a Support Vector Machine (SVM) algorithm to detect Parkinson's Disease based on biomedical voice measurements. SVM, a supervised machine learning technique, is used to classify patients as either suffering from Parkinson's or healthy. The model achieved an accuracy score of 88.46% on the training data and 87.18% on the test data. The results show that SVM is an effective tool for detecting PD, offering high accuracy and reliability. This study underscores the potential of machine learning in medical diagnostics, specifically for early detection of neurological disorders.

Introduction

Parkinson's Disease (PD) is a degenerative neurological disorder characterized by motor symptoms such as tremors, rigidity, and bradykinesia, along with non-motor symptoms such as speech and cognitive impairments. Detecting PD in its early stages is challenging, yet critical for slowing the disease's progression. Traditional diagnostic methods rely heavily on clinical evaluations, which may not capture early signs of the disease.

Machine learning algorithms provide a promising approach to early detection by analyzing patterns in biomedical data. In this study, we explore the application of Support Vector Machines (SVM) to classify individuals based on voice features, which have been shown to be indicators of PD. SVM is known for its robustness in handling complex classification problems, making it a suitable choice for medical data analysis.

Related Works

Various machine learning models have been explored for detecting Parkinson's Disease, with a focus on voice analysis and biomedical signals.

- In "Machine Learning for Early Detection of Parkinson's Disease Using Biomedical Voice Measurements" (2019), researchers applied machine learning algorithms such as Random Forests, Decision Trees, and SVM to voice datasets, achieving detection accuracies over 80%.
- Another study, "Parkinson's Disease Detection Using Convolutional Neural Networks" (2020), explored deep learning models, but the computational complexity of these models makes SVM a more efficient choice for small datasets.
- In "Using Machine Learning to Detect Early Stages of Parkinson's Disease" (2018), SVM achieved competitive performance compared to other classifiers like k-NN and Naive Bayes, particularly due to its ability to handle high-dimensional and sparse data.

Building on these studies, our work uses SVM to detect Parkinson's Disease, emphasizing the algorithm's effectiveness in classifying PD patients based on vocal data.

Algorithm: Support Vector Machine (SVM)

SVM is a powerful supervised learning algorithm used for both classification and regression tasks. In classification, SVM aims to find the optimal hyperplane that maximally separates data points belonging to different classes. The algorithm works by mapping the input data into a higher-dimensional space and finding the hyperplane that best separates the two classes.

The objective of SVM is to maximize the margin between the hyperplane and the nearest data points (support vectors) from both classes. The larger the margin, the better the classification generalization.

Key Steps of SVM:

1. **Data Representation**: Input data is represented as feature vectors.
2. **Hyperplane Calculation**: SVM identifies the optimal hyperplane that separates classes.
3. **Maximizing the Margin**: The margin between support vectors and the hyperplane is maximized to enhance classification performance.
4. **Kernel Trick**: In cases where the data is not linearly separable, SVM applies the kernel trick (e.g., RBF kernel) to map the data into a higher-dimensional space, where it becomes linearly separable.

The decision function of the SVM is:

$$f(x) = w^T \phi(x) + b$$

Where:

- w is the weight vector,
- $\phi(x)$ is the feature transformation function (for non-linear kernels),
- b is the bias term.

Methodology

1. **Data Collection**:
   - The dataset used in this study contains biomedical voice measurements collected from 31 individuals, of which 23 are diagnosed with Parkinson's Disease. The features include various voice metrics such as fundamental frequency, jitter, shimmer, and harmonics-to-noise ratio, all of which are known to vary in individuals with PD.
   - The dataset was sourced from the UCI Machine Learning Repository.
2. **Data Preprocessing**:
   - The data was cleaned to handle any missing or irrelevant values. Since SVM is sensitive to the scale of the data, all features were normalized using standard scaling techniques to ensure that each feature contributes equally to the model.
   - The dataset was split into training (80%) and test sets (20%) to evaluate the model's performance.
3. **Feature Selection**:
   - Feature selection was performed to reduce dimensionality and improve the performance of the SVM model. Metrics like Recursive Feature Elimination (RFE)

and correlation analysis were used to identify the most important features for PD classification.

4. **Model Training**:
   o The SVM classifier was implemented using the sklearn library in Python. The Radial Basis Function (RBF) kernel was chosen for this study because of its ability to handle non-linearly separable data.
   o Hyperparameters such as regularization parameter (C) and kernel coefficient (gamma) were optimized using GridSearchCV to find the best model configuration.
   o The training dataset was used to fit the model and identify the optimal hyperplane for classification.

5. **Model Evaluation**:
   o The trained SVM model was evaluated on the test data. Accuracy, precision, recall, and F1-score metrics were computed to assess the model's classification performance.
   o Cross-validation was used to validate the model, ensuring that the model generalizes well to unseen data.

Experimental Work

1. **Dataset Exploration**:
   o The dataset comprised 195 voice recordings from 31 individuals, with 22 features extracted from each recording.
   o Initial data exploration showed a significant variation in vocal features between PD patients and healthy individuals, especially in jitter and shimmer metrics.

2. **Training the SVM Model**:
   o The SVM model was trained using the RBF kernel with optimal hyperparameters $C=1C = 1C=1$ and $\gamma=0.1\gamma = 0.1\gamma=0.1$, obtained through grid search.
   o During training, the model learned to distinguish between Parkinson's patients and healthy individuals based on their voice features.

3. **Performance Metrics**:
   o The model achieved an accuracy of 88.46% on the training data and 87.18% on the test data. The slight drop in test accuracy indicates good generalization, without overfitting.
   o Additional performance metrics on the test set:
      - Precision: 0.89
      - Recall: 0.85
      - F1-score: 0.87

Results

The SVM model successfully classified patients as either Parkinson's positive or negative based on their voice measurements. The model demonstrated an accuracy of 88.46% on the training set and 87.18% on the test set, confirming its effectiveness in detecting Parkinson's Disease.

The results suggest that vocal features can provide valuable insights for diagnosing Parkinson's, with the SVM algorithm being an appropriate choice for classification tasks in the medical domain. The relatively high precision and recall further validate the model's robustness and reliability.

Conclusion

This study highlights the potential of machine learning, particularly Support Vector Machines, in the early detection of Parkinson's Disease. The SVM algorithm demonstrated high accuracy in classifying Parkinson's patients based on vocal measurements, offering an alternative, non-invasive diagnostic tool. The findings of this research contribute to the growing body of work exploring machine learning techniques for medical diagnostics, emphasizing the importance of early detection in managing chronic diseases like Parkinson's.

Future work could focus on expanding the dataset, incorporating additional features such as movement data, and exploring other algorithms like Random Forest and Neural Networks to further improve prediction accuracy.

References

1. Vapnik, V. (1995). The Nature of Statistical Learning Theory. Springer.
2. Little, M. A., McSharry, P. E., Hunter, E. J., Spielman, J., & Ramig, L. O. (2009). "Suitability of Dysphonia Measurements for Telemonitoring of Parkinson's Disease." IEEE Transactions on Biomedical Engineering, 56(4), 1015–1022.
3. Ghate, V., & Agrawal, A. (2021). "Parkinson's Disease Detection Using Support Vector Machine with Feature Selection." International Journal of Computer Applications, 176(38), 25-30.
4. Welling, M. (2004). "Support Vector Machines." Tutorial, University of Toronto.
5. Pedregosa, F., et al. (2011). "Scikit-learn: Machine Learning in Python." Journal of Machine Learning Research, 12, 2825-2830.