Applied Machine Learning COMS W4995 - Topics in Computer Science
Group 23 Final Project
Exploring Activity Monitoring Through Wearable Technology Data and Machine Learning

Introduction

The exploration of human physiological responses to various activities has become a pivotal area of research, contributing to our understanding of health, well-being, and performance. This report explores the PAMAP2 Dataset, which serves as a valuable resource, offering rich insights into the interplay between physical activities and physiological parameters. The dataset provides a unique opportunity to delve into the intricacies of physiological responses, such as heart rate and body temperature, during activities ranging from sedentary tasks like watching TV and computer work to more dynamic activities like walking, running, and playing sports. Through the application of machine learning and deep learning models, we utilize these data to make predictions on these tasks. This can be utilized in various contexts and environments as we move into a more fitness and health-focused lifestyle using wearable technologies. The ensuing sections will detail the methodology for data preprocessing, model selection, and evaluation, culminating in a discussion on the three models selected to predict physiological responses recorded during the diverse set of activities performed by the participants.

Data

This dataset, collected through the use of wireless Inertial Measurement Units and heart rate monitoring, encompasses 9 diverse participants in 13 distinct activities. The participants underwent a comprehensive protocol, involving routine and optional activities, resulting in a diverse dataset.

Methods

Initial Data Exploration: In the initial exploration of the dataset, it is evident that we are working with a substantial amount of data, comprising a total of 2,864,056 entries distributed across 33 columns. The entries are uniformly spaced at intervals of 0.01 seconds, reflecting the high temporal resolution of the dataset. The primary objective is to predict the 'activityID,' a categorical variable encompassing 13 distinct values, indicating the various activities performed. Since 'activityID' is in categorical values, we need to encode this variable. Furthermore, it is observed that the data samples are represented as float64, indicating that the dataset predominantly consists of numerical values with a high precision level. This initial exploration provides essential insights into the target variable, and overall structure of the data, laying the foundation for subsequent in-depth analyses and model development.

Data Cleaning: From initial data exploration, it is observed that only a minuscule portion pertains to missing heart rate data. Given the small proportion of missing values in this specific variable, a decision has been made to drop the corresponding samples, ensuring minimal impact on the overall dataset. Additionally, we plotted the variable feature correlation and determined no significant collinearity among the variables. This finding suggests there is no compelling need to discard any features due to high correlation.

Data Sampling: The histogram depicting the counts of each ActivityID label provides a clear visual representation of the distribution of activities within the dataset. Notably, 'transient activities' emerge as the most frequently occurring label. In contrast, the remaining categories exhibit comparable frequencies, suggesting a relatively balanced distribution among most activities. Recognizing the imbalanced nature of the 'transient activities' category, under-sample the 'transient activities' category to harmonize the class distribution. This step aims to mitigate potential biases introduced by the imbalanced data and foster a more equitable representation of each activity, thus enhancing the model's ability to generalize across diverse activity types.

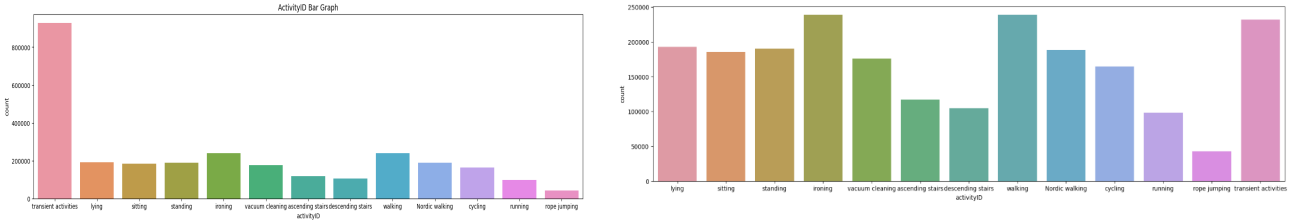

Fig 1. Distribution of Activity ID labels before and after undersampling was implemented

Result

Model Selections: We have chosen K-Nearest Neighbors and a Sequential Neural Network.

KNN

- **Model Training:** We utilized a k-Nearest Neighbors (kNN) algorithm for a range of neighbors from 4 to 10. In each iteration, a new kNN classifier is instantiated and trained on the provided training data. Subsequently, predictions are made on the test data, and the accuracy of these predictions is computed using the scikit-learn accuracy_score function. The calculated accuracy for each iteration, corresponding to different values of k, is stored in the accuracy_scores list. This systematic exploration of varying k values allows for an empirical assessment of the kNN classifier's performance under different neighborhood configurations, aiding in the selection of an optimal parameter for achieving the most accurate and reliable predictions on the given dataset.
- **Model Evaluation:** The provided accuracy scores represent the performance of the k-Nearest Neighbors (kNN) classifier across a range of values for the number of neighbors (k). The accuracy scores for each iteration are as follows: [0.957, 0.953, 0.949, 0.946, 0.942, 0.940, 0.937]. Analyzing the results, it is apparent that the model achieves high accuracy across the range of k values, with accuracy scores consistently above 93%. As the number of neighbors increases, the accuracy slightly decreases, reflecting the potential impact of a more generalized decision boundary.

Neural Network

- **Model Training:** We utilized a neural network model with sequential architecture designed for multi-class classification tasks, particularly suited for scenarios with 13 distinct output classes. Comprising a series of densely connected layers, the model begins with an input layer of 128 units, each employing the rectified linear unit (ReLU) activation function. To prevent overfitting, dropout layers with a rate of 0.5 are introduced after each dense layer. During training, the Adam optimizer is employed to adapt learning rates, while the sparse categorical cross-entropy loss function ensures effective

training for the multi-class classification problem. The model's performance is evaluated using accuracy as a metric.

- **Model Evaluation:** The validation loss for the Sequential Neural Network was 0.629, with a corresponding validation accuracy of 79.7%. In the testing phase, the model exhibited a test loss of 0.685, paired with a test accuracy of 79.7%.

Conclusion

In conclusion, this project delved into the exploration of human physiological responses through an in-depth analysis of the PAMAP2 Dataset. Two distinct models, K-Nearest Neighbors (KNN) and a Sequential Neural Network, were employed for activity prediction. The KNN model demonstrated remarkable accuracy, consistently exceeding 93% across various values of neighbors (k). On the other hand, the Sequential Neural Network exhibited a commendable validation accuracy of 79.7%, showcasing its ability to discern complex patterns in the physiological data.

The conclusion highlights the effectiveness of both models, with the KNN excelling in simplicity and transparency, providing high accuracy with a straightforward algorithm. Meanwhile, the Sequential Neural Network, although more complex, demonstrated competitive accuracy and was well-suited for capturing intricate patterns in the dataset. Overall, leveraging machine learning and deep learning models, we aimed to predict activity labels, facilitating applications in health monitoring and fitness tracking through wearable technologies.

Discussion

During the project's experimentation phase, we explored three tree ensemble models—pruned decision tree, random forest, and XGBoost—for time series classification, fine-tuning their hyperparameters through grid search. However, computational challenges arose, particularly during the grid search, primarily due to the size of the dataset, which was nearly 3 million rows. The extended training times became apparent, with a single model exceeding 300 minutes and occasionally failing to converge within the specified timeframe. This prolonged training can be attributed to the combined impact of exhaustive grid search and the inherent complexity of ensemble models like random forest and XGBoost. These models, designed for enhanced predictive accuracy through tree aggregation, exhibit heightened computational demands, particularly on larger datasets such as the one used in this project. With more time, we would have also preferred to train a Long Short-Term Memory Recurrent Neural Network (LSTM RNN), as they are better suited for time-series forecasting.

Reference

PAMAP2 Physical Activity Monitoring Dataset
https://archive.ics.uci.edu/dataset/231/pamap2+physical+activity+monitoring