西安交通大學

XI'AN JIAOTONG UNIVERSITY

Intelligent decision-making and machine learning Ⅱ

--Research on a wearable device user behavior prediction model based on machine learning

Project：	Intelligent Decision and Machine Learning Ⅱ
Name：	Yuze Shi
Class：	Big Data Management and Application91
Instructor:	Shao-bo Lin
Student ID:	2196113772
Date:	July 4, 2022

Contents

1 Introduction

With the advancement of optical technology, vital sign monitoring functions such as heart rate measurement, photoplethysmography (PPG) and electrocardiogram (ECG) have become popular in smart wearable devices, making it easier for individual users or health professionals to obtain individual vital signs. The data of physical signs can timely and accurately monitor physical health status or assess the possibility of disease onset. Optical solutions have been widely used in wearable devices for health monitoring. The principle is to irradiate the skin, tissues and blood vessels through the light source (currently mainly LED), and due to the different absorption of light by each part, the change of reflected light can be received and measured, and the corresponding algorithm can be used to evaluate the health of the human body. Light is absorbed by hemoglobin in the blood, and based on the absorption rate, the sensor can calculate the pulse rate and oxygen saturation. Green light is mainly used to measure the pulse because it is most easily absorbed by red blood cells. Infrared light is used in places where it is easy to measure the pulse, such as the earlobe, and it also works with red light to measure oxygen saturation. Thanks to the rapid development of LED technology, the breakthrough achieved a green light efficiency improvement of 40%, coupled with the progress in component miniaturization and thermal stability, wearable devices can already achieve higher precision life Sign monitoring, helping users to continuously collect data in daily life, and issue alarms when abnormal.

Classification is an important data mining technique. With the rapid development of the Internet of Things, mobile medical has entered the public's field of vision. The emergence of smart wearable devices has made it relatively easy to obtain human physiological signals, thus promoting the development of mobile medicine. Among them, the relevant mobile medical treatment based on ECG monitoring, the theoretical basis of its system is the classification and recognition of biometric signals. With the development of machine learning in recent years, researchers have begun to use different models and algorithms to classify and identify the activities of objects wearing smart wearable devices.

Based on the heart rate measurement, photoplethysmography (PPG), electrocardiogram (ECG) and other vital sign data provided by the PPG-DaLiA dataset, this paper obtains a sample database that can be used for training and learning through data form transformation and feature extraction. The classification algorithms based on softmax logistic regression, Gaussian kernel-based SVM, decision tree and random forest are designed and implemented, and the classification performance of different algorithms is compared and analyzed from the aspects of training speed, stability and accuracy. Continuous and real-time monitoring of biosignals is critical for better management of patients with chronic diseases, including cardiovascular disease, diabetes, and neurological disorders, and smart devices are showing encouraging signs in healthcare due to their flexibility and compliance result. These wearable devices provide a better understanding of changes within the human body and can help prevent and treat disease. In addition, in the field of sports competition, real-time monitoring of athletes' activities plays an extremely important role in preventing injuries and improving competition status.

2 Model construction based on single subject data

2.1 Data read

Photoplethysmography (PPG) is now widely used for continuous heart rate monitoring. PPG-based heart rate estimation is mainly used for motion artifacts. The original primary purpose of this dataset was PPG-based heart rate estimation. For this task, three sensor modalities are commonly used: a) the PPG sensor itself, b) the 3D accelerometer device embedded in it as a PPG sensor to compensate for motion artifacts, and c) to provide ground truth for ECG evaluation of the heart. A common practice in related work is to use the sliding window method (window length: 8 seconds,

window movement: 2 seconds). This means that all data signals are segmented into this sliding window, and the goal is to determine the heart rate for each 8-second window segment. Fifteen subjects participated in the study, seven males and eight females, aged between 20 and 40. The dataset also provides information on subjects' gender, height, weight, skin type, fitness level, and more.

All the above data are stored in the file SX.pkl. These files store data in dictionary form with the following keys:
1)'activity': includes the activity labels, providing IDs 0…8.
2)'label': includes the ground truth heart rate information. As described above, this is provided as the mean of the ECG-based instantaneous heart rate, given on a sliding window of 8 seconds, shifted with 2 seconds.
3)'questionnaire': includes information about the subject.
4)'rpeaks': the index of the identified and corrected R-peaks, referring to the ECG-signal. the identified R-peaks provide the basis of the heart rate ground truth.
5)'signal': includes all the synchronised raw data, in two fields:
'chest': RespiBAN data (all the modalities: ACC, ECG, EDA, EMG, RESP, TEMP)
'wrist': Empatica E4 data (all the modalities: ACC, BVP, EDA, TEMP)
6)'subject': the current subject's ID
After parsing the pkl file with the parser, extract the value corresponding to each key and build a data feature table.

2.2 Preprocessing and feature engineering

In order to simplify the data processing and the difficulty of modeling, the subject numbered S1 was first selected to start the study.

As mentioned before the features are stored in different frequencies. The first step was to aggregate the data and put them in the same frequency. I chose to put it all to 4Hz as it is the frequency used to record the activity (our target variable).

No feature engineering was needed except for the R-Peaks. Since the indexes were given, I chose to count the number of R-Peaks that occurred during our time period (4Hz so 0.25 seconds).

2.2.1 Preprocessing of Activity

The data set contains eight kinds of subject activities, which correspond to ID1-8 in the data set, which are:Sitting (ID: 1) Ascending and descending stairs (ID: 2) Table soccer (ID: 3) Cycling (ID: 4) Driving a car (ID: 5) Lunch break (ID: 6) Walking (ID: 7) Working (ID: 8). However, in order to make further processing of the dataset more convenient, an activity-signal with 4 Hz sampling rate (which is the lowest sampling rate across all recorded raw sensor data) was created.

It can be seen from Figure 2-1 that the most activity of subject S1 is 'Lunch', accounting for 25.56% of all activities. The activities that account for more than 10% are also unrecorded (24.76%) and 'Working' (12.91%). The frequency of occurrence of the other activities is relatively similar. Because Activity is a label as a model, you need to pay attention to the class imbalance problem. In all activities, there are no classes that exceed the overall $\frac{1}{3}$, and there is no imbalance problem.

2.2.2 Preprocessing of Label

As described above, this is provided as the mean of the ECG-based instantaneous heart rate, given on a sliding window of 8 seconds, shifted with 2 seconds. Because the motion detection window length is 8 seconds, the original data has 4603 rows. To fit the overall data frame, each data was repeated 8 times. If the size of the label is smaller than the activity, just take the average

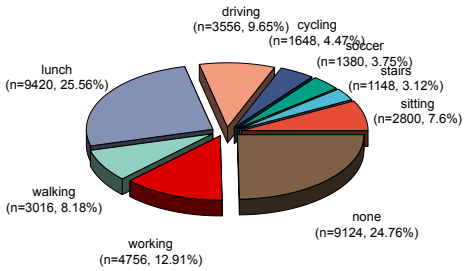

Figure 2.1: pie plot of Activity

value of the label and add it to the data. Based on the heart rate information, a heart rate time series graph can be drawn.

2.2.3 Preprocessing of Signal

Raw sensor data was recorded with two devices: a chest-worn device (RespiBAN Professional) and a wrist-worn device (Empatica E4).

The RespiBAN Professional was used , with the following sensor modalities: ECG, respiration, and three-axis accelerometer. ECG-signal was acquired via a standard three-point ECG. Respiration signal was acquired with an inductive respiration sensor, which is embedded into the RespiBAN chest strap. Three-axis acceleration was acquired via a 3D-accelerometer, which is integrated into the RespiBAN wearable device.Data is organised in a dictionary, corresponding to the sensor modalities.

A triaxial accelerometer records acceleration in three vertical axes (x, y, z) (medial, lateral, front-to-back, vertical) to generate data. By measuring the frequency and magnitude of these movements (that is, how often and how much they occur), the accelerometer can calculate the total gravity to which the wearer is exposed, known as the "composite vector magnitude." Triaxial accelerometers have been used in the past to detect common human activities such as sleeping, fidgeting, lying down, walking, running, and jumping. This technology is already embedded in regular smartphones. Combined with the data in the dictionary, the ACC signal is separated into three channels. Because the ACC signals are collected by the chest and wrist sensors, respectively, there are 6 features after splitting:wrist-ACC-x,wrist-ACC-y,wrist-ACC-z,chest-ACC-x,chest-ACC-y,chest-ACC-z.

The Empatica E4 device was used . The E4 was worn on the subjects' non-dominant wrist. The modalities 'EDA', 'EMG' and 'Temp' include dummy data and should be discarded.

The data distribution density of the six channels is shown in Figure 2.2. It is found that all data distributions are relatively concentrated. The peak density of chest-ACC-y exceeds 7.5, the peak value of chest-ACC-x is about 5.0, and the density of other channels is about 5.0. The absolute value of the peak density is less than 2.5.

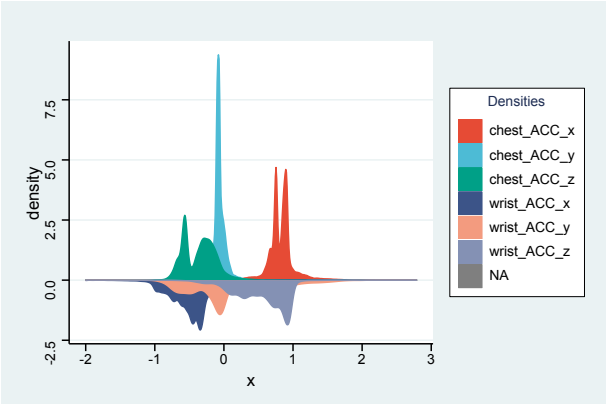

Figure 2.2: density plot of ACC

2.2.4 Preprocessing of Rpeaks

We will cound for each 175 portion (0.25 sec since 700 is 1sec) the number of rpeaks during that period.

Rpeak appears after the time portion The rpeaks will probably end before the time portion so we need to fill the last portions with 0. Finally treat the 'Rpeaks' feature column as a 0-1 variable.

2.3 EDA and Visualization

Since the current modeling analysis is based on the information of a single subject, the characteristics related to the subject itself, such as gender, height, age, etc., are deleted. After completing the above feature construction, data segmentation and combination, data cleaning, feature dummy quantification and a series of feature engineering, the data feature table is constructed. The dimension of the feature table is 36848*12, that is, there are 12 features and 36848 pieces of data. checked The data table has no missing values and outliers, and can be modeled. Summary information for all features is shown in Table 2.1. All features are continuous variables except Rpeaks and data label Activity.

Table 2.1: Feature Summary Table

Feature	Comment	Data Type
ACC	chest:xchest,ychest,zchest wrist:xwrist,ywrist,zwrist	continuous variable
Resp	signal on chest	continuous variable
ECG	signal on chest	continuous variable
BVP	signal on wrist	continuous variable
TEMP	signal on wrist	continuous variable
Rpeaks	count for each 175 portion the number of rpeak	0-1 variable
Label	the ECG-based instantaneous heart rate	continuous variable

Next, a correlation analysis is performed for the numerical variables, and a heat map is drawn by calculating the Pearson correlation coefficient. As can be seen from Figure 2.3, the absolute values of the correlation coefficients between the data features are all lower than 0.5. Among them, the correlation between Label and chest-ACC-y is 0.42, and the correlation with chest-ACC-z is 0.33. From this, it can be concluded that the overall correlation between features is not high, and no feature selection is required for features, which will not affect the performance of later

models. Before building a model to predict subject activity, it is also necessary to observe how

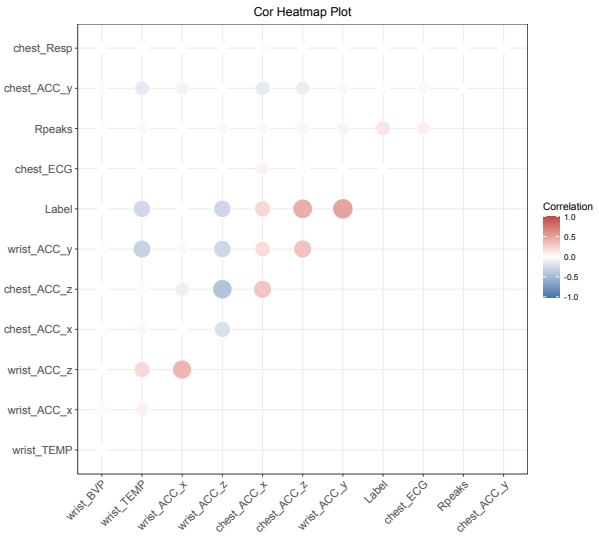

Figure 2.3: heatmap

different activity types change with characteristics under temporal trends. According to experience, activities can be divided into two kinds of vigorous activities and non-vigorous activities. The heart rate of subjects under vigorous activity conditions was significantly higher than that under non-vigorous activity conditions. Figures 2.4 are temperature trends over time for different activities. Figure 2.5 shows the trend of heart rate over time during different activities. The red area represents the time frame when the subject is in the experimental collection activity, and the white area represents the time frame when the subject is in the non-experimental collection activity. It can be seen that during lunch, working, walking and other activities, the body temperature of the subjects is higher than the average. From the single feature of body temperature, the activities can be roughly divided into two categories. When performing activities such as stairs, cycling, walking, etc., the heart rate of the subjects all exceeded 120, that is, they were engaged in vigorous activities.

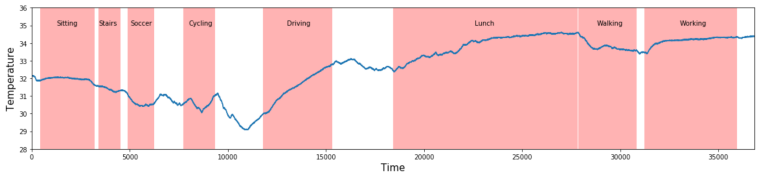

Figure 2.4: Plot of the temperature during the different activities

2.4 Activity prediction model for support vector machines

Firstly, the basic principle of SVM is expounded from three aspects: hypothesis space, optimization strategy, and learning algorithm.

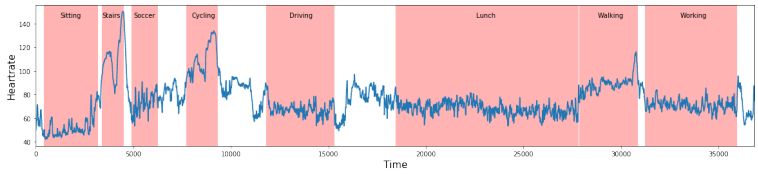

Figure 2.5: Plot of the heart-rate during the different activities

2.4.1 Hypothesis space

Hypothesis space is our a priori assumption about the model form, and the final model we obtain must conform to our model. A priori hypothesis of the form. The hypothesis space of the support vector machine model is the Hilbert space. Hilbert space is a generalization of Euclidean space, and a complete inner product space is called Hilbert space, The inner product space is a normed linear space that defines the inner product.

2.4.2 Optimization strategy

The optimization strategy adopted by the support vector machine is the maximum interval strategy. The farther the data is from the classification surface, the more accurate the prediction. Therefore, when designing the classification surface, it needs to be kept away from the data. The core mathematical model applied when training a binary classifier between two classes of samples i, j:

$$\begin{array}{ll} \min & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{k}\sigma_i \\ \text{s.t.} & y_i(w\varphi(x)+b) \geqslant 1-\sigma_i \\ \text{s.t.} & \sigma_i \geqslant 0 \quad i=1,2,\cdots,k \end{array} \tag{2.1}$$

In the formula: w and b are the parameters to determine the hyperplane; $y_i$represents the category label; $\phi(x)$ represents the kernel function; C represents the penalty factor, which is used to control the loss caused by wrong samples; i represents the relaxation factor, which can be There is a certain range of flexibility in the classification boundaries. Equation 2.1 is transformed by maximizing the geometric distance between the sample point and the hyperplane. Accordingly, the classification problem is transformed into, in the case of allowing a certain error, to solve the optimal w and b, and then determine the optimal classification hyperplane. Due to the complexity of solving the above problem, the Lagrange multiplier is used to convert it into a dual problem, and then the optimal classification function between the two categories i and j is obtained.

2.4.3 Optimization algorithm

A learning algorithm for efficiently solving SVMs is the SMO algorithm. The idea of the SMO algorithm is similar to that of the coordinate ascent algorithm. The coordinate ascent algorithm achieves the purpose of optimizing the function by updating one dimension in the multivariate function each time, and after multiple iterations until convergence , we need to optimize a series of α values, and iterate continuously until the function converges to the optimal value.

SVM was originally proposed as an effective method to solve the binary classification problem, and it cannot directly solve the multi-classification problem. In this study, activity detection with 8 categories is a multi-classification problem. How to use SVM to extend the binary classification method to the multi-classification method has become one of the important research contents of this paper. There are two ideas for SVM multi-classification processing: the first is to solve multi-classification problems at one time, apply the two-classification idea of SVM to multi-classification, optimize the classification function, and build a multi-value classification model, so

that the model has direct Ability to handle multiple classifications. The second is to decompose the multi-classification problem into multiple binary classification problems, and construct multiple binary classifiers to form a multi-classifier. This paper adopts the second idea, which is to construct a one-to-one method of multiple binary classifiers to solve the multi-classification problem.

2.4.4 Model building

The SVM one-to-one multi-classification method was first proposed by Knerr, and it is also the most commonly used multi-classification method. For the classification problem in this paper, the main idea of establishing a detection model is: to build an SVM binary classifier between any two categories, a total of 28 classifiers are needed; to generate each binary classifier, only one of the two types of data in the training data is needed. , and then the combined 28 binary classifiers can solve the multi-classification problem.

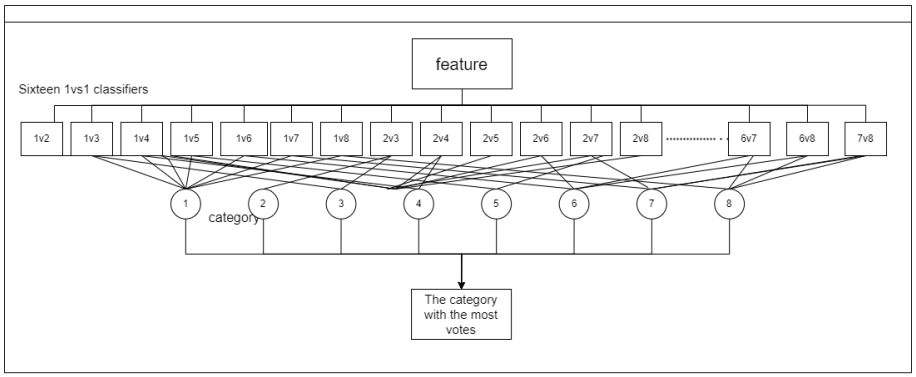

Figure 2.6: SVM multi-classification strategy diagram

After the classifier is constructed, it enters the classification and voting decision-making process. Use all binary classifiers to discriminate the test set sample x_i in turn. If $f_{ij}(x_i) > 0$, the positive example will get one vote, and if $f_{ij}(x_i) < 0$,the negative example will get one vote. After the traversal is completed, the category to which the sample x_i to be tested belongs is judged according to the votes obtained. Figure 2 shows the discriminative process of the five-category SVM of arrhythmia signals. The line between the classifier and the class indicates the class that the sample may be judged for. The category with the most votes is finally output.

Because the data samples used in this paper are non-linearly separable samples This, so it is necessary to use the kernel function to achieve from low-dimensional to high-dimensional space mapping between. The choice of SVM kernel function is important for its detection of arrhythmia constant performance plays a crucial role, applying different Kernel function, the detection effect is very different. Commonly used kernel functions are: Linear kernel function, Gaussian kernel function, polynomial kernel function, sigmoid kernel function. In correlation analysis of multi-classification problems, quasi- Accuracy (ACC), specificity (SPE) and sensitivity (SEN) were evaluated as The index of the performance of the price model is calculated as follows:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{2.2}$$

$$\mathrm{SPE} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{2.3}$$

$$\mathrm{SEN} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2.4}$$

where: TP stands for true positive; TN stands for true negative; FP stands for false Positive; FN stands for false negative. ACC is a correctly classified heart rhythm proportion. SPE is the proportion of normal heart rhythms that are correctly classified example. SEN is the proportion of abnormal heart rhythms that are correctly classified, balancing The ability of the quantitative model to identify cases. How to choose the right verification letter for specific domain data now The number is still inconclusive, so this paper will test the four commonly used kernel functions.

The experimental results are shown in Table 1. Choose the Gaussian kernel function, The detection accuracy, specificity and sensitivity on the test set are 85.28%, 85.08% and 86.22%, the detection effect is the best. Figure 3 is SVM multi-class confusion matrix, from which it can be seen that only category V is classified The correct rate is high, and many samples in the S and Q categories are misjudged as N class, the reason for this result is that the ECG data passes through the Gaussian kernel After the high-dimensional mapping of functions, there are inevitably many This linear inseparability phenomenon results in unsatisfactory detection accuracy.

Table 2.2: Index comparison of different kernel functions

Kernel function	ACC/%	SPE/%	SEN/%
Linear Kernel	73.08	74.34	67.03
Gaussian kernel	85.28	85.08	86.22
polynomial kernel	77.44	76.32	82.81
Sigmoid kernel	72	73.23	66.08

2.5 Activity prediction models for decision trees and random forests

2.5.1 Hypothesis space

The prior form of the decision tree model can be expressed as follows:

$$\hat{\mathbf{y}} = \mathbf{f}[\mathbf{x}] = \mathrm{w}_{\mathrm{q[x]}}$$
$$q := \left\{ R^d \to \{1, 2, 3, \ldots, T\} \right\}$$

where q[x] is a function that maps from the feature space to the node number space. The key to the decision tree model is to divide the feature space into disjoint sub-regions, and samples that fall in the same sub-region have the same predicted value. In order to determine the complete structure of a decision tree, it is necessary to clarify the following two aspects: one is how to divide the sub-region, and the other is how much the predicted value of the sub-region is.

2.5.2 Optimization strategy

The objective function is what standard we use to evaluate the quality of a model. The objective function determines our preference for choosing a model from the hypothesis space.

$$J[f] = L[f] + \Omega[f] = \sum_{i=1}^{n} 1\left[f\left[x_i\right], y_i\right] + \alpha T$$

The objective function of a decision tree can be used to evaluate the quality of a decision tree. This objective function should include two aspects. The first is a loss term that reflects how well the decision tree fits the sample data points, and the second is a regularization term that reflects the complexity of the decision tree model.

The regularization term can take the number of leaf nodes of the model. That is, the more disjoint sub-regions the decision tree model divides, the more complex the model is considered.

For the loss item, if it is a regression problem, the loss item can take the squared loss, and if it is a classification problem, we can use the impurity as a measure. Since all samples on the same leaf node of the decision tree take the same predicted value, if the true label of these samples has only one value, then the sample on this leaf node is very "pure", and we can directly specify the predicted value. For the value of label on this leaf node, the prediction error is 0. On the contrary, if the label values of different samples on the leaf nodes are very messy, it is difficult to adjust the so-called consensus, then no matter how we specify the predicted value on the leaf node, there will always be a large prediction error.

There are generally three methods to measure impurity, namely information entropy, Gini impurity, and classification error rate.

2.5.3 Optimization algorithm

The optimization algorithm refers to how to adjust the value of our model structure or model hyperparameters so that the value of the objective function of the model is continuously reduced. The optimization algorithm determines what steps we use to find a suitable model in the hypothesis space. For decision trees, the optimization algorithm includes tree generation strategy and tree pruning strategy. The tree generation strategy generally adopts the greedy idea to continuously select features to segment the feature space. Tree pruning strategies are generally divided into pre-pruning and post-pruning strategies. Generally speaking, the decision tree generated by the post-pruning strategy has better effect, but its computational cost is also higher.

Figure 2.7 is the decision tree multi-class confusion matrix, from which the categories can be seen 3 The classification accuracy rate is 0.9, 0.08 samples are misclassified as class 0, and the classification accuracy rates of other classes are close to 1. The reason for this result is Due to the serious over-fitting phenomenon of the model, it may be due to the unreasonable use of the pruning strategy, resulting in the low generalization ability of the model.

Figure 2.8 shows the importance order of decision tree features. The top three features are body temperature, chest-ACC-y, and Label. The importance of BVP, ECG, and Rpeaks are all lower than 0.05, which can be deleted and classified.

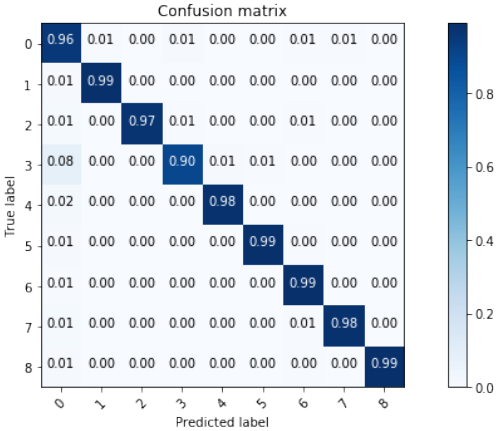

Figure 2.7: Normalized confusion matrix of decision tree

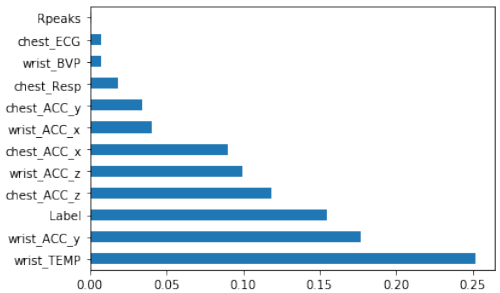

Figure 2.8: feature importance of decision tree

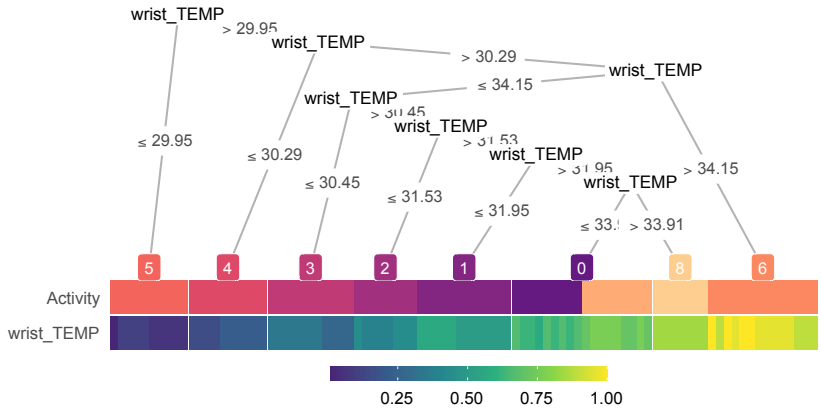

Figure 2.9: visualization of decision tree

2.6 Activity prediction model for the sofmax classifier

2.6.1 Hypothesis space

This section is based on the idea of logistic regression, another machine learning classification algorithm, Build a softmax regression model to classify subject activity. logistic Regression introduces the topic of algorithm optimization and is often used in binary classification, so the selection of class labels The value is usually 0 or 1. Softmax regression is an extension of the former on multi-classification problems. In the classification example in this paper, the dependent variable is the class label, and the independent variable The quantity is the different eigenvalues extracted.Through softmax regression analysis, we can find the best set of parameters, i.e. the weights of the independent variables, to know which eigenvalues are relevant. The key factors of the response signal, so as to achieve the purpose of classification and identification. By minimizing the cost function can implement a suitable softmax regression model.

2.6.2 Optimization strategy

Cost function is an important term often involved in machine learning algorithms. The process of training a model is essentially the process of optimizing the cost function. Usually , all functions that can measure the difference between the predicted value of the model and the true value can be to be called the cost function. If the number of samples is large, the cost function can be the sum of the values is averaged, and the result is recorded as J(θ). For each algorithm, the cost The function is not unique, in logistic regression analysis, the most commonly used is cross entropy (cross entropy) cost function. Assuming the feature vector is x and the class label value is y, The number of training set samples is m, then the cost function of logistic regression is:

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{j=0}^{1} 1\left\{y^{(i)} = j\right\} \log p\left(y^{(i)} = j \mid \boldsymbol{x}^{(i)}; \theta\right)\right]$$

Among them: θ is the parameter that the model needs to train, so that it can minimize the cost function. Since softmax regression is an extension of logistic regression on multi-classification problems, the two The cross-entropy cost function is very similar, the main difference is the cost function of softmax Several values of the class label are accumulated.

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{j=1}^{k} 1\left\{y^{(i)} = j\right\} \log \frac{\mathrm{e}^{\theta_{JT}(i)}}{\sum_{l=1}^{k} \mathrm{e}^{\theta \boldsymbol{T}_x(i)}}\right]$$

Where: $1/\sum_{j=1}^{k} e^{\theta T_x(i)}$ is the normalization of the probability distribution, so that the probability sum of all values is 1 . The problem of minimizing $J(\theta)$ involves optimization theory.

2.6.3 Optimization algorithm

Among the optimization algorithms, various gradient descent algorithms are most commonly used. The direction of the gradient determines the direction of parameter decline in the training process, and the learning rate determines the step size of each step change. With the partial derivatives and learning rate in hand, the parameters can be updated. In order to prevent overfitting, a weight decay term is added to the cost function, which can make $J(\theta)$ a convex function in the strict sense, and ensure that the algorithm can converge to the global optimal solution.

2.7 Algorithm comparison

There are five common criteria for judging machine learning methods: generalization ability, computational cost, stability, interpretability, and privacy protection.

Generalization ability refers to the ability of the learned function to predict unknown data, which is usually evaluated by test error. The generalization ability of SVM and softmax is slightly better than that of decision tree and random forest on low-dimensional data in this paper, but the performance of high-dimensional data is not as good as random forest. In addition, because only one subject is selected for data modeling, decision trees and random forests have serious overfitting.

The computational cost is divided into time cost, space cost and communication cost. The time cost is measured by the time complexity of the algorithm. The space cost refers to the memory space required to run the algorithm. The communication cost refers to the consumption of data transmission. Ranked from fastest to slowest training speed are softmax algorithms, random forests, decision trees, SVMs.

Stability is divided into stability for hyperparameters, samples, algorithms, and environments. The parameter stability of the above four algorithms is very good.

Interpretability is mainly aimed at the interpretability of the model and the interpretability of the features. The interpretability of decision trees is the strongest. It directly gives the ranking of feature importance. Because the softmax algorithm is a generalization of logistic regression, its interpretability is also high. While SVM and random forest are less interpretable.

Privacy protection generally refers to preventing sensitive information from being "reversely" obtained by visitors in certain ways when data is legally accessed, and avoiding leakage and abuse of sensitive information caused by "reverse" deduction of data. The privacy protection capabilities of the above four algorithms are average and need further improvement.

Figure 2.10 is the ROC curve comparison of the four algorithms. The calculated auc values are: RF: 0.958, DT: 0.956, Softmax: 0.68, SVM: 0.569.

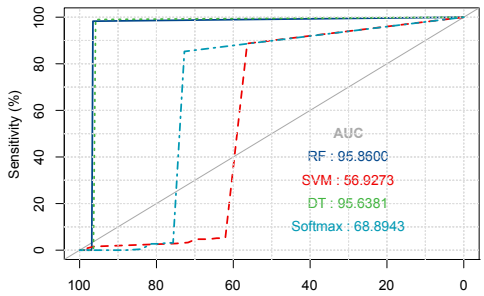

Figure 2.10: ROC

3 Model construction based on all subject data

3.1 Data reading

The data reading and preprocessing method for each remaining subject is the same as that of the first subject, first extracting the values corresponding to each key of the dictionary in each pkl file, and then splicing all subject features into the same table among. The difference is that because all subject data is added to the model training, the characteristics representing different subjects such as gender, weight, height, age, exercise frequency, etc. should not be ignored. The last addition represents each subject number d1-d15.

3.2 Visualization

The dimension of the processed data table is 517956*20, that is, there are 20 features and 517956 pieces of data. After checking that the data table has no missing values, outliers, modeling can be done. The following is a visual analysis of the characteristics of the subjects themselves. A total of 15 subjects participated in the experiment, 7 males and 8 females, with a wide age distribution, including young, middle-aged and old. The 'SPORT' feature indicates how often does the subject do sports; on a scale 1-6 where 1 refers to less than once a month and 6 refers to 5-7 times a week. Figure 3.4 shows the correlation between subject weight and height under different age conditions. It can be found that there is a positive linear relationship between weight and height, and the smaller the age, the more concentrated the distribution.

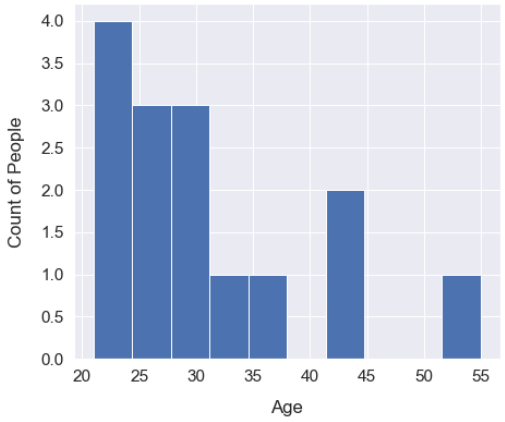

Figure 3.1: AGE

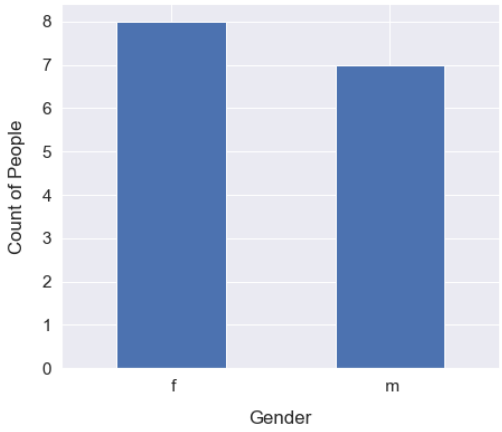

Figure 3.2: Gender

3.3 Model building

In order to facilitate the comparison of the effects of the models, the model used when modeling a single subject is still used. In order to study the performance differences of different algorithms, different evaluation indicators are selected to evaluate: sensitivity (Sensitivity, Sen), positive predictive value (Ppv), specificity degree (Specificity, Spe) and total accuracy (Accuracy, Acc). Acc is used to measure to measure the overall performance of the system; Sen, Ppv and Spe are used to measure the performance of the algorithm for each classification performance. In order to better measure the classification accuracy, this paper uses the F1 index to measure the classification accuracy. numbers are calculated. TP(True Positive), TN(True Negative), FP(False Positive) and

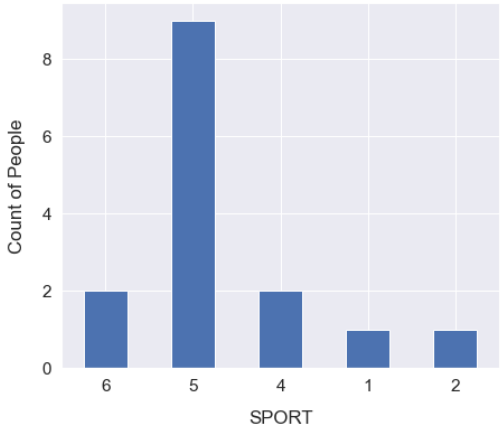

Figure 3.3: SPORT

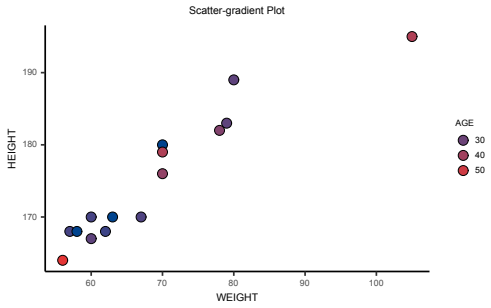

Figure 3.4: scatter plot of height and weight

FN (False Negative) represent the true positives and true negatives belonging to the i-th category, respectively. Calculation of F1 The formula is as follows:

$$F1 = \frac{2 * \text{Sen}_i * Ppv_i}{\text{Sen}_i + Ppv_i} \tag{3.1}$$

3.4 Algorithm comparison

Table 3.1 compares the classification performance of the four algorithms in the categories of Lunch, working, and walking. It can be seen that decision tree and random forest dominate the overall classification accuracy, and the Acc value reaches 97.5% and 96.8%; the Sen values of the decision tree in Lucnh class, working class and walking class respectively reached 99.0%, 79.2% and 25.2%. All algorithms perform poorly on walking, probably due to the small sample size of the walking class.

In general, if only the classification accuracy of the algorithm is considered, it is better to use the random forest or softmax algorithm after reducing the model overfitting problem. The SVM does not perform well on the dataset of this paper, and the training speed is the slowest. If the dataset continues to expand in the later stage, the SVM classification accuracy will be further reduced. Combining interpretability and computational cost, it is more reasonable to select the softmax algorithm if you want to study the importance of features. In addition, the computational complexity of the Softmax algorithm is relatively small, and the cache configuration requirements of the wearable device are not high, which can realize real-time monitoring.

Table 3.1: Comparison of algorithm performance

algorithm	Acc	Lunch			working			walking		
		Sen	Ppv	F1	Sen	Ppv	F1	Sen	Ppv	F1
SVM	0.747	0.864	0.842	0.884	0.792	0.637	0.706	0.252	0.16	0.196
DT	0.975	0.99	0.987	0.968	0.94	0.955	0.947	0.833	0.461	0.594
RF	0.968	0.97	0.994	0.971	0.943	0.962	0.952	0.740	0.685	0.418
Softmax	0.858	0.834	0.902	0.941	0.93	0.982	0.933	0.833	0.851	0.869

3.5 Further research

After modeling and model evaluation based on the information of a single subject and the information of all subjects, it is found that the models have overfitting problems. By fear of our model overfitting since we have very few patients, we will use only 13 patients to train and validate our model. We will then see how the model scores on the 2 other patients.

We can see that using different patients for the test makes the score drop drastically. The model has a hard time generalizing and the previous models surely have overfitted since they've all been tested on the same patients they've been trained. Having only 15 patients, it is very difficult to build a model that will better perform on unseen data.

4 Research conclusions

With the rapid development of science and technology, the application scenarios of smart wearable devices are becoming more and more extensive. The two most widely used fields are health monitoring and disease treatment. Health monitoring is the most basic function of smart wearable devices. Smart wearable devices can be used for real-time monitoring of human health indicators such as body temperature, dynamic ECG, pulse wave, blood pressure, blood oxygen, blood sugar, and sleep status. For example, a remote ECG monitoring system can collect ECG data, analyze the signal quality, and screen for symptoms such as arrhythmias. The wearable breathing sensing system can record the breathing wave data of the human body's chest and abdomen in sleep state, analyze the sleep state, and judge the sleep quality. Chronic disease treatment is a new medical method based on health monitoring. Chronic diseases such as diabetes, hypertension, Parkinson's disease, and heart failure have seriously threatened human health. Many medical device manufacturers and start-ups have deployed in the field of chronic disease treatment, and developed intermittent tremor monitoring systems, artificial pancreas systems, etc. Therapeutic wearable devices are currently mainly researched and developed prototypes, with micro-dose, real-time feedback, and remote early warning as their main features. After technical review and clinical trials, they can be put into the market to help people shorten the diagnosis and treatment process and save medical costs.

In the field of competitive sports, various smart devices have been used to collect physical information of athletes, and real-time monitoring of athletes' health status and potential diseases has been realized. The early sports wristbands did not accurately record the number of steps, but most high-end sports wristbands are now developing in the direction of excellence. Many athletes participating in the Olympic Games and their coaches use the sports wristbands to improve their performance. This is mainly due to the improvement of the algorithm, the improvement of the sensor and the addition of heart rate monitoring equipment.

The model explored in this paper to predict possible activities based on human body information collected by wearable devices has broad application prospects in the construction of digital football systems. In football games, the possible activities of players can be roughly divided into 10 to 12 categories, and each activity corresponds to different performances of players on the field. In a standard 90-minute game, high-level players can remain physically active for 75 minutes in good condition. If the inactive state of the player is detected during the game for an abnormal duration, the coaching staff should consider replacing it immediately or help the player adjust.

For example, football players now wear sports vests on the upper body during training and competition. There is a mounting bag on the back of the sports vest, where the professional motion sensing module is placed to make it close to the back, and a heart rate belt is installed on the chest, which can monitor the athlete's total running distance, sprint running times, and high-intensity running ratio. , maximum heart rate and average heart rate, heart rate load, metabolic load and other physical skills data and analysis, these sports data will be transmitted to the PAD or laptop in the hands of coaches or professional researchers through the wireless base station on the

sidelines. When the athlete puts on the sports vest, the coach only needs to turn on the computer,

and the performance of the players will be presented in a digital form. Coaches can not only monitor the training competition in real time through these data, but also control the training intensity. It is worth mentioning that these data can also play a role in monitoring injuries. If a player's physical skill data suddenly becomes abnormal, for example, the acceleration generated by kicking the ground is much lower than the usual level, then there may be a hidden danger of injury somewhere on the player's body, and they should be taken to the team doctor in time. A comprehensive inspection can kill problems in time before they become apparent. In addition, football is a sport that combines aerobic and anaerobic training, and aerobic training and anaerobic training have different heart rate requirements. Through the use of wearable devices, players can be better controlled within a suitable heart rate range to train.

5 Summary

Behavior classification and recognition based on biological signals is an important basis for health monitoring and diagnosis of major diseases. Efficient classification algorithms can greatly save the time and energy of doctors and guardians, and provide important auxiliary information. In this paper, after preprocessing the data provided by the PPG-DaLiA dataset, the eigenvalues suitable for modeling are selected. This paper starts with a single subject and initially observes the performance of the selected classification algorithm through modeling. The performance was then remodeled and tested on the full subject dataset.

This paper mainly designs and implements four classification algorithms: softmax regression, SVM, random forest and decision tree. The experimental results show that random forest has better classification performance, and considering the interpretability and computational cost, the softmax algorithm will be more suitable.

On the basis of the research in this paper, there are still many works to be further solved and improved. The current research mainly focuses on the classification and identification of certain types of activities, and the number of samples is relatively small. It can establish cooperative relations with relevant medical institutions, expand the sample database in large quantities, and classify and identify based on more biological signals. On the basis of the existing technology, the parameter design and optimization of the model structure are discussed to further improve the accuracy and stability of the wearer's behavior monitoring. The classification algorithm is transplanted to the wearable device to build a cloud server, so that the mobile medical device can provide more humanized services.