UDC 004.391.

HUMAN ACTIVITY RECOGNITION BASED ON RANDOM FOREST

Y. XUEYING

Belarusian State University of Informatics and Radioelectronics, Republic of Belarus

Received February 26, 2022

Abstract. With the improvement of pattern recognition algorithm, human activity recognition (HAR) based on smart phone sensor data has become a highly concerned and rapidly developing research field. According to the basic principle of HAR, this paper proposes an activity recognition system based on random forest classifier. Data collection, data preprocessing and feature extraction are specified. The random forest classifier is summarized. The accuracy rate, TPR, FNR, PPV and FNR were used to evaluate HAR system proposed in this paper. Experimental results show that this study can accurately distinguish six basic activities, and the average accuracy of the random forest algorithm is 94,6 %.

Keywords: human activity recognition, classifier, random forest algorithm.

Introduction

Human Activity Recognition (HAR) has become a very active research topic in the field of machine learning [1-2]. The purpose of this work is to use sensor data and machine learning methods to detect human movement. Advanced mobile devices, such as smart phones, are usually integrated by multiple sensors. Accelerometers, gyroscopes, and magnetometers are commonly embedded. Mobile devices can obtain a large amount of user-related information by monitoring and tracking the user's movement. Classic pattern recognition steps include data preprocessing, feature extraction/selection, classifier training, and sample recognition. The research is mainly as follows:

1 Bagging algorithm [3] was proposed in 1996. The new training dataset generated by random sampling contains the same number of data instances as the original training dataset D_i, and the data is repeated due to uneven selection. In literature [4], the D_i $(i=1,2,...,T)$ the model building method of decision tree CART [5] algorithm is used for improvement. There are T decision trees in the decision forest to provide information for their decisions.

2 AdaBoost algorithm is a classical algorithm among Boost algorithms [6]. In this kind of algorithm, the base classifier is derived from the training data set D with weight factors added.

3 Random Subspace algorithm [7] re-selects data attribute subset and randomly selects attribute subset f from all feature space m in the process of each iteration to form m. The decision tree is trained from the new training data set. The accuracy of Random Subspace algorithm is between Bagging algorithm and AdaBoost algorithm, but the diversity is not as high as the above two algorithms.

This paper puts forward a kind of human activity recognition based on random forest algorithm, the target is to adopt mobile phone embedded sensor data through human activity recognition technology, achieved the identification of common 6 class activities. The experimental results show that the human activity recognition based on random forest algorithm has very high accuracy.

The processing procedures of accelerometer data

According to the classical pattern recognition process, we introduce in detail the data processing process as the pre-part of activity recognition method, including: data acquisition, data pretreatment, data segmentation, feature extraction and selection.

The user activity data is collected to analyze and verify the activity identification method proposed in this paper. The quality of data will directly affect the final result of activity identification. The difficulty of data collection can be greatly reduced by using sensors embedded in portable intelligent devices to collect data. Simple activities are characterized by high monotonicity and repeatability and obvious feature distribution.

Therefore, based on the experience of previous studies and the basic characteristics of simple human activities, we selected the sensors built into smart phones to obtain the original signals, Acceleration sensor, Gravity sensor, Gyroscope, Magnetic sensor

During the data acquisition process, due to interference from factors such as external environment and human error, the original sensor data usually contains noise (missing values, incorrect values, or abnormal values, etc.). Therefore, the original data needs to be preprocessed. The commonly used data preprocessing methods are mainly data filtering and windowing segmentation.

In data processing, the function of a filter is to remove unwanted parts of the signal, such as random noise, or to extract useful parts of the signal, such as the components lying within a certain frequency range. In this paper, Chebyshev Type II sensor is used to remove the influence of gravity. Chebyshev Type II filters have flat passbands (no ripple), making them a good choice for DC and low frequency measurement applications.

The Chebyshev high-pass filter shown in Figure 1 is used to separate the components of bulk acceleration (BA) and gravitational acceleration (GA).

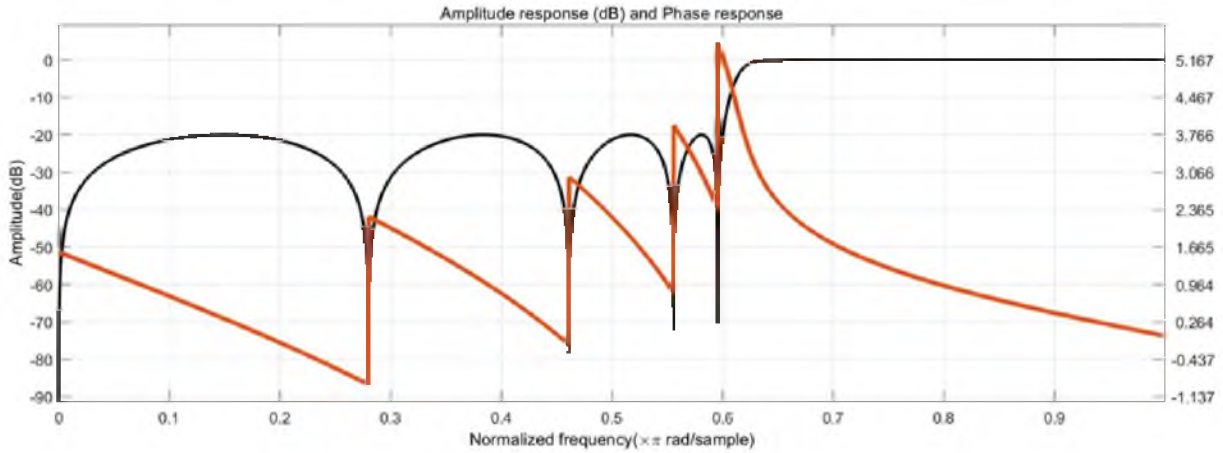
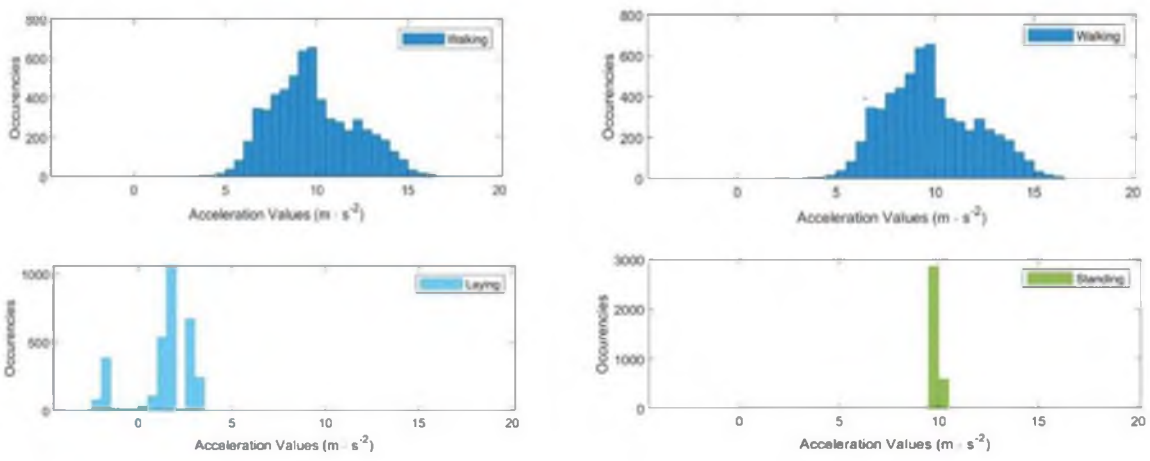


Figure 1. The amplitude and phase response of 9th-order high-pass Chebyshev Type II filter

Due to the data from the sensor is time domain data points in a row, high dimension, not suitable for directly used for feature extraction, otherwise it will affect the accuracy of the system and efficiency. In this paper, a 50 % overlap slider Window is selected to windotize the data.

Feature extracting. Extracting key features to represent human activities is the key to activity analysis and behavior analysis. Feature selection has a very important impact on the recognition effect. Features with significant discriminative power can greatly improve recognition accuracy or reduce model complexity. Active features usually include time domain features, frequency domain features and some other domain features. In this paper, we extracted 66 temporal features from the ACC data of 6 sports activities, as shown in Figure 2.

It can be seen from Figure 2, *a* that walking and lying down can be easily distinguished by using the feature vector consisting of histogram values of ACC signal. It can be seen that walking and standing can be easily distinguished by using the standard deviation, as shown in Figure 2, *b* shown.



a *b*

Figure 2. Histograms of ACC data for 3 classes of human physical activities:
a – Walking and Lying; *b* – Walking and Standing

Random forest (RF) is a supervised machine learning algorithm that has become one of the most commonly used algorithms due to its accuracy, simplicity and flexibility. It has nonlinear properties that enable it to be used for classification and regression tasks and is highly adaptable to a range of data and situations. RF use Bagging (picking one observation sample instead of all) and random subspace methods (picking one feature sample instead of all features, in other words – attribute bagging) to grow a tree. RF has simplicity, diversity, robustness and reliability, and can greatly improve model accuracy by adjusting hyperparameters and selecting key features.

RF learning algorithm is composed of three parts like any other machine learning algorithm. Training, validation and testing. RF learning provides us with several Spaces (optimal combination of features) in which we can perform optimal domain partitioning and achieve high classification accuracy.

The base classifier of RF is decision tree, which belongs to a single classifier. A decision tree can provide a set of rules to distinguish between different characteristics of data. The basic idea of decision tree algorithm is to use the training set of known categories to train classifiers and generate rules. These rules are used to classify and mine unknown data sets.

The decision tree model is a tree structure, including three types of nodes (Figure 3).

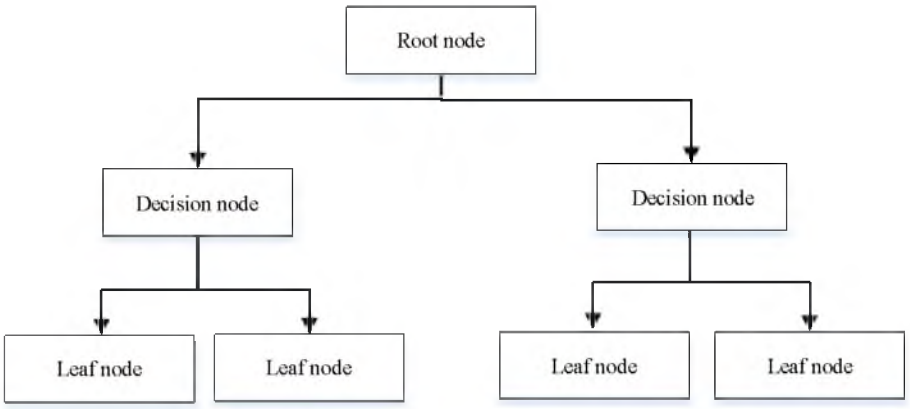


Figure 3. Decision tree structure used in RF classifier

The nodes include root decision node, intermediate decision node and leaf node. Among them, the decision child node represents the determination of feature attributes, and the leaf node represents the determination of feature attributes. The child node represents a category, and the path of the connected node represents the output of the characteristic attribute within a certain range of values. The decision-making process of decision tree is the process of testing the characteristic attributes of the data to be classified, selecting the output branches, and arriving from the root node to the leaf node.

The most critical step in constructing decision tree is to determine split attribute. Split attribute refers to dividing and constructing different branches according to a certain characteristic attribute at

non-leaf nodes. The key lies in attribute selection metric, which is a criterion for selecting split attribute. There are many node splitting algorithms, which generally adopt the top-down recursive divide and conquer method and the greedy strategy without backtracking.

This paper adopts CART algorithm, which adopts the splitting rule *Gini* index minimum principle when nodes are split. *Gini* coefficient calculation formula is as follows:

$$Gini(P) = 1 - \sum_{i=1}^{C} (p_i)^2, \tag{2}$$

where $P = (p_1, p_2, \ldots, p_n)$; p_i is the probability of an object being classified to a particular class; C is the number of physical activities and i is an index of the class.

The *Gini* index is applied to categorize the target variables as "success" or "failure". Since it only performs binary splitting, a higher *Gini* index means greater inequality and heterogeneity.

The construction process of random forest mainly includes three steps. First, sample the data set and generate a training set for each decision tree; then use each training set to build a decision tree. The process of generating a decision tree does not require pruning processing; and finally generate a random forest and perform a classification algorithm.

The growth of each decision tree is as follows:

1. Random record selection: each tree training accounts for roughly two-thirds of the total training data was (63,2 %). Cases were randomly selected and replaced from the original data. The sample will be used as a training set for growing trees.

2. Random variable selection: randomly selected from all the predictor variable m, for example, some predict variables, and use these m the best segmentation to segment nodes.

The combination of multiple decision trees established through the above two steps will become a random forest. The decision tree uses voting to participate in decision-making, and the category with the most votes will be the final output of the random forest algorithm. The random forest pseudocode used for classification application.

Activity recognition performance evaluation

The random forest algorithm is used to verify the existing model using a verified subset of the research data set. The 10-fold cross-verification method is used for the training subsets to verify the cross-verification model. The experiment used different random seeds for multiple experiments. Use the obfuscation matrix to verify the performance of the model. The results are shown in Table 1.

Table 1. **Confusion matrix of recognized activities**

Human Activity Recognition	Walking	Walking upstairs	Walking downstairs	Sitting	Standing	Laying
Walking	1668	15	39	0	0	0
Walking upstairs	21	1502	21	0	0	0
Walking downstairs	44	54	1308	0	0	0
Sitting	0	0	0	1624	152	1
Standing	0	0	0	206	1699	1
Laying	0	0	0	0	0	1944

The confusion matrix represents the dispositions of the set of instances, on which a classifier is tested. The example is given in Table 2.

Table 2. **The estimations of activity recognition performance**

Human Activity Recognition	TPR (%)	FNR (%)	PPV (%)	FDR (%)	Precision (%)
Walking	96,9	3,1	96,2	3,8	96,9
Walking upstairs	97,3	2,7	95,6	4,4	97,3
Walking downstairs	93,0	7,0	95,6	4,4	93,0
Sitting	91,4	8,6	88,7	11,3	91,4
Standing	89,1	10,9	91,8	8,2	89,1
Laying	100	0	99,9	0,1	100

The highest values in the result percentage of the confusion matrix indicate a higher success in recognition performance i.e. The error percentage is minimal. It can be seen from the confusion matrix that the recognition of human activities based on random forests can accurately identify basic human activities: walking, walking upstairs, walking downstairs, Sitting, Standing, and Laying.

TPR (true positive rate), FNR (false negative rate), PPV (positive predictive value) and FDR (false discovery rate) were used to demonstrate the identification performance of using random forest, the experimental results show that this study can accurately distinguish the six basic activities, and clearly shows that the random forest algorithm in the activity recognition experiment of higher accuracy, the average accuracy is 94,6 %.

Conclusion

With the development of wearable devices, human activity recognition through sensor data has become an important research field. All kinds of classifiers in machine learning show excellent performance in their own fields. This paper reviews and summarizes human activity recognition based on random forest classifier. Firstly, the set of embedded sensors for data acquisition and their functions, Chebyshev Type II high-pass filter for data processing and feature extraction using time domain features are introduced. Secondly, the random forest classifier, its base classifier, and the decision tree node splitting algorithm are described and analyzed. Finally, with TPR, FNR, PPV, FDR and accuracy as evaluation indexes, the human activity recognition system is evaluated through experiments. Research shows that the recognition of human activities based on random forest has high average accuracy (94,6 %) for 6 classes of physical activities using time features.

References

1. Thakur D., Biswas S. // Smartphone based human activity monitoring and recognition using ML and DL: a comprehensive survey. 2020. P. 433–444.
2. Beddiar D. R., Nini B. // Vision-based human activity recognition: a survey. 2020. P. 509–555.
3. Breiman L. // Machine Learning. 1982. Vol. 1. P. 123–140.
4. Han Jiawei. Data mining: concepts and techniques. Morgan Kaufmann Publishers Inc, 2005.
5. Loh W. Y. // Wiley Interdisciplinary Reviews Data Ming and Knowledge Discovery. 2011. Vol. 1. P. 14–23.
6. Freund Y., Schapire R. E. // Experiments with a new boosting algorithm. 2017. Vol. 1. P. 12–15.
7. Fawagrehk G. // Systems Science and Control Engineering. 2014. P. 602–609.