Human Activity Recognition Using Gaussian Naïve Bayes Algorithm in Smart Home

Jiajie Shen and Hongqing Fang
College of Energy and Electrical Engineering, Hohai University, 8 Focheng West Road, Nanjing 211100, Jiangsu, China
Email: chineseyouth@163.com; fanghongqing@sohu.com

Abstract. Human activity recognition (HAR), as an important research issue, aims to identify human activities in smart homes. In this paper, we apply Gaussian Naïve Bayes (GNB) algorithm to HAR and evaluate the model based on smart environment sensor data. Experimental results show that the effective selection and processing of features are helpful to improve the accuracy of activity recognition of the model. Compared with NB whose accuracy rate is 82.7%, GNB has a better accuracy rate of 89.5% and even has a higher recognition accuracy in almost every category of activities. Selecting the feature variables as good and useful as possible to get a better model in the process of activity recognition is conducive to the correct classification of samples by machine learning algorithm and improves the classification performance of the model.

1. Introduction
With the rapid increase of life expectancy, ageing is becoming a hot topic and much more attention has been paid to the topic of ageing of population [1]. For example, with the highest proportion of the elderly population as of 2015, Japan also has the second largest number of older adults among the developed countries. In recent years, China’s ageing population is increasingly prominent as well. In 2015, the proportion of the elderly aged 60 or over accounted for 15.2% of the total population in China [2]. As mentioned above, we can draw a conclusion that the research of HAR is of great necessity and importance.

Smart homes which provide a better and more intelligent home environment for human beings give more convenient services for the elderly at home [3] with the application of various activity recognition algorithms based on smart environment sensor data.

CASAS, funded by Washington State University, is an intelligent research project. And the targets of the CASAS project are to create an energy-saving environment and maximize residents’ comfort. The smart homes [4] are typical examples of external sensing in which the sensors are fixed in predetermined points of interest to sense the environment. Then executing devices make corresponding actions after the recognition of human activities using certain algorithms. Lots of classification algorithms, such as Support Vector Machine [5], Hidden Markov Model [6] and Decision Tree [7] which are used to create the smart homes have been reported.

2. Method

2.1. Smart Apartment Testbed and Experiment Data
The smart apartment testbed is used to obtain the experiment data. As figure 1 shows, there are a living room, one kitchen, three bedrooms, a storage and a bathroom in the apartment where motion sensors are
distributed on the ceiling to detect behaviors and temperature sensors are used to acquire ambient temperature information.

![Figure 1. Sensors distribution in the smart apartment.](image)

The data collected from the smart apartment are processed and labeled to form the dataset. And the representation of dataset contains five columns including date, time, sensor ID, sensor status and label. An example of activity Leave_home is expressed in the following table.

| Date of activity | Time of activity | Sensor ID | Status | Label         |
|------------------|------------------|-----------|--------|---------------|
| 2009-06-13       | 07:44:17.008     | M025      | ON     | Leave_home begin |
| 2009-06-13       | 07:44:23.042     | M014      | OFF    |               |
| 2009-06-13       | 07:44:25.045     | M027      | ON     |               |
| 2009-06-13       | 07:44:30.027     | M025      | OFF    |               |
| 2009-06-13       | 07:44:36.056     | M027      | OFF    | Leave_home end |

As seen above, the sensor events of a detected activity are shown with the parameters of concrete date of event, time of event, activity label, sensor ID and status. This activity started on July 13, 2009 at 7:44:17. And the activity triggered the first sensor M025 whose value is ON. M027 is also triggered and M014 is closed during the activity. The activity ended on July 13, 2009 at 7:44:36.

The training and testing process of HAR can be seen in figure 2. K-fold cross validation [8] method is used in the whole training and testing process where K’s value is 3.

2.2. Feature Selection
Selecting features which affects algorithm recognition efficiency is directly related to the classification performance of HAR. According to the sensor data and the actual situation, the following features can be obtained.

1. The duration time of the current activity. It is converted into a value from 0 to 2 that represents 3 thresholds in order to discriminate the length of time between various activities.
2. The beginning day of the activity. It is translated into a value from 0 to 6 that represents the day on which the current activity occurred.
3. The ending day of the activity, which represents the ending day of week.
4. The beginning hour of a day, which is discretized to a value whose scale is 24 from 0 to 23.
(5) The ending hour of a day.

(6) Previous activity.
(7) First sensor whose state is ON of the activity.
(8) Last sensor whose state is ON of the activity.
(9) The number of ON sensors, which denotes the activity’s length measured in the quantities of sensors whose state are ON.
(10) Average sensor ID of each activity, the expression is shown in equation (1).

\[
\bar{S}_{ij} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} S_{ijk}
\]

\(\text{where } \bar{S}_{ij} \text{ is the average value of all triggered sensor IDs that are ON at the } j_{th} \text{ occurrence of activity } i, \ n_{ij} \text{ represents the quantities of sensors that are ON, } S_{ijk} \text{ denotes the } k_{th} \text{ sensor ID of the activity.}
\]

2.3. Gaussian Naïve Bayes Algorithm Applied for Activity Recognition
NB is a classification method based on Bayes theorem [9-10] and the independent assumption of conditions. For a given training data set, use Bayes’ theorem to find the output \(Y\) with the largest posterior probability which is obtained from the likelihood function and the prior probability. Following Bayes theorem, the posterior distribution over the parameter space is proportional to the likelihood function times the prior distribution.

The conditional probability in the form of \(P(C1|C2)\) represents the occurrence probability of \(C1\) when the condition \(C2\) is satisfied. According to the independent assumption of conditions and the conditional probability formula, the conditional probability formula is:

\[
P(c|x) = \frac{p(c)}{P(x)} \prod_{i=0}^{d} P(x_{i}|c)
\]

\(\text{Since } P(x) \text{ is the same for all categories, the following Bayesian classifier expression can be obtained:}
\]

\[
h_{nb}(x) = \arg \max_{c} P(c) \prod_{i=0}^{d} P(x_{i}|c) \quad c \in Y
\]

where \(c\) belongs to \(Y\) which is the collection of all activity categories, \(Y = \{c_1, c_2, \ldots, c_N\}\), the number of categories of activities is represented by \(N\), \(d\) denotes the number of features. \(x_i\) represents the \(i_{th}\) feature.
As one of the most principal and widely used methods of numerical analysis, Gaussian distribution plays a key role in the probability calculation of continuous feature variable. And its mathematical expression is as follows:

\[
P(x_i|c) = \frac{1}{\delta_{c,i}(2\pi)^{\frac{D}{2}}} e^{-\frac{(x_i-\mu_{c,i})^2}{2\delta_{c,i}^2}}
\]  

(4)

where, \(P(x_i|c) \rightarrow N(\mu_{c,i}, \delta_{c,i}^2)\) that means the probability of \(i_{th}\) feature value of activity \(c\) meets the normal distribution.

In order to improve activity recognition accuracy and training reliability, we perform correction on the likelihood function. The equations are stated in the following (5) and (6):

\[
P(c) = \frac{D_c+1}{D+N}
\]

(5)

\[
P(x_i|c) = \frac{D_{c,x_i}+1}{D_{c}+N_i}
\]

(6)

where, \(D_c\) is the number of activity \(c\) in training data, \(D_{c,x_i}\) is the number of feature \(x_i\) in activity \(c\). \(N\) represents the number of activity categories and \(N_i\) is the number of possible values of feature \(i\).

3. Tests Results

3.1. Activities for Classification

There are 10 activities in total. And they were repeatedly executed in the intelligent living environment through 2 volunteers to acquire sensors data serving for GNB and NB algorithms. These activities are shown below:

1. Go to washroom (activity label: A)
2. Have breakfast (activity label: C)
3. Go to bed (activity label: E)
4. C_programming (activity label: G)
5. Have dinner (activity label: I)
6. Do laundry (activity label: K)
7. Leave home (activity label: M)
8. Have lunch (activity label: P)
9. Night wandering (activity label: R)
10. Take medicine (activity label: T)

3.2. Results and Comparison

Tables 2-3 show the confusion matrix obtained from the human activity recognition classification. The two algorithms are GNB model with 10 features (table 2) and NB model with 5 features (table 3).

Test results using (table 2) GNB and (table 3) NB show that the accuracy achieved from GNB which has 10 features and gets correction is higher than NB which has 5 features without any correction in almost every category of activities (except A8).

| Activity label | A   | C   | E   | G   | I   | K   | M   | P   | R   | T   |
|----------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| A              | 0.900 | 0 | 0.033 | 0 | 0 | 0 | 0 | 0.067 | 0 | 0 |
| C              | 0 | 1.000 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| E              | 0.030 | 0 | 0.960 | 0 | 0 | 0 | 0 | 0.010 | 0 | 0 |
| G              | 0 | 0.022 | 0.154 | 0.783 | 0 | 0 | 0 | 0.041 | 0 | 0 |
| I              | 0 | 0 | 0 | 1.000 | 0 | 0 | 0 | 0 | 0 | 0 |
| K              | 0 | 0 | 0 | 0 | 0.600 | 0.400 | 0 | 0 | 0 | 0 |
| M              | 0 | 0 | 0 | 0 | 0 | 1.000 | 0 | 0 | 0 | 0 |
| P              | 0 | 0.081 | 0 | 0 | 0 | 0 | 0 | 0.919 | 0 | 0 |
| R              | 0.227 | 0.265 | 0 | 0 | 0 | 0 | 0 | 0.508 | 0 | 0 |
| T              | 0 | 0.022 | 0.023 | 0 | 0 | 0 | 0 | 0 | 0.955 | 0 | 0 |
Table 3. Confusion matrix: NB.

| Activity label | A   | C   | E   | G   | I   | K   | M   | P   | R   | T   |
|----------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| A              | 0.367 | 0 | 0.033 | 0 | 0 | 0 | 0 | 0 | 0.600 | 0 |
| C              | 0 | 0.896 | 0.104 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| E              | 0.009 | 0 | 0.879 | 0.034 | 0 | 0 | 0 | 0 | 0.078 | 0 |
| G              | 0 | 0 | 0.413 | 0.565 | 0 | 0 | 0 | 0.022 | 0 | 0 |
| I              | 0 | 0 | 0 | 0.024 | 0.976 | 0 | 0 | 0 | 0 | 0 |
| K              | 0 | 0 | 0 | 0.300 | 0 | 0.300 | 0.200 | 0.200 | 0 | 0 |
| M              | 0 | 0 | 0.015 | 0.027 | 0 | 0.015 | 0.928 | 0.015 | 0 | 0 |
| P              | 0 | 0.027 | 0 | 0.027 | 0 | 0 | 0 | 0.108 | 0.838 | 0 |
| R              | 0.015 | 0.045 | 0 | 0 | 0 | 0 | 0 | 0 | 0.940 | 0 |
| T              | 0 | 0 | 0.273 | 0 | 0 | 0 | 0 | 0 | 0 | 0.727 |

4. Conclusion
This paper applies Gaussian Naive Bayes (GNB) model to HAR based on the sensor data. From the test results we learn that the performance of accuracy generated by GNB classifier is much better than NB classifier. The accuracy produced by GNB is 89.5%, while NB is 82.7%. Therefore, reasonable selection and processing of features have a huge impact on recognition accuracy. The optimal parameters of features must be determined through the experiment.

Acknowledgement
The authors are very grateful to Dr. Diane J. Cook and anonymous reviewers who provided most valuable suggestions for promoting the quality of this paper.

References
[1] Saibene A and Gasparini F 2019 Cognitive and physiological response for health monitoring in an ageing population: A multi-modal LNCS 11938.
[2] Angela C. Lyons, John E. Grable and So-Hyun Joo 2018 A cross-country analysis of population aging and financial security The Journal of the Economics of Ageing 12 96-117.
[3] Pal D, Triyason T, Funilkul S and Chutimaskul W 2018 Smart homes and quality of life for the Elderly: Perspective of competing models IEEE Access 6 8109-22.
[4] Guo K, Li Y, Lu Y, Sun X, Wang S and Cao R 2016 An activity recognition-assistance algorithm based on hybrid semantic model in smart home Int. J. Distrib. Sens. Netw. 12 8.
[5] An F 2018 Human action recognition algorithm based on adaptive initialization of deep learning model parameters and support vector machine IEEE Access 6 59405-21.
[6] Sarkar and Jehad A M 2014 Hidden markov mined activity model for human activity recognition Int. J. Distrib. Sens. Netw. 2014 949175.
[7] Fan S, Jia Y and Jia C 2019 A feature selection and classification method for activity recognition based on an inertial sensing unit Information 10 (10) 290.
[8] Qi C, Diao J and Qiu L 2019 On estimating model in feature selection with cross-validation IEEE Access 7 33454-63.
[9] Zhao T, Deng L, Wang W, Elson D S and Su L 2018 Bayes theorem-based binary algorithm for fast reference-less calibration of a multimode fiber Opt. Express 26 (17) 20368-78.
[10] Zhao X, Yang J and Zhou J 2013 An improved TANC method based on Bayesian equivalence theorem Inf. Technol. J. 12 (17) 4336-9.