CLASSIFICATION AND PREDICTION OF BASIC HUMAN PHYSICAL ACTIVITY USING LR MACHINE LEARNING ALGORITHM

Mr. S. Ashok Kumar, Research Scholar, School of Computer Science, Park's College (Autonomous), Tirupur, India
Dr. K.P. Rajesh, Associate Professor & Head, School of Computer Science, Park's College (Autonomous), Tirupur, India

Abstract

The machine learning and deep learning algorithms are broadly used in various research areas such as medical industry and especially for predicting the physical human activities. Human activity recognition is the ability of the machine to predict the basic movement of a human body based on sensor data set provided to them for learning. The proposed activity is such a classification and prediction of human basic activities such as walking, moving upstairs, moving downstairs, simply sitting, standing or lying. The data set received from the recognized repositories are taken, preprocessed and used with the supervised machine learning algorithm Logistic Regression(LR) algorithm. The human activity recognition was carried out with text based dataset. The preprocessed dataset are prepared and are fed to the created algorithm and the results were analyzed and discussed. The classification and predication accuracy of the algorithm was found to be 95.8% for the datasets taken.

Keywords: Human activity, Machine Learning, Logistic Regression, Supervised learning

1. Introduction

The Machine learning is programming computers to optimize a performance criterion using an example data or using the past experience. The model may be a predictive model to make predictions in the future, or it can be a descriptive model which is used to gain knowledge from data, or it can be of both types. Machine learning is mostly worried using the correct features used to build the right models that can solve the correct tasks.

The human activity recognition is the most popular and active research area by using machine learning algorithms. This was done using sensors, accelerometer, gyroscope etc. The accelerometer provides the maximum functionality and followed by the gyroscope [9]. The recognition of human activity is majorly used in the health care system which is installed in the residence, hospitals and other medical based centers. These are also used in disease management and disease prevention projects. The proposed activity is such a recognition and prediction of human basic activities such as walking, moving upstairs, moving downstairs, simply sitting, standing or lying.

2. Background Study

Chieh-ChenWu, Wen-ChunYeh et al., developed and compared classification models to predict fatty liver disease. Classification models such as random forest (RF), Naïve Bayes (NB), artificial neural networks (ANN), and logistic regression (LR) were developed to predict Fatty liver disease [1].

Varun Arvind, Jun S. Kim, Eric K. Oermann et al. proposed that NN and LR algorithms outperform other classification algorithms for predicting individual postoperative complications.

With growing size of medical data the training of machine learning on these large datasets promises to improve risk prognostication with the ability of continuously learning making them excellent tools in complex clinical scenarios[4].

Helen R.Marucci-Wellman, Helen L.Corns Mark R.Lehto proposed methods build off our prior results by integrating Logistic Regression and Support Vector Machine algorithms with Naïve Bayes and that results in high accuracy, that significantly reduce the human resources required to accomplish the task. They also suggested that if resources are constrained at a low level then the best

approach for accuracy will be to combine the manual coding along with codes assigned by the LR algorithm[8].

Negar Golestani and Mahta Moghaddam propose that recognizing human physical activities using wireless sensor networks has attracted significant research interest due to its broad range of applications.. There proposed that there are serious challenges in designing a sensor-based activity recognition system that operates in and around a lossy medium such as the human body to gain exchange between computational complexity and accuracy. They also introduce a wireless system based on magnetic induction for recognizing the human activity to deal with such challenges and constraints. The magnetic induction system was integrated with the machine learning LR techniques which helps to detect wide range of human movement[12].

3. Logistic Regression Algorithm

Supervised Learning Logistic Regression (LR) is a popular machine learning algorithms. It is used for predicting the definite dependent variable using a given set of independent variables. Logistic regression predicts the output of a definite dependent variable. Therefore the outcome must be a definite or discrete value. Therefore it can be either yes-no, 0-1 or true-false in case of definite value and it gives the probabilistic values which lie between 0 and 1 in case of discrete value. Linear Regression algorithm is mostly used for solving various Regression problems but Logistic regression is used to solve the classification problems**.**

In Logistic regression, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1). The curve from the logistic function point towards the possibility of something. Logistic Regression is a supervised machine learning algorithm because it has the ability to provide the probabilities and classify the new data either by using the continuous and discrete datasets. Logistic Regression can be used for classification of the observations by using different types of data and can easily determine the useful variables used for the classification problems. Logistic Regression is classified into three types as:

- Binomial**:** In binomial LR, there can be only two possible types of the dependent variables such as 0 or 1.
- Multinomial**:** In multinomial LR there can be 3 or more possible unordered types of the dependent variable, such as cat, dogs or sheep
- Ordinal: In ordinal LR there can be 3 or more possible ordered types of dependent variables, such as low, medium or high.

3.1 Logistic Regression Equation

The Logistic Regression equation can be achieved from the linear regression equation. The mathematical steps to get the Logistic Regression equations is given below, we know that the equation of the straight line is:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \ldots\ldots + b_nx_n$$

In Logistic Regression y can be between 0 and 1 only, so we will divide the above equation by (1-y):

$$\frac{y}{1-y}\ ;\ 0 \text{ for } y=0, \text{ and infinity for } y=1$$

Since we have to change the range between -[infinity] to +[infinity], then take logarithm of the equation it will be as

$$\log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \ldots\ldots + b_nx_n$$

The above equation is the equation for Logistic Regression

ISSN: 2278-4632
Vol-12 Issue-11 No.03 November 2022

4. Preparing the Data

In this research, WISDM Smartphone and Smart watch Activity and Biometrics Dataset is used to predict the human activity. This research work was developed for the pre-trained ML models to classify the different activities of human. The dataset i.e., WISDM Smartphone and Smart watch Activity and Biometrics Dataset is downloaded from the UCI repository.

Smartphones contains accelerometers that measure acceleration in all three spatial dimensions namely x, y and z and the raw accelerometer signal data sourced from WISDM Lab are downloaded as raw data file. This data is collected from different users as they performed some day-to-day human activities such as walking, sitting, standing, jogging, ascending and descending stairs for a specific period of time. The dataset has 564 columns and our target variable (class-label) is one of the six 'activity' which we intend to predict. The fig 4.1 shows an insight of the dataset taken for training

tBodyAcc-mean()-X	tBodyAcc-mean()-Y	tBodyAcc-mean()-Z	tBodyAcc-std()-X	tBodyAcc-std()-Y	tBodyAcc-std()-Z	tBodyAcc-mad()-X
0.25717778	-0.02328523	-0.014653762	-0.938404	-0.92009078	-0.66768331	-0.95250112
0.28602671	-0.013163359	-0.11908252	-0.97541469	-0.9674579	-0.94495817	-0.9867988
0.27548482	-0.02605042	-0.11815167	-0.99381904	-0.96992551	-0.96274798	-0.99440345
0.27029822	-0.032613869	-0.11752018	-0.99474279	-0.97326761	-0.96709068	-0.99527433
0.27483295	-0.027847788	-0.12952716	-0.99385248	-0.96744548	-0.97829499	-0.9941114
0.27921995	-0.018620399	-0.11390197	-0.99445523	-0.97041688	-0.96531629	-0.99458514
0.27974586	-0.018271026	-0.10399988	-0.99581919	-0.97635361	-0.97772468	-0.99599613

angle(tBo	angle(tBo	angle(X,gravityMean)	angle(Y,gravityMean)	angle(Z,gravityMean)	subject	Activity	ActivityName
-0.82589	0.271151	-0.72000927	0.27680104	-0.057978304	2	5	STANDING
-0.20438	-0.10683	-0.81653306	0.17052513	-0.09101897	2	4	SITTING
0.097469	0.224834	0.76371827	-0.46991662	-0.51785124	2	6	LAYING
0.813487	-0.64668	-0.68485322	0.3154051	0.003481795	2	1	WALKING
0.98929	-0.16834	-0.67627143	0.32173046	0.012499478	2	3	WALKING_DOWNSTAIRS
0.527328	-0.68917	-0.30691258	0.30394638	0.4524172	2	2	WALKING_UPSTAIRS
0.084313	-0.40584	-0.62680985	0.32843315	-0.098609873	2	5	STANDING

Fig. 4.1. Sample Dataset for Training

4.1 Data Cleaning and Preprocessing

Before we start training our models, the dataset needs to be cleaned and organized. Hence preprocessing of data set is required to get the desired output. The following are the steps performed for processing the dataset so that the desired output can be reached.

1. The null values are identified and dropped.
2. The datatype of the 'z-axis' column is changed to floating values.
3. Also drop the rows where the timestamp value equal to 0.
4. Sort the data in the ascending order based on activity columns.

After preprocessing there are 7352 rows of data available for training the LR algorithm and 2947 rows of data is kept for testing the algorithm for classification and predictions. The implementation part of the work is explained below. Python and Juypter interface were used for coding and implementation. Initially set the required packages by import using them and also set the input directory.

All the human activity based data is preprocessed and stored in the file DATA.csv. This is fed to the python code for processing. The data is read from the ht excel file and verified with coding. The sample code is presented. The length of the raw data taken is 7352. Necessary labels are added for processing.

```
# Set dataset file name
input_data_file = 'DATA.csv'
# Display list of files in input directory
os.listdir('input')
# To get the input data path
data_path = os.path.join(os.curdir,input_dir,input_data_file)
data_path
# To read data from excel file
raw_data = pd.read_csv(data_path)
raw_data.head()
# save the labeled data
raw_data.to_csv('input/LS.csv',index=False)
# Read the labeled dataset
def load_data():
data = pd.read_csv('input\LS.csv',encoding='cp1252')
return data
Ldf = load_data()
```

Fig. 4.1.2 Code for reading dataset from files

ISSN: 2278-4632
Vol-12 Issue-11 No.03 November 2022

4.2 Splitting dataset for Training and Testing

The dataset is divided so that the training and testing data are different. To analyze the class label distribution among the training data, the number of samples by activity is plotted against the activity. The following chart depicts the same. The activities such as standing, laying, sitting, walking, walk downstairs and walk upstairs are plotted again the number of dataset taken.

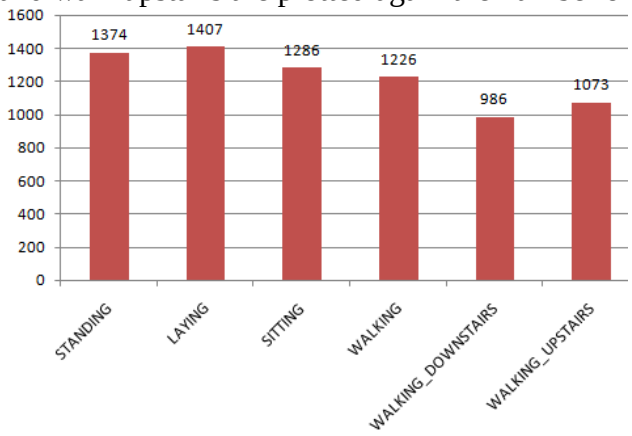

Fig. 4.2.2 Comparison of data taken from training based on the activity class

It can be noted from the above graph that except walking downstairs label all the other labels where fed with data that is almost above 1050 samples for the purpose of training. The variation in dataset was planned so as to identify the prediction rate and accuracy of the algorithm training. 2947 data sets are used to test and check for classifying.

5. Training and Classification with ML-LR Algorithm

Once the data is fed and trained and tested the end result of the logistic regression algorithm for predicted the basic physical human activity is 0.95795. Thus, we have got nearly 96% accuracy on the test data and with error rate of 3.4% which is comparatively a small number. Now let's observe the confusion matrix to see how our model performs on each of the six class label that is chosen.

The confusion matrix explains how well the algorithm was able to classify and predict based on its previous learning's. The confusion data will be the error rate of the algorithm.

Standing	492	1	3	0	0	0
Laying	25	444	2	0	0	0
Sitting	4	13	403	0	0	0
Walking	0	3	0	431	57	0
Walk_Down	0	0	0	16	516	0
Walk_Up	0	0	0	0	0	537
	Standing	Laying	Sitting	Walking	Walk_Down	Walk_Up

It can be observed from the above confusion matrix that the three most common activities in our dataset are walking, walking downstairs and walking upstairs. In that the walking upstairs was classified perfectly since the training data was more. In the case of walking down stairs which had the less number of training data 57 numbers of dataset were classified as walking instead of walking down stairs. This is expected as these can be fined tuned by providing sufficient data during training so as to accurately predict them. The classification and prediction accuracy was also better than walking down stairs for the other three classes.

6. Conclusion

The human activity recognition was implemented with text based dataset and the results were found on a positive note. The process started with processing the raw data consisting of six relevant

features of physical activity of human. Then we trained the data with supervised LR machine learning network on the preprocessed data. The LR algorithm learned the complex features automatically from the provided data and it is able to predict the class label with high accuracy. The LR algorithm was also able to distinguish and recognize all the activities with better accuracy.

References

1. Chieh-ChenWu, Wen-ChunYeh, Wen-DingHsu, Md. MohaimenulIslam, Phung Anh (Alex)Nguyen, Tahmina NasrinPoly, Yao-ChinWang, Hsuan-ChiaYang, Yu-Chuan(Jack) Li, Elsevier, Computer Methods and Programs in Biomedicine Volume 170, March 2019, Pages 23-29
2. Michalis Vrigkas, Christophoros Nikou, Ioannis A Kakadiaris, "A review of Human activity recognition methods" Review article Front. Robot. AI, Sec. Robot and Machine Vision, 16 November 2015, https://doi.org/10.3389/frobt.2015.00028
3. Jennifer R. Kwapisz, Gary M. Weiss, and Samuel A. Moore (2010). Activity Recognition using Cell Phone Accelerometers, Proceedings of the Fourth International Workshop on Knowledge Discovery from Sensor Data (at KDD-10), Washington DC
4. Varun Arvind, Jun S. Kim, Eric K. Oermann, Deepak Kaji, and Samuel K. Cho, Neurospine. 2018 Dec; 15(4): 329–337. Published online 2018 Dec 17. doi: 10.14245/ns.1836248.124
5. Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra and Jorge L. Reyes-Ortiz. A Public Domain Dataset for Human Activity Recognition Using Smartphones. 21st European Symposium on Artificial Neural Networks, Computational Intelligence, and Machine Learning, ESANN 2013
6. Deepti Sehrawat, Nasib Singh Gill, "IoT Based Human Activity Recognition System Using Smart Sensors, ASTES Journal, Volume 5, Issue 4, Page No 516-522, 2020
7. Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. Cancer Inform. 2007;2:59–77.
8. Helen R.Marucci-Wellman, Helen L.Corns[a]Mark R.Lehto, Elsevier, Accident Analysis & Prevention Volume 98, January 2017, Pages 359-371
9. Aman Kharwal, Human Activity Recognition using machine learning, The cleverprogrammer article, January 10,2021
10. Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. J Biomed Inform. 2002;35:352–9
11. Nour Takiddeen, Imran Zualkernan et. al, "Smart Watches as IOT Edge devices: A Framework and survey", 2019 Fourth International Conference on Fog and Mobile Edge Computing (FMEC), DOI: 10.1109/FMEC.2019.8795338
12. Negar Golestani & Mahta Moghaddam**,** Nature Communications **V**olume 11, Article number: 1551 (2020)**,** Nature communications - Human activity recognition using magnetic induction-based motion signals and deep recurrent neural networks
13. Hao Liu, Huaibin Quing, etal "A promising material for human friendly functional wearable electronics" Science Direct, Materials Science and Engineering reports, Volume 112, February 2017, pages1-22
14. F. Attal, S. Mohammed, M. Dedabrishvili, F. Chamroukhi, L. Oukhellou, Y. Amirat, "Physical human activity recognition using wearable sensors" Sensors, 15(12), 31314-31338, 2015
15. K. Butchi Raju, Suresh Dara, Ankit Vidyarthi, V. MNSSVKR Gupta, Baseem Khan, "Smart Heart Disease Prediction System with IoT and Fog Computing Sectors Enabled by Cascaded Deep Learning Model", Computational Intelligence and Neuroscience, vol. 2022
16. M. Ganesan, N. Sivakumar, IoT based heart disease prediction and diagnosis model for healthcare using machine learning models, in: Proceedings of the IEEE (ICSCAN), 2019, pp. 1–5,
17. Alexander Genkin, David D. Lewis, David D. Lewis. 2007. Large-Scale Bayesian Logistic Regression for Text Categorization. Technometrics. 49(3).
18. K.G. Dinesh, K. Arumugaraj, K.D. Santhosh, V. Mareeswari, Prediction of cardiovas cular disease using machine learning algorithms, in: Proceedings of the International Conference on Current Trends towards Converging Technologies (ICCTCT), 2018, pp. 1–7.

