2022 International Conference on Machine Learning and Intelligent Systems Engineering (MLISE) | 978-1-6654-9246-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/MLISE57402.2022.00087

Prediction of People's Abnormal Behaviors Based on Machine Learning Algorithms

Xiaoyu Song[1],*
[1]School of Communication & Information Engineering
Shanghai University
99 Shangda Road, Baoshan District, Shanghai, 200444, China
* Corresponding author: 825106963@shu.edu.cn

***Abstract*—Some improper behaviors in specific situations may put people in danger, such as smoking in a gas station, therefore they need to be detected. This paper tries to find out the best Machine Learning algorithm to address that kind of prediction problems. Datasets related to behavior detection are collected, whose categories consists of smoking, calling and normal behaviors. Experiments based on several famous algorithms are conducted, including Linear Support Vector Machine (LSVM), Kernel Support Vector Machine (KSVM), Decision Tree Classifier (DT), Random Forest Classifier (RF), K-nearest Neighbors (KNN) and K-Means Clustering. Additionally, Confusion Matrix and Mean Squared Error (MSE) are used to judge the performance of each algorithm. Finally, Principal Component Analysis (PCA) visualizes the outcome of the best algorithm. The results show that Random Forest Classifier (RF) achieves the best performance and is capable of predicting people's abnormal behaviors with an accuracy of 82%.**

Keywords-component; Machine Learning; Abnormal behaviors prediction; Dimensionality reduction

I. INTRODUCTION

Nowadays, people are paying more attention to their health, but there are still a lot of dangerous behaviors that may get people injured. They are extremely threatening in some specific situations. For example, talking on the phone while driving distracts people's attention, which may result in traffic accidents. Also, smoking is prohibited in places such as gas stations and department stores, since they may cause fire even explosion. Avoiding some bad behaviors may save many people's life and therefore governments have already implemented lots of regulation on people's behaving improperly and they need to be detected in time. However, it is impossible to detect all these behaviors simply by human being. Fortunately, Machine learning and Computer vision is becoming more prevailing and can be used by humans. By studying the relationship between data, computers can develop the ability to classify the photos by itself. So, if some smoking and calling images can be put into computers for learning, they can be used to help detect the improper behaviors.

Machine learning developed significantly in different fields in recent years [1-3]. In the previous studies, there are some studies that have already tried to apply the machine learning into the field of computer vision about human. By designing a convolutional neural network, the computer managed to distinguish different human's behaviors [4]. What' s more, Zhu et al. also gave out an algorithm based on deep learning to monitor students' behaviors during the test [5]. In terms of smoking behavior detecting, Zhang et al. have developed a machine learning algorithm in the method of decision tree [6]. Their model achieved 84.11% accuracy with the best performance.

However, there are still few studies about the prediction of calling behaviors, especially applying algorithms based on Machine Learning methods. For instance, smoking, talking on the phone is hard to detect even by our naked eyes as well. The phone may be too small that is blocked by people's hand, thus making the problem more complicated. In [7], Zheng used Machine Learning algorithms based on Support Vector Machine (SVM) as well as Convolutional Neural Network (CNN) to predict people's walking upstairs and downstairs behavior, which achieved 93.5% as the highest accuracy. However, this paper would like to compare the mainstream machine learning algorithms in detecting the smoking and calling behaviors and figure out which one is the best solution to the problem.

The rest of this paper is divided into following parts: Part 2 shows the datasets this study picked and its problems to be solved method. Then visualization of each class and results of different machine learning algorithms will be presented in Part 3. Finally, the conclusion will be summarized in Part 4.

II. METHOD

A. Dataset description and preprocessing

The dataset this paper uses has three classes: Smoking class, Calling class and Normal class. For the Smoking class, this paper chooses 'Cigarette Smoker Detection' dataset from Kaggle, which has 805 images with different sizes [8]. For the Calling class, it includes 1,227 images with different sizes from TIANCHI DATA SET and 396 images from CSDN, whose sizes are 3456 × 4608 [9, 10]. The Normal class comes from 'Person Face Dataset' from Kaggle, containing 10,000 images of 1024×1024 [11]. The sample images are shown in Figure 1, Figure 2 and Figure 3.

Figure 1. Sample images in Smoking class

978-1-6654-9246-1/22/$31.00 ©2022 IEEE
DOI 10.1109/MLISE57402.2022.00087

Authorized licensed use limited to: Naga Bandla. Downloaded on January 20,2023 at 07:09:33 UTC from IEEE Xplore. Restrictions apply.

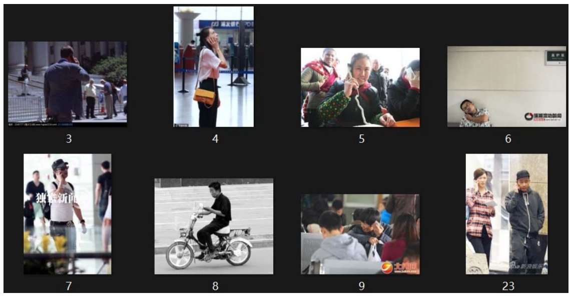

Figure 2. Images in Calling class from TIANCHI DATA SET

Figure 3. Sample images in Calling class from CSDN

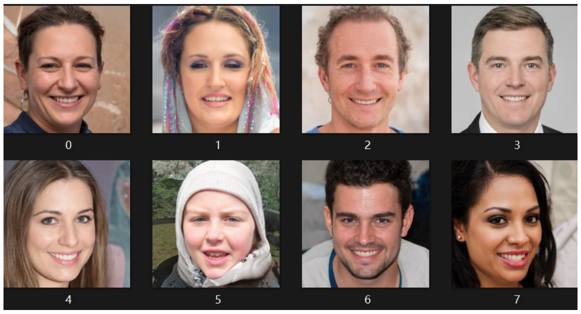

Figure 4. Sample images in Normal class

The preprocessing is consisted of six parts. First, the get_frontal_face_detector function from dlib is used to locate human's face in images. So, their behaviors such as talking on the phone can be detected in a better way. After that, all the images are resized into 64×64. In the third part, images are transformed into gray through the cvtColor function from cv2. In this way, they are more uniform for the machine learning. Then, in order to balance the dataset, about 700 preprocessed images are selected for each class because the Calling class only have around 700 images. What's more, this paper normalizes the dataset by dividing 255. Finally, the dataset is split into train and test parts, whose ratio of the training is 0.8. Figure 5, Figure 6 and Figure 7 shows the processed data.

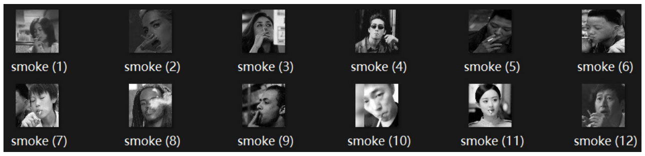

Figure 5. Preprocessed images in Smoking class

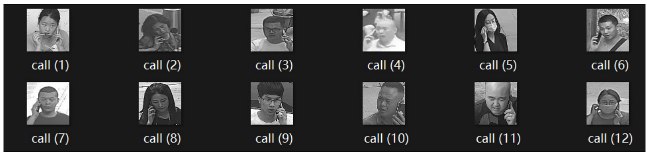

Figure 6. Preprocessed images in Calling class

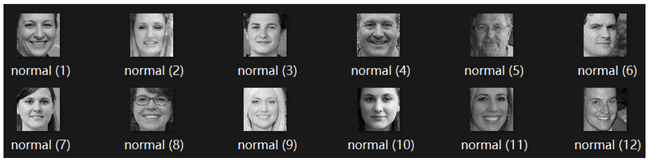

Figure 7. Preprocessed images in Normal class

B. Machine learning algorithms

This paper used several famous Machine Learning algorithms including Support Vector Machine, Decision tree, Random Forest, K-nearest Neighbors, K-Means. There are some introductions about these algorithms, which can be found below.

a) Support Vector Machine (SVM): SVM is a supervised machine learning algorithm that can be used to solve classification or regression problems. It aims at looking for a hyperplane in an N-dimensional space, which can classify the data points. In order to separate two data point from classes, there are a great number of hyperplanes to choose. So, SVM is designed to find the most suitable hyperplane that has the maximum distance between the two data points so that it can classify the data points better.

SVM has a lot of kernels to choose: 'linear', 'poly', 'rbf', 'sigmoid', 'precomputed'. The C in SVM is Regularization parameter and Gamma is the kernel coefficient for 'rbf', 'poly', 'sigmoid'.

This paper chooses LSVM--'linear' kernel and KSVM--'rbf' kernel to compare the models' ability in classifying the dataset. For rest of the parameters, they are default ones, whose C is 1 and Gamma is 'scale'.

b) Decision Tree: Decision Tree is a non-parametric supervised learning algorithm that can be used to solve classification or regression problems. It uses a model that looks like a tree which can show the decisions and the possible consequences. Each branch represents the outcome of the test, and each leaf node represents a class label.

c) Random Forest: Random Forest is an algorithm that can used to solve classification or regression problems by building a number of decision trees. For classification tasks, the outcome of the random forest is the class that are selected by most trees. For regression tasks, the outcome is the average prediction of each tree. The advanatge of Random Forest lies in avoiding decision trees from overfitting to their training set. The main parameter for the Random Forest is 'n_estimators', which is the number of the trees in the forest.

This paper uses a model whose 'n_estimators' is 250 to solve the problem.

Authorized licensed use limited to: Naga Bandla. Downloaded on January 20,2023 at 07:09:33 UTC from IEEE Xplore. Restrictions apply.

d) K-nearest Neighbors (KNN): KNN is a non-parametric supervised learning algorithm that can be used to solve classification or regression problems. In a classification case, a data point is classified to the class that is the most common among its k nearest neighbors. In a regression case, the output is the average of the values of the input's k nearest neighbors.

e) K-Means clustering: K-Means is a clustering algorithm that can be used to solve classification problems. It requires the number of clusters to be given specifically and repeats finding the center of the random k data points until it finds the best way to classify all the data points. The main parameter for the K-Means is 'n_clusters', which is the number of the class that dataset needs to be split into. It is required while designing the algorithm.

This paper uses 3 clusters and compares the result with the original labels of the dataset.

C. *More details*

a) Principal Component Analysis (PCA): PCA is a dimensionality reduction algorithm that can be used in exploratory data analysis and making predictive models. It looks for the principal components and using them to perform a change of basis on the data, sometimes using only the first few principal components and ignoring the rest.

b) Mean Squared Error (MSE): MSE is usually used as a loss function that can test the average of the squares of the errors, which is the average squared difference between the estimated values and the actual value. This paper uses MSE as a loss function to calculate the precision of the models.

c) Confusion Matrix: Confusion Matrix is a table that shows the performance of an algorithm. Each row represents the samples in an actual class and each column represents the samples in a predicted class. This paper uses the confusion matrix function from metrics in Python.

III. Result And Discussion

Confusion matrix and results of the different algorithms (i.e. KSVM, LSVM, Decision Tree, Random forest, KNN and K-means) are showed in Table I, Table II, Table III, Table IV, Table V and Table VI.

TABLE I. Confusion Matrix Of KSVM

CONFUSION MATRIX OF KSVM			
Predict ***Actual***	***Smoking***	***Calling***	***Normal***
Smoking	87	40	5
Calling	66	87	4
Normal	6	3	128

TABLE II. Confusion Matrix Of LSVM

CONFUSION MATRIX OF LSVM			
Predict ***Actual***	***Smoking***	***Calling***	***Normal***
Smoking	105	25	2
Calling	59	91	7
Normal	2	1	134

TABLE III. Confusion Matrix Of Decision Tree

CONFUSION MATRIX OF DECISION TREE			
Predict ***Actual***	***Smoking***	***Calling***	***Normal***
Smoking	94	31	7
Calling	48	92	17
Normal	15	6	116

TABLE IV. Confusion Matrix Of Random Forest

CONFUSION MATRIX OF RANDOM FOREST			
Predict ***Actual***	***Smoking***	***Calling***	***Normal***
Smoking	97	33	2
Calling	32	122	3
Normal	5	2	130

TABLE V. Confusion Matrix Of KNN

CONFUSION MATRIX OF KNN			
Predict ***Actual***	***Smoking***	***Calling***	***Normal***
Smoking	93	9	30
Calling	78	45	34
Normal	2	0	135

TABLE VI. Confusion Matrix Of K-Means

CONFUSION MATRIX OF K-Means			
Predict ***Actual***	***Smoking***	***Calling***	***Normal***
Smoking	33	67	32
Calling	38	74	45
Normal	109	3	25

TABLE VII. Results Of Different Algorithms

RESULTS OF DIFFERENT ALGORITHMS					
	Accuracy	***MSE***	***Precision***	***Recall***	***F1-score***
KSVM	0.71	0.369	0.72	0.71	0.71
LSVM	0.77	0.254	0.78	0.77	0.77
Decision Tree	0.71	0.446	0.71	0.71	0.71
Random Forest	0.82	0.230	0.82	0.82	0.82
KNN	0.64	0.585	0.69	0.64	0.60
K-Means	0.31	1.683	0.33	0.31	0.31

A. *Comparison in Accuracy*

In terms of accuracy, the results in Table VII showed that Linear Support Vector Machine, Kernel Support Vector Machine, Decision Tree, Random Forest performed well. Among them, Random Forest achieves the best result, with an accuracy of 82%. KNN is not bad, while K-means is the worst, only having an accuracy of 31%.

B. *Comparison in Mean Squared Error*

In terms of Mean square error, the results showed that Linear Support Vector Machine, Kernel Support Vector Machine, Decision Tree, Random Forest, KNN performed well, especially the Linear Support Vector Machine, only having a MSE with 0.254. In contrast, K-Means clustering has a MSE of 1.683, which is a quite terrible result.

Authorized licensed use limited to: Naga Bandla. Downloaded on January 20,2023 at 07:09:33 UTC from IEEE Xplore. Restrictions apply.

C. *Discussion*

The results in Accuracy and Mean Squared Error are consistent. Random Forest has the best performance on the task that this paper tries to solve, while K-Means is the worst.

The reason why K-means has such a bad result may lie in the fact that it is an unsupervised algorithm which is good at tackling the classification problems where photos have no label. However, supervised algorithms can actually perform better since all the photos are already well classified.

D. *Visualization of Principal Component Analysis*

Figure 8 shows the outcome of the Random Forest algorithm visualized in the mean of Principal Component Analysis, comparing with the split test part of the dataset, which is also showed in Figure 9. The picture shows the excellent performance of Random Forest directly that most photos are well classified.

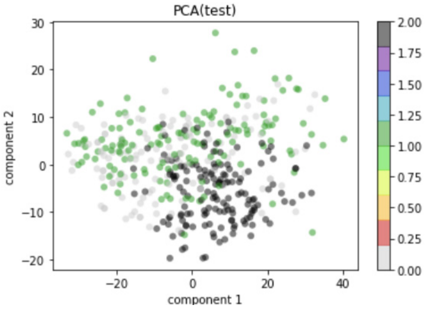

Figure 8. Visualization of the relationship between test_x and test_y.

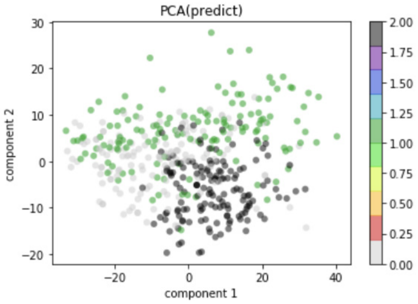

Figure 9. Visualization of the relationship between test_x and predicted_y by Random Forest.

IV. Conclusion

The goal of the study is to detect people's improper behaviors which may put others' life into danger by applying Machine Learning algorithms. This paper focus on studying the performance of different Machine Learning algorithms including Linear Support Vector Machine (LSVM), Kernel Support Vector Machine (KSVM), Decision tree, Random Forest, K-nearest Neighbors (KNN) and K-Means Clustering. Confusion Matrix and Mean Squared Error are applied to help judge whether the model is good or not. Additionally, Principal Component Analysis visualizes the outcome of the best algorithm. The results of the study show that Random Forest is the most suitable method for the problem, while K-Means clustering is the worst. In the future, applications that can be applied into detecting people's behaviors through camera will be developed and the model of Random Forest can work better by adjusting the parameters.

References

[1] Q. Zhou, W. Lan, Y. Zhou and G. Mo, "Effectiveness Evaluation of Anti-bird Devices based on Random Forest Algorithm," 2020 7th International Conference on Information, Cybernetics, and Computational Social Systems (ICCSS), 2020, pp. 743-748.

[2] Y. Guo, Y. Zhou, X. Hu and W. Cheng, "Research on Recommendation of Insurance Products Based on Random Forest," 2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), 2019, pp. 308-311.

[3] Y. Qiu, P. Chen, Z. Lin, Y. Yang, L. Zeng and Y. Fan, "Clustering Analysis for Silent Telecom Customers Based on K-means++," 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), 2020, pp. 1023-1027.

[4] S. Zheng, "Research on Algorithm of Pedestrain Attitude Estimation and Recognition Based on Machine Learning," Shandong University, 2019, .DOI:10.27272/d.cnki.gshdu.2019.000463.

[5] G. Zhu, X. Jiang, F. Xu, "Application of video behavior and action recognition based on machine learning in paperless assessment (in Chinese)," Construction informatization in China, 2019, vol. 10, pp. 56-57.

[6] Y. Zhang, J. Liu, Z. Zhang and J. Huang. "Prediction of daily smoking behavior based on decision tree machine learning algorithm." In 2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC), pp. 330-333. IEEE, 2019.

[7] Y. Zheng, "Research on Machine Learning Algorithm for Human Behavior Recognition," Wuhan University of Technology, 2019. The DOI: 10.27381 /, dc nki. Gwlgu. 2019.000606.

[8] "Cigaretter-somker-detection", kaggle, https://www.kaggle.com/datasets/vitaminc/cigarette-smoker-detection, 2019.

[9] "Smoking and Calling Image Dataset", Tianchi, https://tianchi.aliyun.com/dataset/dataDetail?dataId=89271, 2022.

[10] "Calling dataset", CSDN. https://download.csdn.net/download/weixin_45546050/77837488, 2022.

[11] "Person-face-dataset", kaggle, https://www.kaggle.com/datasets/almightyj/person-face-dataset-thispersondoesnotexist, 2022.

Authorized licensed use limited to: Naga Bandla. Downloaded on January 20,2023 at 07:09:33 UTC from IEEE Xplore. Restrictions apply.