Machine Learning Based People's Anomalous Human Behaviour Forecasting

[1]Tummuri Durga Deekshitha, [2]S. Aruna,

[1]MCA Student, Dept. Of MCA, Swarnandhra College of Engineering and Technology, Seetharampuram, Narsapur, Andhra Pradesh 534280,

happideekshi95@gmail.com

[2]Assistant Professor, Dept. Of MCA,Swarnandhra College of Engineering and Technology, Seetharampuram, Narsapur, Andhra Pradesh 534280,

Abstract:*Some improper behaviour in specific situations mayput people in danger, such as smoking in a gas station; thereforethey need to be detected. This paper tries to find out the bestMachine Learning algorithm to address that kind of predictionproblems.Datasetsrelatedtobehaviourdetectionarecollected,whosecategoriesconsistsofsmoking,callingandnormalbehaviours. Experiments based on several famous algorithms areconducted, including Linear Support Vector Machine (LSVM),KernelSupportVectorMachine(KSVM),DecisionTreeClassifier(DT),RandomForestClassifier(RF),K-nearestNeighbours(KNN)andK-MeansClustering.Additionally,Confusion Matrix and Mean Squared Error (MSE) are used tojudgetheperformanceofeachalgorithm.Finally,PrincipalComponentAnalysis(PCA)visualizes theoutcomeofthebestalgorithm. The results show that Random Forest Classifier (RF)achievesthebestperformanceandiscapableofpredictingpeople'sabnormalbehaviourswith anaccuracyof 82%.*

Keywords-component; Machine Learning; Abnormal behavioursprediction; Dimensionalityreduction

I INTRODUCTION

Nowadays, people are paying more attention to their health,but there are still a lot of dangerous behaviours that may getpeople injured. They are extremely threatening in some specificsituations. For example, talking on the phone while drivingdistractspeople'sattention,whichmayresultintrafficaccidents? Also, smoking is prohibited in places such as gasstations and department stores, since they may cause fire evenexplosion.Avoidingsomebadbehaviour maysavemanypeople'slifeandthereforegove

ISSN: 2366-1313

rnmentshavealreadyimplementedlotsofregulationonpeople'sbehavingimproperly and they need to be detected in time. However, it isimpossibletodetectallthesebehavioursimplybyhumanbeing.Fortunately,MachinelearningandComputervisionisbecoming more prevailing and can be used by humans. Bystudying the relationship between data, computers can developthe ability to classify the photos by itself. So, if some smokingand calling images can be put into computers for learning, theycanbeusedtohelpdetecttheimproperbehaviours.

Machine learning developed significantly in different fieldsin recent years [1-3]. In the previous studies, there are somestudies that have already tried to apply the machine learninginto the field of computer vision about human. By designing aconvolutionneuralnetwork,thecomputermanagedtodistinguish different human's behaviours [4]. What's more, Zhuet al. also gave out an algorithm based on deep learning tomonitorstudents'behavioursduringthetest [5].Intermsof smoking behaviour detecting, Zhang et al. have developed amachine learning algorithm in the method of decision tree [6].Theirmodelachieved84.11%accuracywiththebestperformance. However, there are still few studies about the prediction ofcallingbehaviours,especiallyapplyingalgorithmsbasedonMachine Learning methods. For instance, smoking, talking onthe phone is hard to detect even by our naked eyes as well. Thephone may be too small that is blocked by people's hand, thusmaking the problem more complicated. In [7], Zheng usedMachineLearningalgorithmsbasedonSupportVectorMachine(SVM)aswellasConvolutionNeuralNetwork(CNN)topredictpeople'swalkingupstairsanddownstairsbehaviour,whichachieved93.5%asthehighestaccuracy.However, this paper would like to compare the mainstreammachinelearningalgorithmsindetectingthesmokingandcalling behaviours and figure out which one is the best solutiontotheproblem.

II METHOD

Datasetdescriptionandpre-processing

Thedatasetthispaperuseshasthreeclasses:Smokingclass,Calling class and Normal class. For the Smoking class, thispaperchooses'CigaretteSmokerDetection'datasetfromKaggle, which has 805 images with different sizes [8]. For theCalling class, it includes 1,227 images with different sizes fromTIANCHI DATA SET and 396 images from CSDN,

ISSN: 2366-1313

whosesizes are 3456 × 4608 [9, 10]. The Normal class comes from‘PersonFaceDataset'fromKaggle,containing10,000imagesof 1024×1024 [11]. The sample images are shown in Figure 1,Figure2andFigure3.

Figure1.SampleimagesinSmokingclass

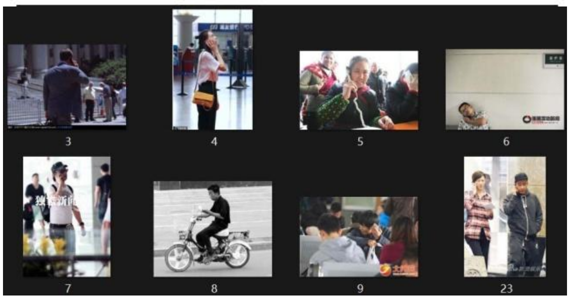


Figure2.ImagesincallingclassfromTIANCHIDATASET

Figure 3. Sample images in calling class from CSDN

Thepre-processingisconsistedofsixparts.First,theget_frontal_face_detector function from dib is used to locatehuman's face in images. So, their behaviours such as talking onthe phone can be detected in a better way. After that, all theimages are resized into 64×64. In the third part, images aretransformed 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 pre-processedimages are selected for each class because the Calling classonlyhavearound700images.What'smore,thispapernormalizes the dataset by dividing 255. Finally, the dataset issplit into train and test parts, whose ratio of the training is 0.8.Figure5,Figure6andFigure7showstheprocesseddata.

Figure 5. Preprocessed images in Smoking class

Machinelearningalgorithms

ThispaperusedseveralfamousMachineLearningalgorithms including Support Vector Machine, Decision tree,RandomForest,K-nearestNeighbours,K-Means.Therearesome introductions about these algorithms, which can be foundbelow.

Support Vector Machine (SVM): SVM is a supervisedmachinelearningalgorithmthatcanbeusedtosolveclassification or regression problems. It aims at looking for ahyper plane in an N-dimensional space, which can classify thedata points. In order to separate two data point from classes,there are a great number of hyper planes to choose. So, SVM isdesignedtofindthemostsuitablehyper planethathasthemaximum distance between the two data points so that it canclassifythe datapointsbetter.

SVM has a lot of kernels to choose: ‘linear', ‘poly', ‘rbf',‘sigmoid', ‘recomputed'. The C in SVM is Regularizationparameter 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 thedataset.Forrestoftheparameters,theyared efaultones,whoseCis1andGammais‘scale'.

DecisionTree:DecisionTreeisanon-parametricsupervisedlearningalgorithmthat canbeusedtosolveclassificationorregression problems.Itusesamodelthatlooks like a tree which can show the decisions and the possibleconsequences. Each branch represents the outcome of the test,andeachleaf noderepresents a class label.

Random Forest: Random Forest is an algorithm thatcanusedtosolveclassificationorregressio nproblemsbybuildinganumberofdecisiontre es.Forclassificationtasks,the outcome of the random forest is the class that are selectedby most trees. For regression tasks, the outcome is the averagepredictionofeachtree.Theadvantage ofRandomForestliesin avoiding decision trees from over fitting to their training set.The main parameter for the Random Forest is ‘n_estimators',whichis thenumberofthetreesinthe forest.

This paper uses a model whose ‘n_estimators' is 250 tosolvetheproblem.

III RESULTANDDISCUSSION

Confusionmatrixandresultsofthedifferental gorithms(i.e.KSVM, LSVM, Decision Tree, Random forest, KNN and K-means) are showed in Table I, Table II, Table III, Table IV,TableVandTableVI.

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

ComparisoninAccuracy

In terms of accuracy, the results in Table VII showed thatLinearSupportVectorMachine,KernelSupportVectorMachine,DecisionTree,RandomForestperformedwell.Among them, Random Forest achieves the best result, with anaccuracy of 82%. KNN is not bad, while K-means is the worst,onlyhavinganaccuracyof31%.

ComparisoninMeanSquaredError

Interms of Meansquare error,the resultsshowedthatLinearSupportVectorMachine,KernelSupportVectorMachine, Decision Tree, Random Forest, KNN performed well,especially the Linear Support Vector Machine, only having aMSE with 0.254. In contrast, K-Means clustering has a MSE of1.683,whichisaquiteterribleresult.

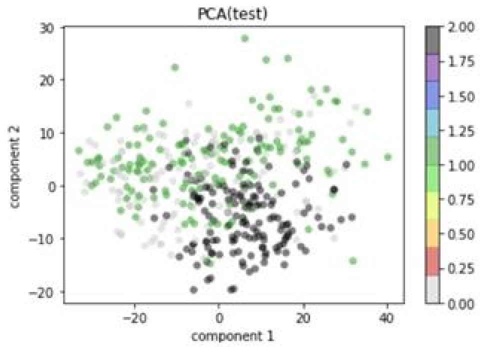
Discussion

TheresultsinAccuracyandMeanSquaredErrorareconsistent.RandomForesthasthebestperformanceonthetaskthatthispapertriestosolve,whileK-Meansistheworst.

The reason why K-means has such a bad result may lie inthe fact that it is an unsupervised algorithm which is good attacklingtheclassificationproblemswherephotoshavenolabel.However, supervised algorithms can actually perform bettersinceallthephotosarealreadywellclassified.

VisualizationofPrincipalComponentAnalysis

Figure8showstheoutcomeoftheRandomForestalgorithmvisualizedinthemeanofPrincipalComponentAnalysis,comparingwiththesplittestpartofthedataset,whichisalsoshowedinFigure9.Thepictureshowstheexcellent performance of Random Forest directly that mostphotosarewellclassified.



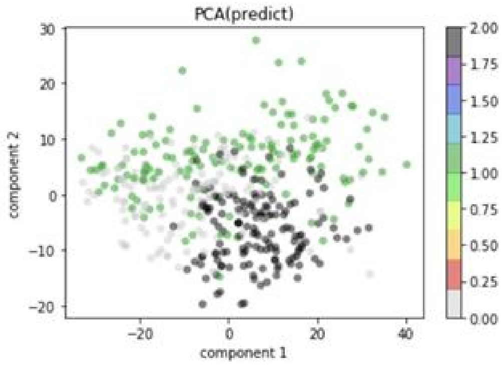


Figure 8.Visualizationoftherelationshipbetween test_xandtest_y.

VII CONCLUSION

Thegoalofthestudyistodetectpeople'simproperbehaviourswhichmayputothers'lifeintodangerbyapplying MachineLearningalgorithms.Thispaperfocusonstudyingtheperformanceofdifferent MachineLearningalgorithmsincludingLinearSupportVectorMachine(LSVM),KernelSupportVectorMachine(KSVM),Decisiontree,RandomForest, K-nearest Neighbours (KNN) and K-Means Clustering.Confusion Matrix and Mean Squared Error are applied to helpjudge whether the model is good or not. Additionally, PrincipalComponentAnalysisvisualizesthe outcomeofthebestalgorithm. The results of the study show that Random Forest isthemostsuitablemethodfortheproblem,whileK-Meansclustering is the worst. In the future, applications that can beapplied into detecting people's behaviours through camera willbe developed and the model of Random Forest can work betterbyadjustingtheparameters.

REFERENCES

1.Q. Zhou, W. Lan, Y. Zhou and G. Mo, "Effectiveness Evaluation ofAnti-birdDevicesbasedonRandomForestAlgorithm," 20207thInternationalConferenceonInformation,Cybernetics,andComputationalSocialSystems(ICCSS),2020,pp.743-748.

2. Y. Guo, Y.Zhou, X. Hu andW. Cheng, "Researchon RecommendationofInsuranceProductsBasedonRandomForest," 2019InternationalConference 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,

"ClusteringAnalysisforSilentTelecomCustomersBasedonK-means++," 2020IEEE4thInformationTechnology,Networking,ElectronicandAutomationControlConference(ITNEC),2020,pp.1023-1027.

4. S. Zheng, “Research on Algorithm of Pedestrian Attitude Estimation andRecognitionBasedonMachineLearning, ”ShandongUniversity,2019,.DOI:10.27272/d.cnki.gshdu.2019.000463.

5. G.Zhu,X.Jiang,F.Xu,“Applicationofvideobehaviourandactionrecognitionbasedonmachinelearninginpaperlessassessment(inChinese),” Construction informatization in China, 2019, vol. 10, pp. 56-57.

6. Y. Zhang, J. Liu, Z. Zhang and J. Huang. "Prediction of daily smokingbehaviour based on decision tree machine learning algorithm." In 2019IEEE9thInternationalConferenceonElectronicsInformationandEmergencyCommunication(ICEIEC),pp.330-333.IEEE,2019.

7. Y.Zheng,“ResearchonMachineLearningAlgorithmforHumanBehaviourRecognition,” WuhanUniversityofTechnology,2019.TheDOI: 10.27381/,dcnki.Gwlgu.2019.000606.

8. Prasadu Peddi (2015) "A review of the academic achievement of students utilising large-scale data analysis", ISSN: 2057-5688, Vol 7, Issue 1, pp: 28-35

9. Prasadu Peddi (2015) "A machine learning method intended to predict a student's academic achievement", ISSN: 2366-1313, Vol 1, issue 2, pp:23-37.