Recognizing Humans from Their Behavioral Patterns

Sonia
Department of Computer Science
Indian Institute of Technology Guwahati
Assam, India 781039
Email: s.sonia@iitg.ernet.in

***Abstract*—Authentication of human in an unobtrusive manner is important in modern technology to enrich society. Authentication of a person in an environment like smart home, car driving,etc., can lead to better man-machine conjugation. Advancement in sensor technology and machine learning makes it possible to obtain and analyze a significant amount of data. This paper explores the important aspects of human behavior daily routine in the home. Traditional approaches rely on the extraction of statistical features for machine learning algorithms. However, with an increased number of sensors, feature extraction may not be possible. Therefore, human identification in a smart home is performed using deep learning technique on the data obtained from multiple sensors placed at the various place and devices in the homes like a bed, bathroom, door. In this paper, both statistical, as well as deep neural networks, are implemented for human activity data. For humans in smart home deep learning has upper hand as compared to other machine learning techniques. Thus proposed approach can identify humans based on analysis of sensor data using various machine learning technique using unobtrusive sensory data.**

I. Introduction

In this era of the internet of things, environments are becoming smart to improve the living standards of human beings. The smartness in the living space of human beings has proved its advantage in several domains such as elderly care, traffic management, intruder detection, automation of devices in smart homes,etc. As these smart systems intend to add benefits to the humans, who are an integral part of the environments, therefore, these smart systems perform better and become more intelligent if these are familiar with the human behavior. To analyze human behavior human can be assumed as a machine with multiple different states in his brain. Over the time state transitions happen. Behavior modeling is to model transition by using machine learning based model or mathematical statistics. For a machine, to behave according to individual user specification, it should be able to determine the current active state of human and predict the next state. Enev et al. [1] investigate, "the potential to identify individuals" using multi-sensor data. In their work, it is shown that drivers can be uniquely identified with an accuracy of 87 percent (99 percent with top 5 sensors). For different purposes, human behavior has been studied by researchers in different environments. In [2], human behavior pattern is studied for abnormality detection in human behavior in the smart home environment by using the longest common subsequence algorithm. Furthermore, the wellness determination process as presented by authors is a novel framework which verifies the behavior of elderly at three different stages of daily living (usage of appliances, activity recognition and forecast levels) in a smart home monitoring environment. Similarly, for elderly care, the behavior of a person is analyzed with the help of wireless sensors embedded in their living space by authors in [3]. To monitor the health status of an elderly person, the similar work is performed by authors in [4]. However, authentication is a different problem than identification. There are cases where fake identification is used in an illegal manner such as many bad drivers assume the phony identity of the good driver to impersonate for insurance purpose; fake identity is used to occupy someone else living space, etc. Thus we propose a behavioral biometric which authenticate a person based on his/her natural behavior. Such a method helps to prevent fake identify. In the modern world, many advancements are made in machine learning. The same approach can lead to better authentication if authentication is treated as a two-class classification (i.e., yes and no) for each subject. For the elongated analysis of behavioral authentication data from more number of sensors embedded in astute homes is analyzed. Data is received from intellective homes. As the data amount, in this case, is immensely high because of the increased number of sensors, so instead of statistical analysis recurrent Neural Network (RNN) is utilized for behavioral analysis in smart homes.

Some advantages of the proposed method for authentication are:-

- Privacy preserving; no credentials or image is taken.
- Very minimal infrastructure and implementation cost.
- Requires no active participation of the user.

Several classifiers (i.e., machine learning methods) have endeavored, and performance comparison is presented. Results show the proficiency of the approach utilized.

II. Data and Basic Analysis

For the experimental purpose and to monitor the health of elders multiple sensors are embedded in the 50 old age homes. Sensors embedded in the smart house are non-intrusive for the privacy concerns of the residents. There are 15 different

978-1-5386-9151-9/19/$31.00 ©2019 IEEE

sensors used to serve the purpose of activity monitoring. Total six months of data are collected for 50 subjects. To collect the data in smart homes, sensors are connected to the Arduino board and via Arduino data is getting buffered in the gateway for further processing. Along with sensory value, timestamp, sensor status and position of the sensor are also stored. Data is annotated based on the sensor response, sensor position, and sensor status. Timestamp includes the data and time of the data buffered. Sensor response gives the output of sensor at a given instance of time. Sensor Position gives the information about the physical position of the sensor. Sensor status gives the data whether the sensor is in working condition or not?

A. Data Representation

In smart houses, sensors like PIR sensors, light sensor, etc. placed at different places of a home gives output 0 and 1. Therefore, for the further analysis of raw data buffered from sensors(embedded in houses) needs to be presented in a structured way. To accomplish the task a 24 hrs routine of a subject, in terms of buffered sensory data, a vector is defined. To explain the vector, let $S_1, S_2, ...S_N$ represents the data from N number of sensors and $t_1, t_2, ...t_{86400}$ represents time in seconds. Now, the vector can be represented as $V_x^i = [(t_1, (S_1, .., s_N)), (t_2, (S_1, .., S_N)), ..., (t_{86400}, (S_1, .., s_N))]$ where, V_x represents the routine of the x^{th} day of the i^{th} subject and $x\epsilon(1, 2, 3.....), i\epsilon(1, 2, ..., 50)$. Now, to statistically explore the dataset, multiple features were extracted from the data. For everyday data, the values of different features are computed. Features such as mean time of activity, the median of the activity time, skewness, standard deviation, max, min, the total duration of activity of each activity for every hour are calculated. Each of these 'feature' is important for human authentication problem. But all features are not important for every human. Thus we have multiple features for each human. All these features are analyzed further towards human authentication. K fold cross-validation validates the proposed method for authentication with $k = 10$. Robustness is confirmed by examining sensitivity, specificity, and ROC (Receiver operating characteristic) analysis; results are given in section methods and results. We have nine months of data. Of these, initial seven months data is used for analysis and classifier selection, methodology validation. Remaining two months data are kept separately for further validation of final model (i.e., feature and classifier). The final model is validated on last two-month dataset. Entire feature computation can be done in the cloud or in the device itself.

III. Method and Results

In this section, the issue resolved is authentically verifying a person's identity. To test the authenticity of the claim following question is answered:
Is the person present in the house is 'A' or not?

For the experimentation, dataset consisting (initial seven months data) of the equal mix of routines of A and routines of others is considered. The model validated using ten-fold cross-validation. It is to be noted that this is not a biometric identification system. In real life scenario, it is highly probable that occasional impersonation (where the identities of different humans are switched) takes place. Thus for a caregiver, it is important to know whether the reported person is authenticated or not. Initially, it is hypothesized that person-specific features (for every person) can lead to better results and that personalized features can be used to define his/her natural living style. Also, those features can serve as a distinguishing style for that person. Towards that goal, we tried to create a personalized feature set for every person. For feature ranking purpose, R statistical software is used in the training data. To rank these features, Boruta package [7] is used. 'Boruta' gives VIM (variable importance measure) for each of the features - corresponding to the given person and also accepts (or rejects) a feature for a classification problem. Using Boruta, the most significant features are selected for each person. It is then found that every human has a different set of notable features; here onwards referred to as personalized feature set. For the next phase of an investigation, only the customized feature set is used for each person. These features define parameters of importance for the respective person which distinguishes him/her from the rest. For each person, the cardinality of personalized feature set varies. A maximum of three features is selected for person H006 whereas for H004 only five features are selected. The average number of selected feature is six with a standard deviation of 11. Different classifiers are tested to authenticate the person. For one person each classifier is trained on initial seven months data. Total 50 persons data is present. For each person, ten-fold cross-validation for different classifiers was performed. Classifiers are used to identify if the same person is living in the house or not; for any given trip. In the present analysis, sensitivity and specificity [8] for each person is also measured along with overall accuracy. Process for selecting a classifier is simple and can be explained as follows: Perform analysis of initial seven months data, referred as past dataset, Overall data is nine months, Compare accuracy, sensitivity, and specificity to choose the best classifier, Then validate the robustness of chosen classifier by AUC analysis (ROC (Receiver operating characteristic) [36]) For every person, model creation process includes validation of the model with k fold cross-validation on training data. Finally, a classifier is chosen which maximized accuracy, sensitivity, and specificity for ten-fold cross-validation. For each person, a separate model is obtained for each classifier. In this work, 11 different classifiers such as SVM (Support vector machine with Sequential minimal optimization is used), (RF)Random Forest [9], SGD (Stochastic gradient descent) [10], PART [11], MLP (Multilayer Perceptron) [12], J48, KNN (i.e. IBK: k-nearest neighbors algorithm) [13], (DT) Decision Table [14], (DS) Decision Stump, (NB) Naive Bayes [15] and AdaBoost are selected to classify a routine to a person or not. A decision stump is a machine learning model consisting of a one-level decision tree. Table I provides median and standard deviation of accuracy for different classifiers. AdaBoost is the second best classifier; SVM, SGD, MLP, and Adaboost perform almost equally well. Each of the obtained median

TABLE I
ACCURACY FOR DIFFERENT CLASSIFIERS WITH PERSON-SPECIFIC FEATURE SET

Classifier	Median Accuracy Accuracy	Standard Deviation
SVM	71.82	7.5
SGD	77.66	8.7
Random Forest	92.01	12.2
PART	78.67	8.2
NaiveBayes	74.24	6.7
MLP	75.24	7.3
J48	85.99	9.5
IBK (kNN)	76.77	8.7
Decision Table	74.52	13.1
Decision Stump	74.37	.9
AdaBoost	79.84	8.1

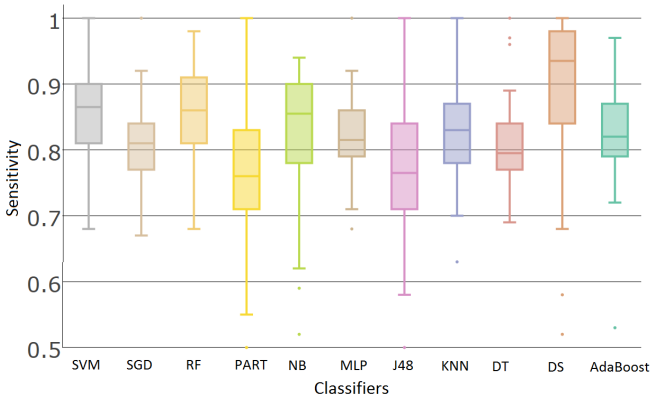

Fig. 1. Feature computation from sensor data for Different Classifiers Obtained For 50 humans. Y Axis Shows the Value of Obtained Sensitivity in Percentage and X Axis Shows the Algorithm Used

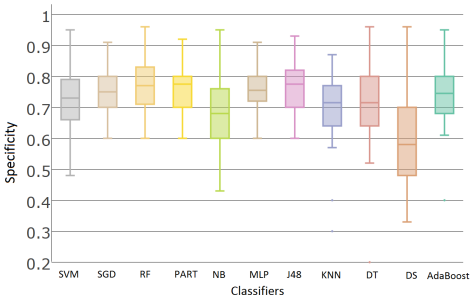

Fig. 2. Specificity for Different Classifiers.

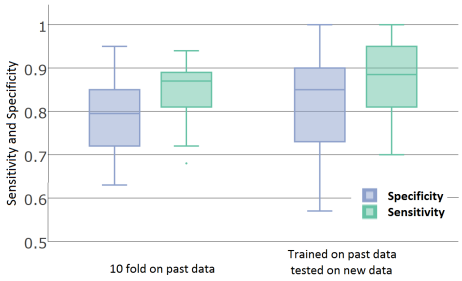

Fig. 3. Specificity and Sensitivity for 10 Fold Cross Validation and on Test data for Random Forest

accuracies is higher than 70%, thereby demonstrating the effectiveness of custom (i.e., person specific personalized) feature set. From Table I, it is clear that Random Forest is suitable for authentication problem. To validate this further, we checked sensitivity and specificity for all these classifiers. For person authentication problem, apart from accuracy figure, sensitivity and specificity should be equally high. Figure 1 shows sensitivity values for different classifiers used to authenticate persons. Random Forest gives a median sensitivity of 0.86. Decision Stump shows the best sensitivity of 0.94 and SVM performs well in comparison to Random Forest. But we need to check specificity also for Decision Stump and SVM.

Figure 2 shows specificity values for different classifiers. Although Decision stump showed the best sensitivity; it has worst specificity at 0.58. Also for both cases, it shows larger variation across humans, therefore, rendering it as unsuitable. In this regard, however, Random Forest performs quite well. Also from Figure 2 it is clear that for human authentication, specificity values for the classifiers are slightly less compared to respective sensitivity.

Now to finally check the validity of the proposed personalized feature set, we trained our models on past data and checked how it performs on new data using Random Forest. Past data comprises almost 80 percent of total data (1st seven months) and new trip data (last two months) constitutes 20 percent of overall collected data. Figure 3 shows a comparison of sensitivity and specificity for Random Forest applied to new data and ten fold cross-validation. It is seen that sensitivity and specificity improves on test sets compared to 10 fold cross validation. The same is true for accuracy. Median accuracy for 10 fold is 92.01 % whereas on test data median accuracy obtained is 94%. In real life, training data availability may be less. Thus, it is important to see how the proposed system will work with less amount of training data. Hence, authors trained model (Random forest with the personalized feature) on 20 percent data(1st two months) and tested the accuracy on remaining 80 percent data(last seven months); to investigate the robustness. It is found that the results are promising. As per results, sensitivity is higher than specificity with median (of both) crossing 0.75. Even with limited training data, the median accuracy crosses 75 percent. Thus the method is found to be robust, reliable and accurate for human authentication. Even if the analysis is done with less amount of data very good predictions can be done. Median is chosen as a proper representation as it is a robust measure of central tendency and less prone to outliers that mean [16].

TABLE II
ACCURACY SENSITIVITY AND SPECIFICITY MEASURE AND VARIATION FOR CONDUCTED EXPERIMENT

Experiment Id	Sensitivity		Specificity		Accuracy (%)	
	Median	Standard Deviation	Median	Standard Deviation	Median	Standard Deviation
i	0.86	0.122	0.76	0.12	92.01	12.02
ii	0.94	0.09	0.85	0.17	94.00	11.23
iii	0.	0.19	0.76	0.13	76.21	8.32

TABLE III
ACCURACY FOR DIFFERENT CLASSIFIERS WITH IN-HOUSE SENSORY DATA

Classifier	Accuracy
One-class SVM	72.52
Multi-class SVM	76.77
Random Forest	64.01
HMM	80.52
LSTM	93.47

A. Without Feature Extraction

However, statistical methods are dependent on feature selection which is a hit and trial method. In this section, two approaches are explored for person authentication and person identification. For person identification 2-layer LSTM [17] is utilized. For the training of LSTM, data of 9 months is inadequate to train the build deep learning-based system. Consequently, a new dataset is generated based on the currently available data. As, data accumulated emanates from December to May, therefore, considering the fact of different sunrise and sunset timings over the year, the age of the targets and global warming incipient data set is generated with a continuous shift of 30 seconds in the daily activity routine or 0.005-degree change of temperature. This way, generated data set is of 150 months ascertaining the distributed arbitrariness in the data to avoid the biasing. Training of the system is performed in a batch-wise manner, and cross-validation is performed to evade the over-fitting of the data. The output layer of the LSTM predicts the class of the given input vector. Testing is performed one versus all and all versus all. For all versus all, as data is available for 50 targets (human beings) therefore, 50 classes are defined. The second approach used for comparison is Support vector machine(SVM). SVM model is trained in two ways for the comparison of the various methods. One support vector is implemented for one class classification which is utilized for one versus all testing and second SVM model is applied for 50 classes which are being used for multi-class testing. Other approaches used for comparisons are random forest and HMM which are commonly used for smart homes for activity prediction and classification. Unlike statistical methods, these methods take vectors (defined in the data representation) as an input. Results of the different approaches are calculated for the comparisons of their performances as shown in table III. Results are shown regarding accuracy which is calculated by using equation 1.

$$Accuracy = (TP + TN)/(TP + FP + TN + FN) \quad (1)$$

where, TP = true positives, FP = false positives, TN = true negatives, FN = false negatives As shown in table III LSTM gives better accuracy as compare to other techniques such as SVM, random forest and HMM.

IV. CONCLUSIONS

Thus using our proposed method can authenticate a human successfully. Also, further investigation is needed in future for exact identification. Also, this method can work with minimal sensing and without the much-dedicated sensor. Deployment is very easy and scalable, and only non-imaging sensory data logging is needed for effective implementation. Using a modern smartphone or any other embedded device placed in the proposed system can work without manual intervention. For future authentication can be improved further by adding another level of authentication on top of this method. This leads to the identification of frauds or impersonators effectively. Also, such classification quantifies a human for natural behavior. However, in smart home behavior analysis can be further extended to monitor the individual health status.

REFERENCES

[1] M. Enev, A. Takakuwa, K. Koscher, and T. Kohno, "Automobile driver fingerprinting," *Proceedings on Privacy Enhancing Technologies*, vol. 2016, no. 1, pp. 34–50, 2016.

[2] K. Park, Y. Lin, V. Metsis, Z. Le, and F. Makedon, "Abnormal human behavioral pattern detection in assisted living environments," in *Proceedings of the 3rd International Conference on PErvasive Technologies Related to Assistive Environments*. ACM, 2010, p. 9.

[3] N. K. Suryadevara, S. C. Mukhopadhyay, R. Wang, and R. Rayudu, "Forecasting the behavior of an elderly using wireless sensors data in a smart home," *Engineering Applications of Artificial Intelligence*, vol. 26, no. 10, pp. 2641–2652, 2013.

[4] M. T. Moutacalli, A. Bouzouane, and B. Bouchard, "The behavioral profiling based on times series forecasting for smart homes assistance," *Journal of Ambient Intelligence and Humanized Computing*, vol. 6, no. 5, pp. 647–659, 2015.

[5] M. Phelan, "Driver authentication system and method for monitoring and controlling vehicle usage," Jun. 2 2015, uS Patent 9,045,101.

[6] S. Ota, "Apparatus for authenticating vehicle driver," Mar. 20 2006, uS Patent App. 11/384,678.

[7] M. B. Kursa, W. R. Rudnicki *et al.*, "Feature selection with the boruta package," *J Stat Softw*, vol. 36, no. 11, pp. 1–13, 2010.

[8] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," 2011.

[9] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE transactions on pattern analysis and machine intelligence*, vol. 20, no. 8, pp. 832–844, 1998.

[10] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*. Springer, 2010, pp. 177–186.

[11] E. Frank and I. H. Witten, "Generating accurate rule sets without global optimization," 1998.

[12] D. W. Ruck, S. K. Rogers, M. Kabrisky, M. E. Oxley, and B. W. Suter, "The multilayer perceptron as an approximation to a bayes optimal discriminant function," *IEEE Transactions on Neural Networks*, vol. 1, no. 4, pp. 296–298, 1990.

[13] B. Lantz, *Machine learning with R*. Packt Publishing Ltd, 2013.

[14] R. Kohavi, "The power of decision tables," *Machine learning: ECML-95*, pp. 174–189, 1995.

[15] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[16] C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata, "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median," *Journal of Experimental Social Psychology*, vol. 49, no. 4, pp. 764–766, 2013.

[17] A. Graves, S. Fernández, and J. Schmidhuber, "Bidirectional lstm networks for improved phoneme classification and recognition," *Artificial Neural Networks: Formal Models and Their Applications–ICANN 2005*, pp. 753–753, 2005.