Vol.2019-DPS-178 No.20
Vol.2019-CSEC-84 No.20
2019/3/4

Towards a Semantic-based Mobile Health Monitoring Mechanism

Sigdel Shree Ram, Bernady Apduhan

Graduate School of Information Science
Kyushu Sangyo University
Fukuoka, 813-8503, Japan
k18gjk01@st., bob@is.(kyusan-u.ac.jp)

Abstract: The increasing population of aging people and the lack of human resources have created great challenges in society. Here, we consider machine learning in edge computing and semantic technologies to detect and improve the predictability in mobile health monitoring of human activity. With multi-modal sensor data, we conducted pre-processing to sanitize the data and extracted the ECG data. We used and compare the performance of random forests and SVM machine learning algorithms to identify the patterns of body activity. We achieved approximately 95% accuracy with random forest which was better than SVM, at 93%. While observing the confusion matrix we were able to identify the majority of mismatched data belonging to initial value of sensors while recording a particular activity. The preliminary experiments provided promising results and insights on the data semantization process to improve the prediction accuracy of human activity.

Keywords: Edge Computing, Edge Intelligence, Mobile Health, Machine Learning, Random Forest, Support Vector Machine, Personalized Health Monitoring, Semantization on Edge

1. Introduction

The large number of IoT (Internet of Things) devices for a certain application will generate a huge volume of data, so-called Big Data. Even though Big Data already existed before IoT came out, it was not until this time that activities to unleash and harness the hidden wealth of this huge data which will likewise generate a new breadth of knowledge were given much attention.

Typically, Big Data processing is done on the cloud, but the large communication overhead to transfer data from the IoT devices to/from a data center has raised crucial issues especially for applications which requires near-real time responses. Specifically, the problem of cloud centric computing, i.e., narrow bandwidth, long latency, intermittent connection, delay in decision making, etc., paves the way of edge computing as a new revolutionary way of correcting the aforementioned deficiencies. Edge computing is not a replacement of cloud computing, but rather as a supplement. That is, much of the processing maybe be done at the edge nodes which are close in distance to the edge devices or source, but when the processing capacity of edge nodes may not be sufficient at certain times, some or part of the application processing can be offloaded to the data center. The short data transfer rate between the source (IoT devices) and edge node, can be considered as part of the solution for applications requiring near-real time responses. However, the limited processing and storage capacities of edge nodes are hindrances to processing requirements.

In this paper, considering the above-mentioned processing and storage limitations of edge nodes, we study the use and performance of machine learning techniques to augment the data classification accuracy, in an effort to establish our platform to develop a semantic-based mobile health monitoring system.

The paper is organized as follows. Section 2 describe our research objectives, and section 3 cites some related researches. Furthermore, section 4 describe our experiment environment and the datasets, whereas section 5 describe our system architecture. The experiment methods are explained in section 6, and section 7 describe the experiment results and observations. Section 8 gives a summary and concluding remarks of the study. Last but not the least, section 7 cites some of our future work.

2. Research Objective

The objective of this research is to study and identify the most viable supervised machine learning algorithm (Random Forests (RF), Support Vector Machines (SVM)) which will suit to our proposed semantic-based mobile health monitoring system. We use datasets and carefully evaluate the experiments' results to quantify the data classification accuracy. These preliminary experiments are done to gain better understanding of machine learning algorithms and to gain insights to implement edge analytics in an effort to realize a personalized mobile health monitoring system.

3. Related research

In Mahmut Taha Yazici, et.al, [1], their research used Random Forest, Support Vector Machine (SVM) and Multi-Layer Perceptron for testing data sets on Raspberry pi. They have compared accuracy, processing capacity, power consumption between three algorithms in Raspberry pi. They have concluded that when the size of dataset is small, SVM is slightly faster than Random Forest; but when the size of data is large, Random Forest was faster with higher accuracy. While we share similar objectives, our tests were conducted on a PC which provides more processing power and generate our envisioned model.

Whereas, François-Élie Calvier, et.al. [2], proposed a method to bridge existing knowledge models with ad hoc taxonomies to address the problem of textual documents classification. This method allows the expert user to match their needs by optimising text document classification. This technique is used on web based textual documents. In contrast, in our study, we have implemented the basic concept of semantization to

improve the data pre-processing and in identifying the critical cases.

4. Experiment Environment and Dataset

In our experiments, we have used a Mac PC (macOS High Sierra version 10.13.6, 3.1 GHz Intel Core i5 CPU, and 8GB RAM). Python 3.7 programming language and Anaconda navigator were used as a programming platform with Spyder 3.3.1 IDE.

The datasets ([4], [5], [6]) consist of data generated from multi-modal sensors placed on the subject's chest, right wrist and left ankle. These three sensors record the body movement and turns of diverse human activity. The activity set consist of following labels:

1 = Standing Still (1min)
2 = Sitting and relaxing (1min)
3 = Lying down (1 min)
4 = Walking (1 min)
5 = Climbing stairs (1 min)
6 = Waist bends forward (20x)
7 = Frontal elevation of arms (20x)
8 = Knees bending (crouching) (20x)
9 = Cycling (1 min)
10 = Jogging (1 min)
11 = Running (1 min)
12 = Jump front & back (20x)
NOTE: In brackets are the number of repetitions (Nx) or the duration of the exercises (min).

Attribute information:
Attribute information consist of 23 different kind of information which was recorded from 3 sensors.
The meaning of each column is detailed next:
Column 1: acceleration from the chest sensor (X axis)
Column 2: acceleration from the chest sensor (Y axis)
Column 3: acceleration from the chest sensor (Z axis)
Column 4: electrocardiogram signal (lead 1)
Column 5: electrocardiogram signal (lead 2)
Column 6: acceleration from the left-ankle sensor (X axis)
Column 7: acceleration from the left-ankle sensor (Y axis)
Column 8: acceleration from the left-ankle sensor (Z axis)
Column 9: gyro from the left-ankle sensor (X axis)
Column 10: gyro from the left-ankle sensor (Y axis)
Column 11: gyro from the left-ankle sensor (Z axis)
Column 13: magnetometer from the left-ankle sensor (X axis)
Column 13: magnetometer from the left-ankle sensor (Y axis)
Column 14: magnetometer from the left-ankle sensor (Z axis)
Column 15: acceleration from the right-lower-arm sensor (X axis)
Column 16: acceleration from the right-lower-arm sensor (Y axis)
Column 17: acceleration from the right-lower-arm sensor (Z axis)
Column 18: gyro from the right-lower-arm sensor (X axis)
Column 19: gyro from the right-lower-arm sensor (Y axis)
Column 20: gyro from the right-lower-arm sensor (Z axis)
Column 21: magnetometer from the right-lower-arm sensor (X axis)
Column 22: magnetometer from the right-lower-arm sensor (Y axis)
Column 23: magnetometer from the right-lower-arm sensor (Z axis)
Column 24: Label (0 for the null class)

*Units: Acceleration (m/s^2), gyroscope (deg/s), magnetic field (local), ecg (mV)

5. System Architecture

Our proposed personalized health monitoring system consist of multi-modal sensors to record the human body activity and vital signs of a body, i.e. Electrocardiogram value, motions, etc., as shown in Fig. 1. These data will be pre-processed. To outline the scope of semantization, we conducted a small test experiment to find out the critical values. By extracting such information, we can be more accurate to predict as well as ~~can~~ to react sooner on critical situations. After pre-processing, the data will be passed through the trained models to predict the activity. Furthermore, after the prediction, semantization can be applied to augment the prediction accuracy of the subject's activity.

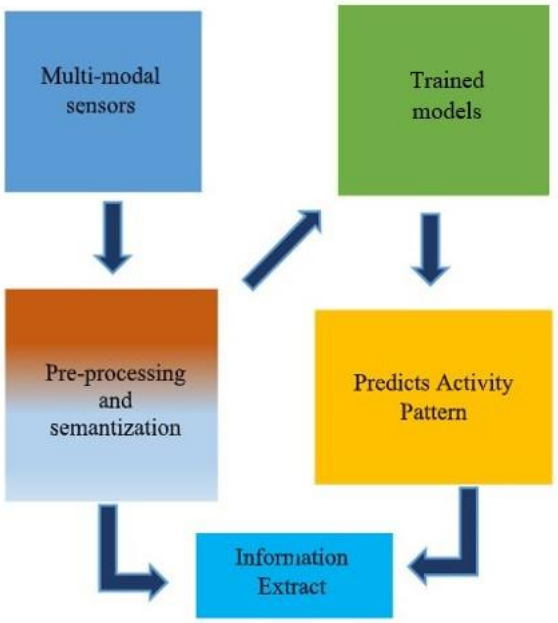

Fig. 1. Block diagram of a personalized health monitoring system.

6. Experiment Methods

The following are the procedures and steps in our experiments.

6.1 Data pre-processing

To make the input data easier to work with machine learning and semantization, data pre-processing was done.

6.1.1 Importing Libraries and datasets

First of all, libraries and datasets were imported, so that it will be easier to manipulate the data. Adding labels and additional information as well as finding out the boundary of the datasets were easily accomplished.

6.1.2 Missing data

To get the quality output, we checked for missing data. There are three famous ways of adjusting the missing data: mean calculation, median calculation and the most repetitive data of the missing data type. We used mean value calculation so that

there won't be much of the differences like when we use median value or highest repetitive value.

6.1.3 Splitting datasets

To implement machine learning algorithm, it is important to split the datasets into two, i.e., one for training purposes and the other for testing. Generally, datasets are split having training data's quantity higher than that of testing. So that the algorithm will be well trained. To this, we have implemented a 70:30, i.e., 70% of datasets is used for training purposes and the rest of the data for testing purposes.

6.1.4 Scaling data

Our dataset consists of different kinds of data with huge differences. So to optimize, we need to do the scaling of the data. There are four ways to scale the data: Rescaling (min-max normalization), Mean normalization, Standardization and Scaling to unit length. We have used Standardization to scale the data. The process of converting different variety of data into the same type considering the same basic features is standardization. In context to our datasets we have different types of datasets with large ranges like ECG and motion acceleration.

6.2 Training and Testing

After pre-processing of data, we train and test the algorithm one at time. We have used Support Vector Machines and Random Forest supervised machine learning algorithms in these experiments.

6.3 Data Analysis

During the analysis process, we concentrate on the accuracy check with respect to processing speed. Accuracy check between real output of the testing data and output of predicted data. Furthermore, to understand the concentration of the predicted data and real data we have used confusion matrix.

7. Experiments Results and Observations

These experiments were done to understand the implementation of machine learning on edge devices while considering the classification accuracy as a major measuring component. We calculate and compare the outcomes of two of algorithms, to find out the best algorithm for similar kind of datasets. The following measures were taken to analyse the outcome.

7.1 Data Classification Accuracy

The observation results were different with various parameters of machine learning algorithms. The maximum accuracy that was achieved with Random Forest was approximately 95% whereas with SVM, it was approximately 93%. The result was highly influenced by the size of dataset. While considering larger dataset, the processing speed of Random Forest was faster than SVM.

7.2 Confusion Matrix

Confusion Matrix is a matrix which compares the real output and output obtained using an algorithm on a test dataset. In other words, a confusion matrix is a technique for summarizing the performance of a classification algorithm. Although we were able to achieve 95% accuracy with the accuracy check but classification accuracy alone can be a misleading metrics for accuracy. Classification accuracy only provides a numerical value of whole datasets but to carefully analyze and understand every activity and its accuracy, confusion matrix is necessary.

We can see on the following figures that the concentration of the data is widely confused, i.e., mismatched in 0 activity rather than other activity. And the data size of the 0 activity is very large than other classes. So, we can know that the accuracy can be further improved. From this information the data received before the actual recording of the data is misleading the outcome. If we ignore the initial data i.e. activity with label 0, then we can have more accurate result.

The label and datasets are then organized into a tabular form, or as a matrix, shown as below in Fig. 2.1 and Fig. 2.2

Each row of the table corresponds to a predicted activity whereas each column of the table corresponds to an actual activity. The diagonally concentrated value are correctly predicted values whereas the majority of mismatched data are concentrated on 0 activity we can see on rows and columns of 0. But if we compare the data between random forest and SVM in activity 2, 3 and 4, random forest is able to classify the data properly than support vector machine.

	0	1	2	3	4	5	6	7	8	9	10	11	12
0	23771	212	295	182	104	23	79	84	127	203	88	33	19
1	13	615	0	0	0	0	0	0	0	0	0	0	0
2	0	0	572	0	0	0	0	0	0	0	0	0	0
3	0	0	0	614	0	0	0	0	0	0	0	0	0
4	5	0	0	0	591	0	0	0	0	0	0	0	0
5	130	0	0	0	0	501	0	0	2	0	0	0	0
6	14	0	0	0	0	0	606	0	0	0	0	0	0
7	127	0	0	0	0	0	0	475	2	0	0	0	0
8	100	0	0	0	0	0	0	0	565	0	0	0	0
9	43	0	0	0	0	0	0	0	0	587	0	0	0
10	12	0	0	0	0	0	0	0	0	0	590	3	0
11	23	0	0	0	0	0	0	0	0	0	1	627	0
12	68	0	0	0	0	0	0	0	0	0	0	4	146

Fig. 2.1. Confusion matrix after using Random Forest Algorithm

	0	1	2	3	4	5	6	7	8	9	10	11	12
0	24558	53	140	99	58	3	29	36	39	81	68	40	16
1	102	526	0	0	0	0	0	0	0	0	0	0	0
2	125	0	447	0	0	0	0	0	0	0	0	0	0
3	105	0	0	509	0	0	0	0	0	0	0	0	0
4	41	0	0	0	555	0	0	0	0	0	0	0	0
5	54	0	0	0	0	579	0	0	0	0	0	0	0
6	17	0	0	0	0	0	603	0	0	0	0	0	0
7	16	0	0	0	0	0	0	587	1	0	0	0	0
8	39	0	0	0	0	0	0	0	626	0	0	0	0
9	54	0	0	0	0	0	0	0	0	576	0	0	0
10	52	0	0	0	0	0	0	0	0	0	551	2	0
11	40	0	0	0	0	0	0	0	0	0	6	605	0
12	88	0	0	0	0	0	0	0	0	0	0	6	124

Fig. 2.2. Confusion matrix after using SVM algorithm

7.3 Visual Analysis

To make a visual analysis, we have plotted the data of one particular data attributes of predicted and actual data value i.e. X-axis Acceleration from chest sensor. First, we have plotted the test result and then predicted result and later on combined them both to see visual differences. There is not much of the difference between the following graphs but even in the picture (Fig. 2.3) we can see a lot of data on 0 activity i.e. data value recorded without any activity which are very highly miss matching to 12th activity in both SVM and Random Forest algorithm.

Fig 2.1 and Fig 3.1 represents the testing data's activity set of random forest and SVM. In both algorithms the same data set was used so, there is not much of difference. But in Fig. 2.2 and Fig. 3.2, we see the changes occurred on testing data result while using support vector machine and random forest. Then we have combined both the visual representation of the activity sets than we can see the differences between the real activity and predicted activity.

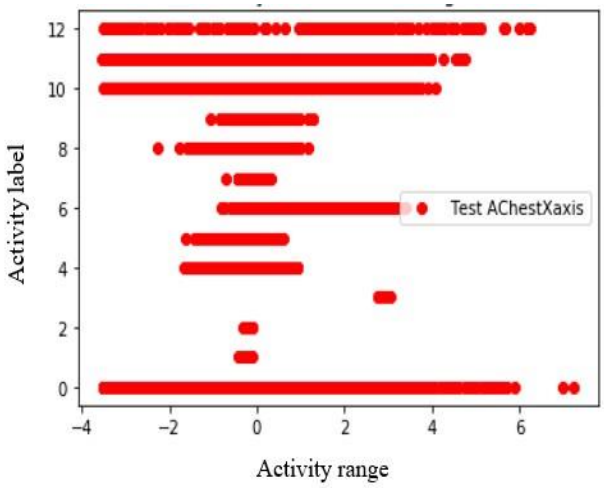

Fig. 2.1. Visual representation of testing data: X-axis acceleration recorded from chest sensors data using RF algorithm

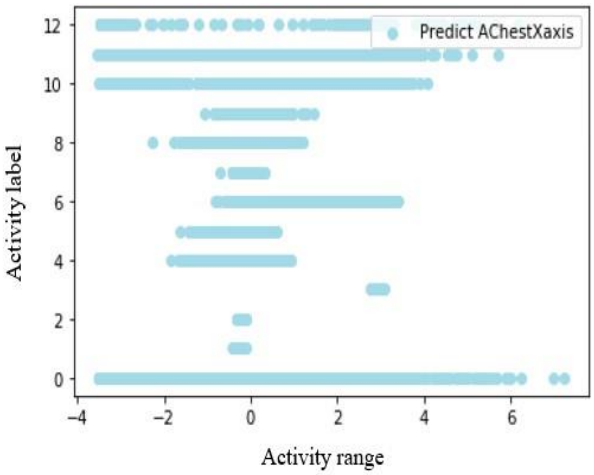

Fig. 2.2. Visual representation of predicted data: X-axis acceleration recorded from chest sensors data using RF algorithm

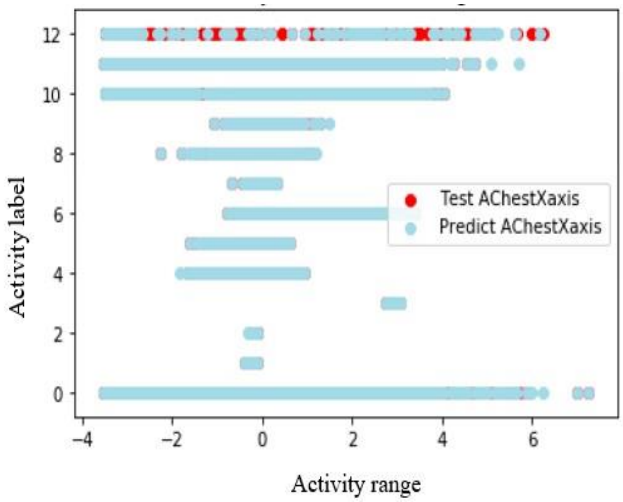

Fig. 2.3. Visual representation of testing and predicted data combined: X-axis acceleration recorded from chest sensors data using RF algorithm

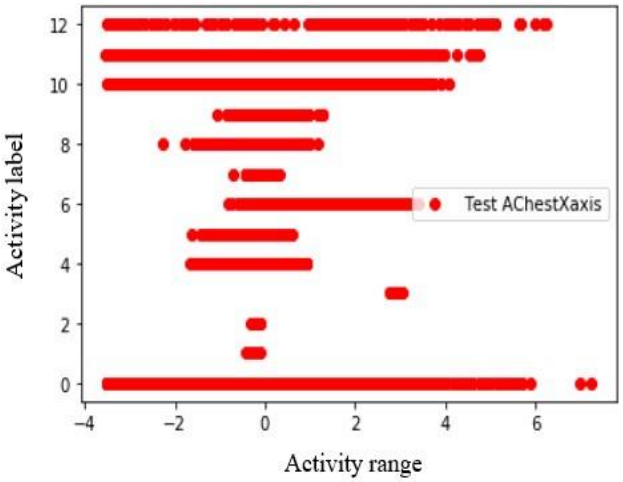

Fig3.1. Visual representation of testing data: X-axis acceleration recorded from chest sensors data using SVM algorithm

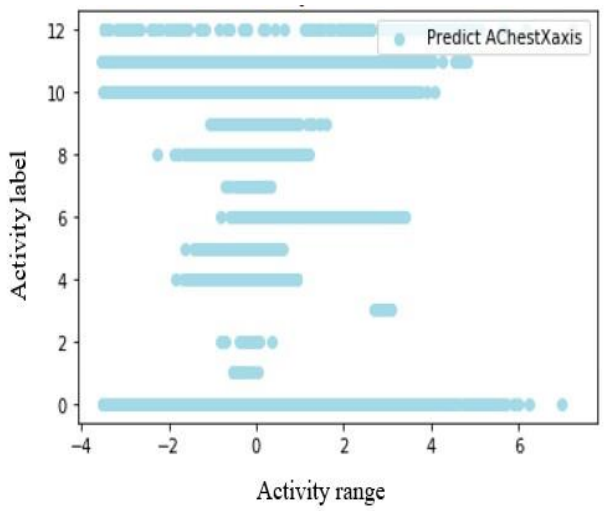

Fig. 3.2. Visual representation of predicted data: X-axis acceleration recorded from chest sensors data using SVM algorithm

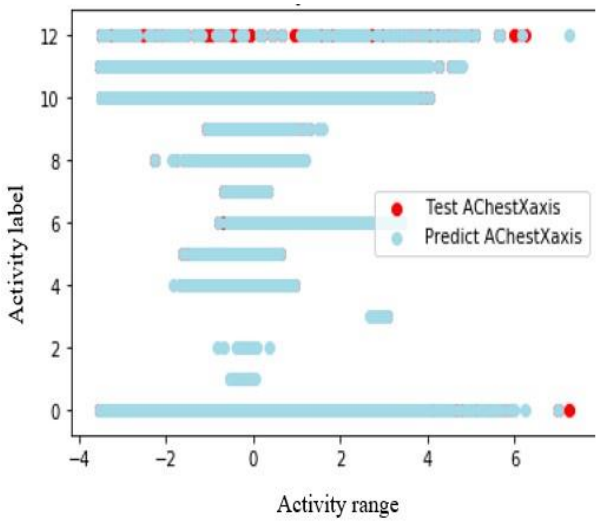

Fig. 3.3. Visual representation of testing and predicted data combined: X-axis acceleration recorded from chest sensors data using SVM algorithm

7.4 ECG values analysis with respect to different activity set

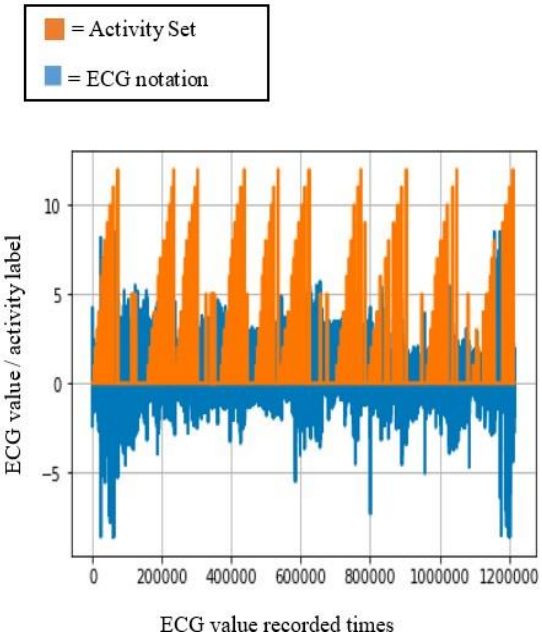

Fig. 4.1. Showing the ECG value with respect to activity sets before sorting.

The Y-axis represents the scaled ECG value, whereas X-axis represents the number of recording. Blue line represents the ECG notation whereas orange line represents the activity label. To understand the characteristic of the datasets and data concentrations, we have sorted the data values with respect to activity sets. Which can be observed in following diagram.

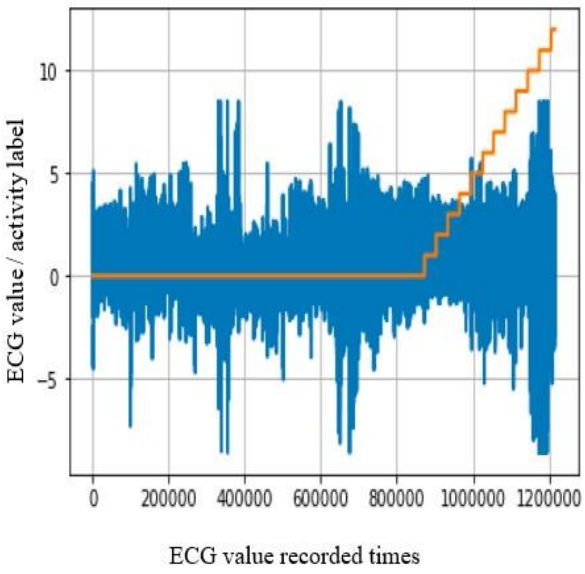

Fig. 4.2. Showing the ECG value with respect to activity sets after sorting.

In the above graph, we have compared the ECG values with respect to activity sets. We can observe the datasets consist of large number of ECG value with no activity. Although there is no activity but the ECG value is on fluctuation. Therefore, we

can say even while classifying and recording the data that we can still make a huge improvement on increasing the prediction accuracy and consistency of data.

From this ECG value we generate a boundary for lower limit and upper limit so that we can implement the semantization to identify the critical situation. With semantization, a lot of information can be extracted from these data, which will further help us to be more specific and accurate on our analysis. For example, if the ECG value is below or above the limit than system can notify the nurse or doctor.

8. Summary and Concluding Remarks

In this paper, we conducted experiments with mobile health data to determine which particular machine learning algorithm is best suited to our proposed personalized mobile health monitoring system. Experiment results depicts the characteristics of Support Vector Machines (SVM) and Random Forests (RF) machine learning algorithms with respect to our test dataset. The RF achieved a 95% data classification accuracy compared to 93% using SVM. Moreover, the RF exhibited faster processing speed with larger datasets. Confusion matrix was used to determine the behavior of the dataset with respect to the algorithm used. The aforementioned results provided clearer understanding of the algorithms and insights on realizing an intelligent edge analytics for mobile health monitoring.

9. Future Work

This study is still in its preliminary stage and much work has still to be done. We plan to study deeper on the algorithms while developing the framework on the data semantization for intelligent edge analytics.

References

[1] Natalia Diaz Rodriguez, M. P. Cuellar, Johan Lilius, and Miguel Delgado Calvo-Flores. 2014. A survey on ontologies for human behaviour recognition. ACM Computing Survey, 46, 4, Article 43, 33 pages, March 2014. DOI: http://dx.doi.org/10.1145/2523819

[2] Francois-Elie Calvier, Michel Plantie, Gerard Dray, Sylvie Ranwez. Ontology Based Machine Learning for Semantic Multiclass Classification. TOTH: Terminologie Ontologie: Theories et Applications 2013, Chambery, France, pp.100, June 2013.

[3] Garvita Bajaj, Rachit Agarwal, Pushpendra Singh, Nikolas Georgantas, Valerie Issarny. A study of existing Ontologies in the IoT-domain. Submitted to Elsevier JWS SI on Web Semantics for the Internet/Web of Things, 2017.

[4] Banos, O., Villalonga, C., Garcia, R., Saez, A., Damas, M., Holgado, J. A., Lee, S., Pomares, H., Rojas, I. Design, implementation and validation of a novel open framework for agile development of mobile health applications**.** BioMedical Engineering OnLine, vol. 14, no. S2:S6, pp. 1-20, 2015.

[5] Banos, O., Garcia, R., Holgado, J. A., Damas, M., Pomares, H., Rojas, I., Saez, A., Villalonga, C. mHealthDroid: a novel framework for agile development of mobile health applications. 6th International Work-conference on Ambient Assisted Living an Active Ageing (IWAAL 2014), Belfast, Northern Ireland, December 2-5, 2014.

[6] Dua, D. and Karra Taniskidou, E. (2017). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.