International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 8, No.6 (Nov-2019)

http://dx.doi.org/10.12785/ijcds/080612

# Evaluation of Semi-Supervised Clustering and Feature Selection for Human Activity Recognition

**Shadi Abudalfa[1] and Hani Qusa[2]**

[1]*University College of Applied Sciences, Palestine*
[2]*Higher Colleges of Technology, United Arab Emirates*

*Received 1 Jun. 2019, Revised 20 Aug. 2019, Accepted 20 Oct. 2019, Published 1 Nov. 2019*

**Abstract:** A lot of concern is shifted nowadays toward human activity recognition for developing powerful systems that assist numerous humans such as patients and elder people. Such smart systems automatically recognize human activities by learning the Activities of Daily Living (ADLs) and then making a suitable decision. Activity recognition systems are currently employed in developing many smart technologies (e.g., smart homes) and their uses have been dramatically increased with availability of Internet of Things (IoT) technology. Researchers have used various machine learning techniques for developing activity recognition systems. Nevertheless, there are some techniques have not been sufficiently exploited in this research direction. In this work, we present a framework to evaluate performance of one of these techniques. The presented technique is based on employing semi-supervised clustering and feature selection for grouping data collected by sensors located in smart homes. Employing semi-supervised clustering technique decreases the need for preparing a huge amount of labelled data that is required for learning activity recognition systems. Additionally, the presented technique improves performance of data clustering by decreasing a risk of grouping data into clusters that do not correspond to targeted activities. Moreover, we take into consideration importance of decreasing computational time complexity for making the presented technique applicable to smart systems located in homes. We conducted various experiments to evaluate performance of employing semi-supervised clustering and feature selection for human activity recognition by using partially labelled data. Experiment results have shown that the presented technique provides remarkable accuracy.

**Keywords:** Evaluation, Semi-Supervised, Data Clustering, Feature Selection, Activity Recognition, Smart Homes

## 1. INTRODUCTION

The wealth of sophisticated sensors motivated researchers to develop techniques that recognize human activities for assisting numerous humans [1]. Human activity recognition is an important task in implementing numerous smart technologies such as smart homes [2]. Such task received a high interest these days with availability of Internet of Things (IoT) technology [3][4].

Activity recognition is one of tasks that are included under umbrella of pattern recognition. Thereby, many machine learning techniques are employed for human activity recognition. Machine learning techniques are categorized mainly into supervised, unsupervised, and semi-supervised learning.

Supervised learning uses only labeled data to train models employed for activity recognition [5]. Building accurate recognizer needs a large amount of labeled data. However, labeling adequate data for human activity recognition is time consuming and may lead to many errors.

To get rid of annotating data when training models, another category of machine learning techniques called unsupervised learning are employed for activity recognition [6]. Such techniques use only unlabeled data and most of them are developed to perform data clustering that divides data into groups (clusters) in which each group has similar properties.

Accuracy of data clustering methods is still limited since we do not have enough information when applying them to unlabeled data. Additionally, most of clustering algorithms are sensitive to parameter initialization. For example, K-means is sensitive to initializing centroids that represent the clusters. Thus, various solutions have been proposed in the literature to improve accuracy of data clustering. One of research directions is based on mimicking semi-supervised learning [7] which uses both labeled and unlabeled data. This technique is referred to as semi-supervised clustering that employs semi-supervised learning for initializing the centroids from a little amount of labeled data points while the clustering process is applied as usual to unlabeled data [8][9].

*E-mail:shadi_abudalfa@hotmail.com, hqusa@hct.ac.ae*

Semi-supervised clustering [10] is based on feeding clustering algorithms with some background knowledge provided by a domain expert about the structure of the data. There are three dominant types of semi-supervised clustering [11] covered in the literature. The first one uses partially labeled data for achieving data clustering. Thereby, the need to prepare a large amount of labeled data is decreased in comparison with supervised learning. The second type entitled cluster-level constraints since it provides information about the clusters. While, the third type is referred to as instance-level constraints [12], also called pairwise constraints since it provides a background knowledge about pair of points located in the dataset.

In this work, we evaluated performance of employing semi-supervised clustering for recognizing human activities. The main emphasis in our work is using partially labeled data when applying data clustering. Additionally, we examined performance of applying feature selection when preparing the dataset used for human activity recognition.

The rest of paper is organized as follows: Section 2 describes a review for some related studies. Section 3 describes the framework used to evaluate the presented semi-supervised clustering technique. Section 4 explains the experiment environment. Section 5 discusses experiment results and provides adequate analysis. Finally, Section 6 concludes the paper and reveals some suggestions for future work.

## 2. LITERATURE REVIEW

A lot of research works have been proposed in the literature to show performance of employing semi-supervised learning for action recognition [13]. Some of state of art techniques manipulated the problem as digital image processing by collecting data from video camera sensors. For example, Hejin Yuan [14] used skeleton detection with semi-supervised learning for improving accuracy of recognizing human activities. Shen et al. [15] evaluated different semi-supervised learning methods by using same experiment setting. There work is specifically cornered on evaluating semi-supervised learning techniques applied in computer vision and multimedia areas for action recognition.

Another research direction, close to our emphasis in this paper, is based on identifying human actions by collecting data from different sensors such as ultrasonic wave sensors and radio-frequency identification (RFID) [16][17]. For example, Cvetkovic et al. [18] proposed semi-supervised multi-classifier adaptive training algorithm for increasing accuracy of the initial activity recognition classifier. Additionally, Guan et al. [19] proposed a semi-supervised learning algorithm to use unlabeled activity samples by using co-training method.

Recently, some research works are conducted to evaluate specifically performance of semi-supervised clustering methods. Svehla [11] reviews and provides empirical evaluation on several semi-supervised clustering methods with emphasis on methods that utilize active learning of pairwise constraints. Svehla used various datasets selected from the UCI Machine Learning Repository [20] for conducting experimental work.

Based on our knowledge, limited works are done by employing semi-supervised clustering for human activity recognition with same research direction selected in our work. As a result of this, we have been motivated to fill some gaps in this direction.

For instance, Hejin Yuan and Cuiru Wang [21] proposed a method of human action recognition based on semi-supervised K-means clustering. Their proposed solution is presented in the direction of image processing and experimented by using Weizmann dataset [22].

In the other direction, Hashim Ali et al. [23] presented a framework for activity recognition based on semi-supervised clustering approach to avoid using traditional clustering methods. Their presented methodology is composed of three phases: cluster centroid initialization, physical activity classification, and reinforcement learning. Thereby, their presented methodology is a time consuming.

Similarly, our research work deals with initializing cluster centroids by using limited amount of labeled data. While, we also consider decreasing computational time complexity to make the presented technique applicable for smart homes that use low computational resources.

## 3. EVALUATION FRAMEWORK

The presented technique can be evaluated by passing through different stages as shown in Figure 1. It starts by collecting data from sensors located in smart home. After collecting data, the next step is extracting features (attributes) that are suitable to recognize human activities. The next stage is selecting the most dominant features that affect significantly on recognizing targeted activities. In this stage, we firstly remove redundant features that have same attribute values. Then, we apply feature selection to reduce number of features used for recognizing targeted activities.

After that, we scale all attribute values to the same range. After conducting this step, the dataset will be ready to use. This step is very important and significantly affects on the accuracy of activity recognition when the original dataset includes attribute values with different scales. Mainly, this phase removes the mean and scales features to unit variance. Accordingly, the required score for each sample (data point) $X$ is calculated by using equation 1.

$$Z = (X - M) / S \qquad (1)$$

Where $M$ and $S$ are the mean and standard deviation of the training samples respectively.

For evaluating performance of semi-supervised clustering, we need to divide the dataset into training and testing sets. Then, we select limited amount of labeled data from training set for initializing centroids used by clustering algorithms. In this work we specifically use K-means [24] and C-means [25] algorithms for evaluating performance of partially labeled semi-supervised clustering. These algorithms use partitional strategy to minimize the distance between data points in same cluster and represent each cluster by one data point (centroid). Our selection to these algorithms depended on reducing computational time complexity [26]. Thereby, the presented technique can be efficiently implemented with smart home systems that have limited resources.

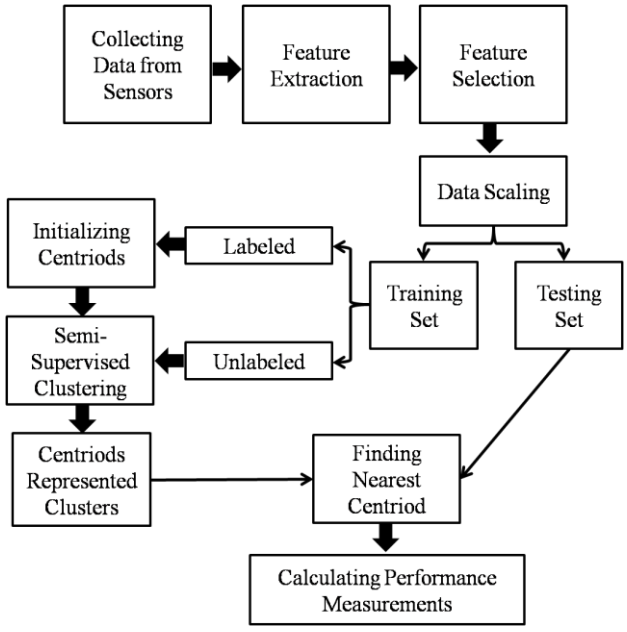



Figure 1. Framework presented for evaluating performance of applying semi-supervised clustering to activity recognition

The presented technique initializes centroids by using small amount of labeled data. The process of initializing centroids is based on randomly selecting each initial centroid from its corresponding class included in the labeled data. This step helps in avoiding fully random selection of initial centroids when applying traditional clustering algorithms. Therefore, this step decreases a risk of grouping data into clusters that do not correspond to the targeted activities.

Furthermore, it is interesting to clarify that the presented technique uses small amount of labeled data for applying centroid initialization. Thereby, the presented technique reduces the implementation cost by precluding the need for preparing a huge amount of labeled data that is used to provide acceptable accuracy. After conducting process of centroid initialization, the initial centroids are used for applying the clustering algorithm to the rest of training data (as unlabeled data).

Finally, the presented framework finds which data points in the testing set are closest to each output centroid and then calculates the performance. To find the closest centroid, we should select a suitable distance measure [27]. Thus, we need to evaluate different distance measures and then select the distance measure that provides the best accuracy.

## 4. Experiment Setup

We experimentally evaluated performance of employing clustering algorithms with the presented semi-supervised clustering technique along with applying feature selection. This section describes the experiment environment used for conducting our experiment work. Next, we describe the dataset, evaluation measures, and settings used for conducting our experiments. The section describes also libraries and tools used in this work.

### *A. Dataset*

Our experiment work is conducted by using a public dataset entitled ActivitiesExercisesRecognition [1] that includes 1150 samples (data points) collected by Kévin Chapron et al. affiliated to the LIARA laboratory at University of Québec in Chicoutimi. The dataset covers ten classes of activities collected from ten participants as shown in Table I. This table includes same description provided by Chapron et al.

Table I. Activities included in the used Dataset.

| # | Activity | Description |
|---|---|---|
| 1 | ExoFente | Typical Front Lunge |
| 2 | ExoMarche | Front Step Down |
| 3 | ExoSitUp | Sit-to-Stand |
| 4 | ExoSquat | Typical Squat on chair |
| 5 | shortRunning | Continuously running for 30 seconds |
| 6 | shortSeat | Staying sit in a chair for 30 seconds |
| 7 | shortSitting | Transition Stand-to-Sit |
| 8 | shortStanding | Transition Sit-to-Stand |
| 9 | shortSwitch | Fast rotation of the wrist |
| 10 | shortWalking | Continuously walking for 30 seconds |

Additionally, we use same features extracted by Chapron et al. The feature set includes 105 features reported by Chapron et al. as shown in Table II. Moreover, we used 70/30 split method for dividing the dataset into training and testing sets. Thereby, the training set (70%) includes 805 samples (data points). While, the testing set includes the rest 345 samples.

[1] https://github.com/LIARALab/Datasets/tree/master/ActivitiesExercisesRecognition

### B. Evaluation Measures

Empirical results obtained from experiments provide a good way to evaluate performance of applying semi-supervised clustering for human activity recognition. We use classification accuracy and F1-score [28] for evaluating the presented technique. The F1-score is basically used with binary classification. Thus, we use the macro-average F1-score [29] for evaluating multiclass classification (more than two classes) which is covered in this research work.

TABLE II. FEATURE SET USED DATASET.

| Domain | Feature | Count |
|---|---|---|
| Temporal | Mean value for each axis | 9 |
| | Mean value of mean values | 3 |
| | Standard Deviation for each axis | 9 |
| | Mean value of standard deviation | 3 |
| | Skewness Value for each axis | 9 |
| | Mean value of Skewness value | 3 |
| | Kurtosis Value for each axis | 9 |
| | Mean value of Kurtosis value | 3 |
| | Zero Crossing Rate for each axis | 9 |
| | Mean value of Zero Crossing Rate | 3 |
| | Correlation between every axis | 18 |
| Frequential | DC Component for each axis | 9 |
| | Energy for each axis | 9 |
| | Entropy for each axis | 9 |
| | **Total** | 105 |

It is noteworthy that we cannot use accuracy and F1-socre to evaluate performance of traditional (unsupervised) clustering algorithms since resulted clusters may not be corresponded to right activity classes included in the used dataset. Such case needs to use other evaluation measures to compare the clustering solution to the ground truth (known classes) when applying traditional clustering algorithms. For example, we can use the Adjusted Rand index [30]. Whereas, applying the presented technique can be evaluated by using accuracy and F1-socre measures since we use labeled data for initializing the centroids that usually correspond to the targeted classes.

### C. Experiment Implementation

All experiments were implemented by using the Python[2] programming language version 3.7.1 64-bit. For handling data and reporting experiment results, we used the following open-source libraries:

- NumPy [31] — a package for scientific computing.
- Pandas [32] — a package for loading and analyzing data.
- Pyclustering — a package for applying clustering algorithms.
- Scikit-learn [33] — a popular machine learning library. We used its implementation for applying feature selection and evaluation measures.
- SciPy [34]—we used its implementation for applying distance measures.
- Microsoft Excel — we used it for calculating confidence intervals and plotting graphs.

[2] https://www.python.org/

### D. Parameter Intilizaton

We have evaluated the presented semi-supervised clustering technique with two selected clustering algorithms called K-means and C-means. Before applying data clustering, we applied centroid initialization by using small amount of labeled data selected from training set.

We applied two methods for initializing centroids used with K-means and C-means. The first method is based on assigning each centroid to the mean value of the corresponding class existed in the labeled data. The second method mimics initializing each centroid by selecting randomly one data point existed in the corresponding class. Table III describes all methods used for conducting our experiment work.

TABLE III. METHODS USED FOR EVALUATING THE PRESENTED TECHNIQUE

| Method | Description |
|---|---|
| K-meansM | K-means algorithm with initializing centroids by using corresponding means of classes existed in the partially labeled data. |
| K-meansS | K-means algorithm with initializing centroids by selecting randomly one sample (data point) of the corresponding classes existed in the partially labeled data. |
| C-meansM | C-means algorithm with initializing centroids by using corresponding means of classes existed in the partially labeled data. |

## 5. EXPERIMENT RESULTS AND ANALYSIS

We conducted numerous experiments to evaluate performance of applying the presented technique to the used dataset. This section shows experiment results provided when applying the presented technique through various perspectives.

### A. Data Scaling

Applying data scaling is very important to improve significantly performance of applying presented semi-supervised clustering technique to the used dataset. To apply data scaling, we used a method entitled StandardScaler implemented in Scikit-learn [33] library.

For clarifying effect of applying data scaling, we applied *K-meansM* with selecting 15% of the training set as labeled data for initializing centroids. Then, we used the whole training set (as unlabeled data) for data clustering. The results of applying data scaling with this experiment are 77.4% and 73.1% for accuracy and macro-average F1-score respectively. While, the results degrade

sharply when neglecting data scaling to be 16.5% and 9.2% for accuracy and macro-average F1-score respectively.

### B. Time Complexity

In this work, we mainly focus on evaluating accuracy of activity recognition. Whereas, reducing computational time complexity is acquired in our work by using simple clustering algorithms. From a technical point of view, the computational time consumed when applying the presented technique can be neglected since the implementation cost of the used clustering algorithm (K-means) is very simple. Based on our experiment work, each run takes less than 0.5 second when applying the presented technique to the used dataset. Table IV shows computational time of applying K-means algorithm by using a machine that has 4.00 GB memory and 1.6 GHZ Intel / Celeron (R) processor. The table shows a summary of the provided result along with confidence interval equals to 95% when running the experiment 101 times.

TABLE IV. COMPUTATIONAL TIME COMPLEXITYWITH K-MEANS

| | Computational Time (Seconds) |
|---|---|
| **Maximum** | 0.26 |
| **Minimum** | 0.07 |
| **Avgerage** | 0.08 |
| **Confidence Level** | ± 0.00525 |

### C. Distance Measures

It is interesting to clarify that performance of the presented technique is sensitive to selecting different distance measures when calculating distances between each centroid and the nearest data point. To show this effect, we applied different distance measures with *K-meansM* method by selecting 15% of the training set as labeled data and then clustering the whole training set (as unlabeled data). We used SciPy library for applying different distance measures. The vacillation in performance is clearly shown in Table V.

TABLE V. EFFECT OF USING DIFFERENT DISTANCE MEASURES

| Measure | Accuracy | Macro F1-score |
|---|---|---|
| Cosine | 78.0 | 74.4 |
| Bray-Curtis | 77.4 | 73.1 |
| Canberra | 73.9 | 67.6 |
| Chebyshev | 74.2 | 70.5 |
| City Block (Manhattan) | **78.6** | **75.0** |
| Correlation | 76.5 | 72.7 |
| Euclidean | 78.0 | 74.7 |
| Minkowski | 78.0 | 74.7 |
| squared Euclidean | 78.0 | 74.7 |

Since using City Block (Manhattan) distance measure provides the best result, we used it for conducting all experiments in this work. However, using different distance measure may work better with other clustering algorithm. For example, our experiment work showed that using Correlation distance measure provides the best performance when applying *C-meansM* algorithm. Using *C-meansM* with Correlation distance measure provides 44.3% and 34.2% for accuracy and macro-average F1-score respectively. While, using *C-meansM* with City Block distance measure provides 36.8% and 24.1% for accuracy and macro-average F1-score respectively.

### D. Experiment Work with Semi-Supervised Clustering Methods using Partially Labeled Data

We evaluated effect of changing the ratio that is selected from the training set as labeled data for initializing centroids when applying *K-meansM*. In this experiment, we applied the clustering algorithm to the whole training set (as unlabeled data). We selected *K-meansM* to conduct this experiment because this method provides stationary results when repeating experiment execution by using same settings. We used the following ratios to select labeled data from training set when conducting this experiment:

{0.15, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1}

We started with 0.15 to select enough data points that cover all activity classes existed in the used dataset when representing the corresponding initial centroids. Figure 2 illustrates classification accuracies provided by this experiment. The x-axis shows the values of ratio after scaling them by multiplying each one by 10.

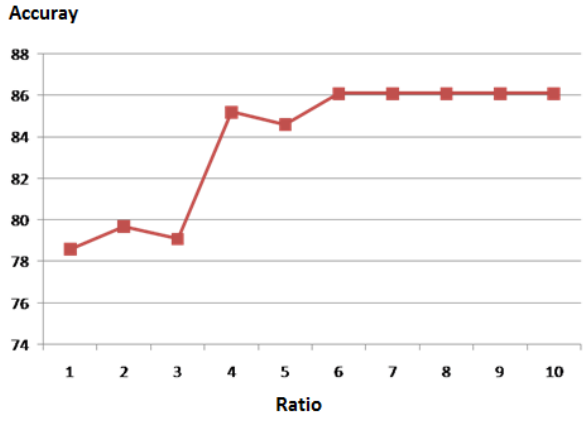



Figure 2. Accuracy of changing the ratio of labeled data selected for centroid initialization when applying K-means algorithm with specific settings

We can notice that there is no significant improve in accuracy when using ratio equals to more than 0.4. Similarly, we repeated this experiment by using *C-meansM* with Correlation distance measure. As shown in Figure 3, the results assent that using only 40% of training data (28% of the whole dataset) provides competitive

results. Therefore, we used this ratio as default value for conducting the rest of experiments.

It is noteworthy that applying the clustering algorithm to the rest part of training set after excluding data points used for centroid initialization performs different results. Based on our experiment work, this scenario provides better results. To show the improvement, we applied *K-meansM* and *C-meansM* to 60% of training data with using the rest 40% as labeled data for centroid initialization. Axiomatically, it is illogical to evaluate all ratios since increasing ratio of labeled data with this scenario hurts applying clustering algorithms by decreasing gradually the data points used for data clustering.

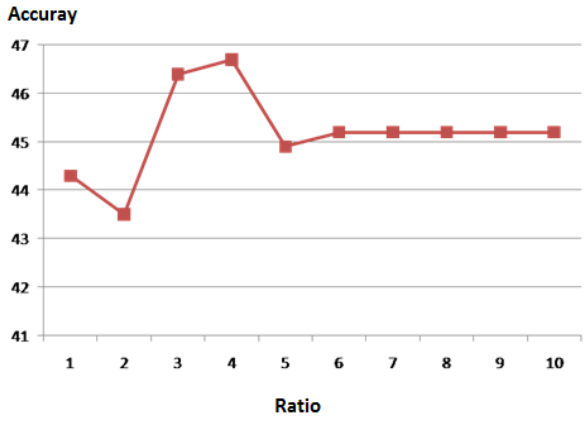



Figure 3. Accuracy of changing the ratio of labeled data selected for centroid initialization when applying C-means algorithm with specific settings

Based on our experiment work, applying *C-meansM* with Correlation distance measure provides 51.0% and 35.7% for accuracy and macro-average F1-score respectively. While, applying *K-meansM* with City Block distance measure provides 88.4% and 86.8% for accuracy and macro-average F1-score respectively.

Since using 60% of training data for data clustering along with using the rest 40% for centroid initialization provides better result, we used this scenario to conduct all experiments that evaluate the rest of strategies applied to show performance of the presented semi-supervised clustering technique.

We conducted an experiment to evaluate performance of the presented technique when applying *K-meansS* method. To find confidence interval with confidence equals 95%, we run this experiment 101 times since the results are non-stationary. This experiment is conducted by using 40% of training set as labeled data for applying centroid initialization while the rest 60% of training set is used as unlabeled data for applying the clustering method. Similarly, same testing set is used for reporting the performance. We used City Block distance measure for finding output labels as described in Section 3. Table VI shows a summary of the provided result along with confidence interval equals to 95%.

TABLE VI. SEMI-SUPERVISED CLUSTERING METHODS USING PARTIALLY LABELED DATA

| | Accuracy (%) | Macro F1-score (%) |
|---|---|---|
| **Maximum** | 89.0 | 87.8 |
| **Minimum** | 8.7 | 6.7 |
| **Avgerage** | 72.3 | 67.3 |
| **Confidence Level** | ± 3.2 | ± 3.3 |

As shown in the table, this method provides competitive results since the maximum reported accuracy equals to 89.0%. Based on the result provided by this experiment, the minimum accuracy (8.7%) is reported only one time. This worse result is expected since we use K-means algorithm which is sensitive to the process of centroid initialization. Thereby, the accuracy will be worse when the presented method selects initial centroids form inadequate samples located in the used dataset. The experiment results show that this case is not dominant since accuracies with values equals to or greater than 80% have been reported 39 times. Therefore, we can note that using the presented technique for activity recognition is a promising research direction.

### *E. Feature Selection*

We evaluated performance of selecting best features that robustly discerns all data points existed in the dataset. For applying feature selection, we firstly removed features that have redundant attribute values. Then, we applied feature selecting by using method entitled SelectKBest implemented in Scikit-learn library. After removing features that have redundant attribute values, number of remaining features becomes 35 features. Figure 4 shows classification accuracies provided when applying *K-meansM* method with changing number of selected features (non-redundant features).

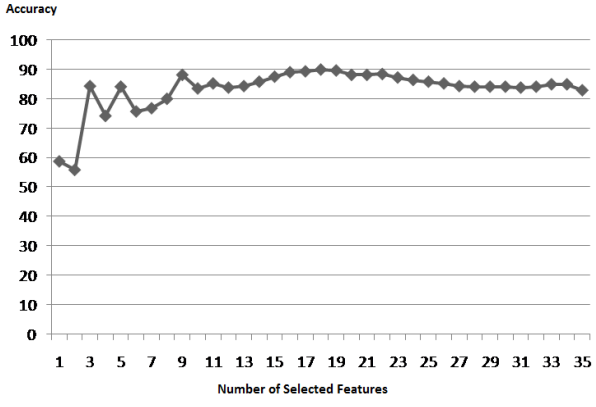



Figure 4. Accuracy of applying feature selection when using *K-meansM*

Based on results shown in Figure 4, the best accuracy is achieved (89.9%) when selecting 18 features. Thus, we selected these 18 features to evaluate performance of

applying *K-meansS* method. Table VII shows a summary of the results provided along with confidence interval equals to 95%. Similarly, this experiment is run 101 times to make a comparison with results reported in Table VI.

TABLE VII. SEMI-SUPERVISED CLUSTERING METHODS USINGPARTIALLY LABELED DATA WITH FEATURE SELECTION

| | Accuracy (%) | Macro F1-score (%) |
|---|---|---|
| **Maximum** | 88.1 | 86.6 |
| **Minimum** | 11.9 | 10.4 |
| **Avgerage** | 81.0 | 78.2 |
| **Confidence Level** | ±2.7 | ±2.7 |

We can notice clearly that using feature selection improves the performance since the average values are increased and the confidence level is decreased. Moreover, accuracies with values equal to or greater than 80% have been increased into 84 times.

### *F. Summary*

Based on our experiment work, we can conclude that the presented semi-supervised clustering technique provides remarkable performance. We can note as well that using data scaling improves the performance significantly. Additionally, experiment results show that applying feature selection increases the performance. Moreover, when using this technique we should select a suitable distance measure for achieving high performance.

## 6. CONCLUSION AND FUTURE WORK

Performance of presented semi-supervised clustering technique for human activity recognition has been evaluated in this paper by using specific strategies. This work helps scholars to assess their solutions proposed in this research direction. The experiment results show that the presented technique performs remarkable accuracy. We concluded from our experiment work that applying data scaling and feature selection along with selecting suitable distance measure provides competitive results. Thereby, using the presented technique for human activity recognition is a promising research direction.

This research can be extended in different ways. It is interesting to show performance of applying the presented technique to more datasets. Additionally, evaluating the presented technique by using more clustering algorithms will be a promise research direction. The future work may evaluate more partial clustering algorithms such as K-medoids [35], K-medians [36] or evaluate other types of clustering algorithms such as hierarchical clustering [37]. Moreover, it is important to evaluate performance of employing more types of semi-supervised clustering such as active semi-supervised clustering for human activity recognition. The performance may be also improved significantly when applying more sophisticated feature selection methods.

## REFERENCES


[1] K. Rasch, "Smart Assistants for Smart Homes," Doctoral Thesis, Royal Institute of Technology, 2013.

[2] M. Chan, E. Campo, D. Estève, and J.-Y. Fourniols, "Smart homes - current features and future perspectives," *Maturitas*, vol. 64, no. 2, pp. 90-97, 2009.

[3] I. Iee, and K. Lee, "The Internet of Things (IoT): applications, investments, and challenges for enterprises," *Business Horizons*, vol. 58, pp. 431–440, 2015.

[4] A. Ramadas, "Smart-Homes Activity Pattern Recognition: A Comparative Study," Master Thesis, University of PORTO, 2018.

[5] J. Yang, M. Nguyen, P. San, X. Li, and S. Krishnaswamy, "Deep Convolutional Neural Networks On Multichannel Time Series For Human Activity Recognition," *In Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)*, Buenos Aires, Argentina, 25–31 July 2015, pp. 3995–4001.

[6] J. Lapalu, K. Bouchard, A. Bouzouane, B. Bouchard, and S. Giroux, "Unsupervised Mining of Activities for Smart Home Prediction," *Procedia Computer Science*, 19, 503-510, 2013.

[7] N. Grira, M. Crucianu, and N. Boujemaa. "Unsupervised and Semi-Supervised Clustering: A Brief Survey," *Report of the MUSCLE European Network of Excellence (FP6)*, 2004.

[8] S. Basu, A. Banerjee, and R. Mooney, "Semi-supervised Clustering by Seeding," *In Proceedings of the 19th International Conference on Machine Learning (ICML-2002)*, Australia, July 2002, pp. 19-26.

[9] X. Wang, C. Wang, and J. Shen, "Semi–supervised K-Means Clustering by Optimizing Initial Cluster Centers," *Web Information Systems and Mining*, vol. 6988 of the series Lecture Notes in Computer Science, pp. 178-187, 2011.

[10] E. Bair, "Semi-supervised clustering methods," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 5, no. 5, pp. 349–361, 2013.

[11] J. Svehla, Active Semi-Supervised Clustering, *Master's thesis*, 2018.

[12] K. Wagstaff, C. Cardie, S. Rogers, and S. Schroedl, "Constrained K-means Clustering with Background Knowledge," *In ICML*, vol. 1, 2001, pp. 577–584.

[13] M. Stikic, K. Laerhoven, and B. Schiele, " Exploring Semi-Supervised and Active Learning for Activity Recognition," *In Proceedings of the 12th IEEE International Symposium on Wearable Computers (ISWC)*, 2008.

[14] H. Yuan, "A Semi-supervised Human Action Recognition Algorithm Based on Skeleton Feature," *Journal of Information Hiding and Multimedia Signal Processing*, vol. 6, no. 1, 2015.

[15] H. Shen, Y. Yan, S. Xu, N. Ballas, and W. Chen, "Evaluation of Semi-Supervised Learning Method on Action Recognition," *Multimedia Tools and Applications*, vol. 74, no. 2, pp. 523–542, 2015.

[16] K. Bouchard, D. Fortin-Simard, S. Gaboury, B. Bouchard, and A. Bouzouane, "Accurate trilateration for passive RFID localization in smart homes," *International Journal of Wireless Information Networks*, vol. 21, pp. 32-47, 2014.

[17] K. Bouchard, F. Bergeron, and S. Giroux, "Applying Data Mining in Smart Home," *Smart Technologies in Healthcare*, pp. 146-177, 2017.

[18] B. Cvetkovic, B. Kaluza, M. Lustrek, and M. Gams, "Multi-Classifier Adaptive Training: Specialising an Activity Recognition Classifier Using Semi-supervised Learning," *Ambient Intelligence LNCS*, vol. 7683, Springer, Heidelberg, 2012, pp. 193–207.

[19] D. Guan, W. Yuan, Y. Lee, A. Gavrilov, and S. Lee, "Activity Recognition Based on Semi-supervised Learning," *In 13th IEEE International Conference on Embedded and Real-Time Computing Systems and Applications (RTCSA)*, Aug. 2007, pp. 469–475.

[20] D. Dua,C. Graff, UCI Machine Learning Repository, 2019, http://archive.ics.uci.edu/ml, Irvine, CA: University of California, School of Information and Computer Science.

[21] H. Yuan, and C. Wang, "A human action recognition algorithm based on semi- supervised kmeans clustering," *Transactions on Edutainment, Lecture Notes in Computer Science*, vol. 6758, 2011, pp. 227–236.

[22] B. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as Space-Time Shapes," *In International Conference on Computer Vision*, Beijing, China, Oct 15–21, 2005, pp. 1395–1402.

[23] H. Ali, E. Messina, and R. Bisiani, "Subject-Dependent Physical Activity Recognition Model Framework with a Semi-Supervised Clustering Approach," *In Proceedings of the 2013 European Modelling Symposium (EMS)*, UK, 20–22 November 2013; pp. 42–47.

[24] C. Ding, and X. He, "K-means Clustering via Principal Component Analysis," *In Proceedings of the 21st International Conference on Machine Learning*, Banff, Canada, July 2004, pp. 225–232.

[25] F. De Carvalho, E. Simoes, L. Santana, and M. Ferreira, "Gaussian Kernel C-means Hard Clustering Algorithms with Automated Computation of the Width Hyper-Parameters," *Pattern Recognition*, vol 79, pp. 370–386, 2018.

[26] T. Velmurugan, and T. Santhanam, "Computational Complexity between K-means and K-medoids Clustering Algorithms for Normal and Uniform Distributions of Data Points," *Journal of Computer Science*, vol. 6, no. 3, pp. 363–368, 2010.

[27] S. Abudalfa, and M. Mikki, "K-Means Algorithm with a Novel Distance Measure," *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 21, no. 6, pp. 1665-1684, Oct 2013.

[28] C. Metz, "Basic principles of ROC analysis," *Semin Nucl Med.*, vol. 8, no. 4, pp. 283–98, Oct 1978.

[29] S. Parambath, N. Usunier, and Y. Grandvalet, "Optimizing F-measures by Cost- Sensitive Classification," *in Neural Information Processing Systems (NIPS)*, 2014, pp. 2123-2131.

[30] L. Hubert, and P. Arabie, "Comparing partitions," *Journal of classification*, vol. 2, no. 1, pp. 193–218, 1985.

[31] T. Oliphant, "A guide to NumPy," vol. 1. *Trelgol Publishing USA*, 2006.

[32] W. McKinney, "Data Structures for Statistical Computing in Python," *In Proceedings of the 9th Python in Science Conference*, vol. 445, Austin, TX, 2010, pp. 51–56.

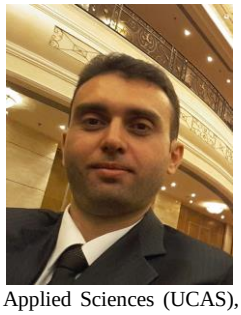
[33] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[34] E. Jones, *SciPy: Open Source Scientific Tools for Python*, 2001, www.scipy.org.

[35] D. Yu, G. Liu, M. Guo, and X. Liu, "An improved K-medoids algorithm based on step increasing and optimizing medoids," *Expert Syst. Appl.*, vol. 92, pp. 464-473, 2018.

[36] J. Li, S. Song, Y. Zhang, and Z. Zhou, "Robust K-median and K-means Clustering Algorithms for Incomplete Data," *Mathematical Problems in Engineering*, pp. 1–8, 2016.

[37] S. Abudalfa, and M. Mikki, "A Dynamic Linkage Clustering using KD-Tree," *International Arab Journal of Information Technology (IAJIT)*, vol. 10, no. 3, May 2013.



**Shadi Abudalfa** received the BSc and MSc degrees both in Computer Engineering from the Islamic University of Gaza (IUG), Palestine in 2003 and 2010 respectively. He completed his PhD program in Computer Science and Engineering at King Fahd University of Petroleum & Minerals (KFUPM), Kingdom of Saudi Arabia in 2018.Abudalfa has a strong teaching and research experience. He is an assistant professor at the University Collage of Applied Sciences (UCAS), Palestine.From July 2003 to August 2004, he worked as a research assistant at Projects and Research Lab in IUG. During same period, he also worked asa teaching assistant at Faculty of Engineering in IUG.Abudalfa isa member of IEEE and his current research interests includeartificial intelligence, data mining, pattern recognition, machine learning,and sentiment analysis.

**Hani Qusa**is a dedicated Information Security Professional (CISSP) with a strong academic background, teaching and research experience. He received his Ph.D. in Computer Engineering in 2013 from University of Rome "La Sapienza" (Italy). Currently, he is an Assistant Professor in the Department of Computer and Information Systems in Higher Colleges of Technology. Equipped with analytical and conceptual thinking skills to problem-solving using IT solutions, he participated in several EU funded Erasmus+ and ENPI projects including BITPAL, STEP, DAEDALUS, and Mid-Mobile. His research interests include Security Intelligence, Big Data Analysis, privacy preserving in collaborative environments, secure multiparty computation protocols, distributed computing, and secure distributed systems.