AUTHOR(S): BASHIR, S.A., DOOLAN, D.C. and PETROVSKI, A.

TITLE: Clustering and nearest neighbour based classification approach for mobile activity recognition.

YEAR: 2016

Publisher citation: BASHIR, S.A., DOOLAN, D.C. and PETROVSKI, A. 2016. Clustering and nearest neighbour based classification approach for mobile activity recognition. Journal of mobile multimedia [online], 12 (1/2), pages 100-124. Available from: http://www.rintonpress.com/xjmm12/jmm-12-12/110-124.pdf

OpenAIR citation: BASHIR, S.A., DOOLAN, D.C. and PETROVSKI, A. 2016. Clustering and nearest neighbour based classification approach for mobile activity recognition. Journal of mobile multimedia, 12 (1/2), pages 100-124. Held on OpenAIR [online]. Available from: https://openair.rgu.ac.uk/

CLUSTERING AND NEAREST NEIGHBOUR BASED CLASSIFICATION APPROACH FOR MOBILE ACTIVITY RECOGNITION

SULAIMON A. BASHIR DANIEL C. DOOLAN ANDREI PETROVSKI

School of Computing Science and Digital Media, Robert Gordon University
Aberdeen, UK.
s.a.bashir@rgu.ac.uk d.c.doolan@rgu.ac.uk a.petrovski@rgu.ac.uk

We present a hybridized algorithm based on clustering and nearest neighbour classifier for mobile activity recognition. The algorithm transforms a training dataset into a more compact and reduced representative set that lessens the computational cost on mobile devices. This is achieved by applying clustering on the original dataset with the concept of percentage data retention to direct the operation. After clustering, we extract three reduced and transformed representation of the original dataset to serve as the reference data for nearest neighbour classification. These reduced representative sets can be used for classifying new instances using the nearest neighbour algorithm step on the mobile phone. Experimental evaluation of our proposed approach using real mobile activity recognition dataset shows improved result over the basic KNN algorithm that uses all the training dataset.

Keywords: Activity Recognition, KNN, Smartphones, Clustering

1 Introduction

Activity recognition is a classification task which utilizes labelled data to train a classification algorithm and produces a model that recognizes new unlabelled data. There are two approaches to model induction in mobile activity recognition [1]. The first approach called offline training collects sample data from subjects who perform the designated activities. The collected data is then used to induce a model on a remote system off the mobile device. The induced model is later deployed into the application for recognition. The second approach called online training involves inducing the model directly on the device using the user's self-annotated data. The advantage of the second approach over the first one is that it can produce more accurate and personalised model for individual user. However, it also leads to duplication of efforts for each user to produce self-annotated data. Contrary to this, offline model eliminates duplication of efforts but rather suffers from less personalised model for each individual [2].

Several studies have evaluated different algorithms both in online and offline modes. These studies have reported KNN to give good performance in terms of accuracy in offline training [3, 4, 5, 6]. But despite this performance, KNN is not being used for online recognition on mobile

phones. The reason for this is that KNN requires all the training data to be kept in memory for comparison operation between each test instance and the entire training data. This is cost prohibitive especially for the resource constraint and real time response requirement of the mobile devices in the face of multitasking and multifarious mobile applications that users run on them. Hence, there is need to make KNN amenable for online recognition of activities.

To make KNN amenable to online recognition of mobile activity recognition, we propose an offline data reduction step that reduces the amount of training instances to a desired percentage of the original dataset. The reduced set serves as a good representation of the training set and ensures better accuracy of KNN in an online setting for activity recognition. The evaluation of the proposed framework shows that it performs better than using the basic nearest neighbour algorithm with all the training data. This framework termed 'CKNN' is an extension of our earlier work presented in [15]. The present work includes extension of the approach and extended experimentation with further results.

The rest of this paper is organised as follows: Section 2 and 3 present some of the related work in using KNN and other algorithms for activity recognition. Section 4 presents an overview of the proposed classification framework. Section 5 describes the methodology while Section 6 presents the result and discussion and finally Section 7 gives the conclusion.

2 Nearest Neighbour and Centroid Based Classifiers

Nearest neighbour algorithm is a classification algorithm based on the premise of computing the distance between a test instance and all the training samples and assigns the class label of the closest point to the test instance. The distance computation is performed by means of various distance functions such as Euclidean, Minkowski, and others. More formally, given a training set $D = ((x_1, y_1), (x_2, y_2), ..., (x_n, y_n))\, x_i \in R^n \, and \, y_i \in R^1$ and new sample pattern x_k to be classified, the pattern will be compared with all the training set D i.e. $f(x_k, D)$ using appropriate distance measure f. The label of the point that has the minimum distance to the pattern is then assigned as its class label. Variations to nearest neighbour algorithm strategy includes K-Nearest Neighbour, Fuzzy K-Nearest Neighbour, R-Nearest Neighbour, Modified K-Nearest algorithm among others [7].

Centroid based classification is an extension of nearest neighbour which provides a reduced set of the training samples. The basic nearest centroid classifier also called minimum distance classifier [8, 9] operates by taking the mean or centroid of all the samples in each class present in the dataset. These means form the prototypes to be used for nearest neighbour classification of new instances. Again, given a training set containing instances with feature vector x_i and target variable y_i as $D = ((x_1, y_1), (x_2, y_2), ..., (x_n, y_n))\, x_i \in R^n \, and \, y_i \in R^1$, we can compute the per-class centroids as $\vec{\mu_l} = \frac{1}{|C_i|}\sum \vec{x_i}$ where C_i is the amount of samples belonging to class $i \in y$. A new instance without a label can be classified to the label $\hat{y} = \arg\min_{l \in \mathbf{Y}} \|\vec{\mu_l} - \vec{x}\|$ of the centroid closest to it. The extension to this simple per class centroid is the method of clustering the entire dataset using the number of classes in the dataset to direct the number of clusters and invariably the number of centroids to be created. Another approach involves the clustering of the data in each class to obtain a number of centroids to be used as the prototype samples for the class. We extend this approach by proposing the notion of percentage of data retention desired to direct the amount of clusters to be created in each class present in the entire samples. The notion of percentage data retention eliminates the decision of K clusters

to be performed in clustering. We need to specify only the desired percentage of data to be used based on the accuracy or the storage requirements of the system.

3 KNN in Activity Recognition

KNN is suitable for activity recognition because some of the factors that limit the efficiency of KNN are not prevalent in the domain of inertial based activity recognition. We can observe that the number of features required in activity recognition are often minimal as a sizeable number of them are adequate for recognizing most activities. Furthermore, the magnitudes of the features are small and in similar proportions. Most significantly, the accelerometer values along the three axes are often within similar range of values. These reasons make nearest neighbour algorithm and its extensions a suitable method for pattern recognition and classification in activity recognition. Activity recognition using different sensor modalities and algorithms has been study extensively. A number of machine learning approaches in activity recognition were reviewed in [10]. A more recent review focusing on mobile phone based activity recognition is presented in [1]. The paper identifies many systems that are based on using smartphone sensors for activity recognition. Also, a comparative study of different classifier algorithms from Weka [11] machine learning tool was performed in [12] using data obtained from smartphone accelerometer. The data collected with phone placed in the shirt pocket was used to compare accuracies of IBK, Naive Bayes, Rotation Forest, VFI, DTNB and LMT algorithms while the data collected when the phone was placed in the palm position was used to compare accuracies of SMO, NNge, ClassificationViaRegression, FT, VFI, IBK and Naive Bayes algorithms. Out of all the algorithms tested, they reported IBK and IB1 to give the best accuracy for the hand's palm data and VFI gives the lowest accuracy. The KNN algorithm was not used directly on mobile phone for activity recognition. This can be attributed to the impracticability of using KNN directly for online activity recognition.

Similarly authors in [5, 6] have all shown the superior performance of KNN in terms of accuracy for mobile AR in an offline evaluation scenario. Kose et al. [13] have proposed an improved KNN algorithm for online activity recognition. Initially, the training dataset consists of 4 features: average, minimum, maximum and standard deviation. The algorithm works by selecting k values from the minimum, maximum and average features across each activity data and the standard deviation of the data in each class. The reduced values and the corresponding class tags are employed during recognition phase. The main drawback of this approach is its feature dependence. The algorithm cannot be applied to a dataset with feature characteristics different from the one used in the algorithm. Our approach does not have this limitation as it is applicable to any feature set. Abdallah et. al. [14] have proposed a cluster based classification algorithm that clustered the training datasets into K clusters of the number of activities. The clusters obtained were refined by removing instances of other classes that were mixed-up in a given cluster having majority instances of another class and then the cluster centroids were calculated. The algorithm employs four features computed from each cluster to classify new data. However, this algorithm does not treat new activity data to be classified as individual instances. Rather, it applies clustering to a window of raw accelerometer data and the clusters obtained are compared to the cluster generated from the training data using Euclidean distance, density, gravitational force and within cluster standard deviation. This approach does not segment between one activity and the other. In

addition, the time required to collect enough samples that can be meaningfully clustered will be high for online recognition system that requires immediate and real time feedback of the recognised activity.

4 Overview of Clustering and Nearest Neighbour (CKNN) Based Classification Framework

One of the key reasons for proposing this algorithm is to develop a change detection and adaptive incremental learning strategy for activity recognition using smartphone accelerometer. In this regard, we need an algorithm that is capable of being updated during online classification. Our framework is well suited for this purpose. The framework is based on the application of clustering and nearest neighbour hybrid technique to create a holistic representation of the original dataset into a compact and more discerning set. The framework has two phases of operation- the offline and online phases. The off-line phase involves extensive experimentation to determine the amount of data to be retained in the dataset to ensure more accurate classification with nearest neighbour approach. The main essence of this stage is to reduce the original dataset as much as possible to yield a more compact and informative reduced set that is suitable for on-line recognition. To achieve this, we apply clustering on the original dataset and introduce the notion of percentage data reduction to guide the clustering routine on the number of cluster to create in each class present in the dataset. The clustering of the data in each class to obtain set of centroids will enable us to have a wide representation of the various patterns that may be present within a given class. This will invariably enable the system to leverage on the wide patterns rather than a single centroid pattern per class to classify new samples. We obtained a set of clusters per class from the clustering stage. After this, we extract three cluster parameters from each cluster. Each of these parameters can be used individually as referenced instances or their predictions can be combined into an ensemble predictions depending the level of accuracy desired and system resources available. The parameters are described as follows:

- Centroid or Mean: The mean vector is obtained from the cluster of data by computing the average of the set of vectors present in the cluster. $\vec{\mu_l} = \frac{1}{|C_i|}\sum \vec{x}_i$ where C_i is amount of samples belonging to cluster $i \in 1, ...n$. The set of centroid vectors obtained from clusters created in each class of data present in the dataset represent a reduced and new representative set from the centroid perspective.

- Maximum vectors are list of vectors each in R^n. Each vector contains the maximum value along each feature of the set of instances in a cluster. This implies that each cluster has a maximum vector. The collection of maximum vectors represents another representation or view of the original dataset. The intuition of using this parameter is that once we have the set of maximum vectors across each type of activity we can be able to differentiate high intense activity with high values from low intense activity with low values.

- Minimum vectors are list of vectors each in R^n. Similar to maximum vectors, each vector contains the minimum value along each feature of the set of instances in a cluster. This also represents another view of the original dataset. The intuition of using this parameter is that once we have the set of minimum vectors across each type of activity

we can be able to differentiate low intense activity with low values from high intense activity with high values.

5 Methodology

Our propose approach of online activity recognition employs a data reduction strategy to reduce the initial training set to a more compact set suitable for in-memory use for online recognition. As shown in Algorithm 1, the algorithm takes the training data and the desired percentage of data to retain as input and produces the Model Data (MD).

Algorithm 1: Offline Training

Input: $\boldsymbol{C_n}$, $\boldsymbol{K_n}$ `// number of classes and percentage of data to retain in each class`
Data: $D = (x^i, y^i)$ $x^i \in R^n$ and $y^i \in R^1$ `// Set of training examples`
Result: MD= {Centroid, Minimum, Maximum} `// Data Representation for each Cluster`

1 **foreach** *class* $C_k \in \{C_1, ..., C_n\}$ **do**
2 $data_{C_k}$ = getData(D, C_k) `// get the data samples belonging to class` C_k
3 $centroidVectors_{C_k}$, ClusteredData= Clustering($data_{C_k}$, K_n) `// cluster the data and return centroid vectors and cluster of data`
4 **foreach** $cluster_k \in ClusteredData$ **do**
`/* compute the maximum and minimum vectors from each cluster */`
5 $maximumVector_{C_k}$ = max($cluster_k$)
6 updateMD($maximumVector_{C_k}$,MD)
7 $minimumVector_{C_k}$ = min($cluster_k$)
8 updateMD($minimumVector_{C_k}$,MD)
9 **end**
10 **end**
11 return MD

The model data (MD) is the set of cluster centroids, minimum and maximum vectors obtained after applying clustering on the dataset. In this algorithm, data samples belonging to each class C_k are clustered (lines 1-3 Algorithm 1) by applying a clustering technique on the data. Possible clustering algorithms include k-Means, DBScan, EM and host of others. However, we employ Bisecting K-Means in the present work. After the clustering step, the list of cluster centres obtained for the class of data are stored in a list.

In addition, we extracted the minimum vectors and maximum vectors from each cluster returned for the current class (Algorithm 1 lines 4-8). The number of clusters created per class is dictated by the percentage of data retention input to the algorithm. This step is repeated for each class in the data. Finally, the set of cluster characteristics i.e. centroid , minimum and maximum vectors obtained from the different clusters of each class and their associated labels are returned from the algorithm. These represent the Model Data (MD) to be deployed for the online recognition on a mobile phone. The key feature of the model is that it is more compact and has a reduced resource overhead in terms of memory requirement and time when compared to the ordinary KNN. In addition, the reduced compact set including centroids can be adapted to evolving sensory stream as new unanticipated changes occurs in

the input data distribution.

During the online phase, new instance can be classified by passing it and the MD to Nearest-Neighbour routine. The routine employs Euclidean distance to compute the K-nearest neighbour to the new instance and assigns the majority label of the K nearest point to it (Algorithm 2). Since we have more than one cluster characteristics in the MD, each is considered separately and a majority voting is performed on the outcome of each comparison. The final class given to the new instance is the majority label returned by all of them.

Algorithm 2: Online Classification

Input: $\boldsymbol{x_{new}}$ new unlabelled instance
$\boldsymbol{k}$ number of nearest neighbours
Data: $\boldsymbol{MD}$ compressed training set with characteristics features
Result: y_{new}=predicted class
1 **foreach** $clusterCharacteristics_i$ *in* MD **do**
2 | $prediction_i$ =nearestNeighbour($clusterCharacteristics_i$, x_{new}, k)
3 **end**
4 The output class is $(y_{new}) = argmax_c\ (prediction_c) \quad c = (1...C)$
5 return y_{new}

5.1 Experiments

In this section we describe the experiments conducted to evaluate the applicability and accuracy of the proposed algorithm. We used the Wireless Sensor Data Mining (WISDM) and Human Activity Recognition Using Smartphone (HARS) dataset for the experiments. The two datasets were released to the public for smartphone based activity recognition evaluations.

5.1.1 Dataset Descriptions

1. WISDM Dataset: The WISDM activity recognition dataset [16] was obtained from the accelerometer of mobile phones. The data were collected from 32 users that performed six designated activities of working, jogging, ascending and descending stairs, sitting and standing. Each data sample in the dataset is represented by 43 features. The features were obtained from the transformation of 200 raw samples of data recorded from the tri-axial accelerometer of a mobile phone. Each 200 worth of samples were recorded within a 10 second window with a sampling frequency of 20Hz. The features used were basic statistical features including standard deviation, average, resultants among others [16] described as follows:

   - $X0..X9, Y0..Y9, Z0..Z9$ are set of bins of values representing fraction of accelerometer samples that fell within that bin.
   - $XAVG, YAVG, ZAVG$ these features represent the average of the x, y, and z values in each recorded 200 samples.
   - $XPEAK, YPEAK, ZPEAK$ these features approximate the dominant frequency along the x, y, and z axis values of the accelerometer within each 200 samples.
   - $XABSOLDEV, YABSOLDEV, ZABSOLDEV$ are the average absolute deviations from the mean value for each axis.

- $XSTANDDEV, YSTANDDEV, ZSTANDDEV$ are the standard deviations for each axis.
- $RESULTANT$ is the average of the square roots of the sum of the values of each axis squared $\sqrt{(x_i^2 + y_i^2 + z_i^2)}$.

The dataset distribution spread across the six activities. The total amount of samples in the dataset and their distributions across each activity are shown in Table 1.

Table 1. Distribution of WISDM Dataset

Activity label	**Instances**	**Percentage(%)**
Walking	2081	38.41
Jogging	1625	29.99
Upstairs	633	11.68
Downstairs	528	9.75
Sitting	306	5.65
Standing	245	4.52
Total	5418	100.00

2. Human Activity Recognition Using Smartphone Dataset-HARS The Human Activity Recognition Using Smartphone Dataset [17] is as set of data collected from a set of 30 volunteers who are within an age bracket of 19-48 years. Each subject performed six designated activities of walking, walking-upstairs, walking-downstairs, sitting, standing, and laying while wearing a smartphone attached to their waists. The data were obtained from gyroscope and accelerometer sensors of the smartphone. Each data sample in the dataset is represented by 561 features containing both time domain and frequency domain features and a corresponding activity label. The features were obtained from 128 fixed-width sliding windows of 2.56sec with 50% overlap. The dataset is partitioned into train and test set in 70% and 30% proportion in random fashion across the different users. The training dataset distribution spread across the six activities are shown in Table 2.

Table 2. Distribution of HARS Training Dataset

Activity label	**Amount of Instances**	**Percentage**
Walking	1226	16.68
Walking_Upstairs	1073	14.59
Walking_DownStairs	986	13.41
Sitting	1286	17.49
Standing	1374	18.69
Laying	1407	19.14
	7352	100.00

5.1.2 Experimental Setup

We performed the experiment in two phases. In the first phase, we examined the accuracy of using individual cluster characteristic. We have three characteristics that were used for

classification decision during the online phase (Algorithm 2). The centroid characteristic is the mean of the data points in a cluster. Minimum characteristic is the minimum values across each feature for the data points in a cluster while the maximum characteristic is the minimum values across each feature for the data points in a cluster. We varied the percentage of data retained ranging from 10-90% retention rate. We did not test the 100% retention rate because this will be equivalent to having all the dataset present. Therefore, we obtained the accuracies of using each characteristic as the number of data retention was varied. In the second phase of the experiment all the characteristics were combined to predict the classes of unseen instances used for testing the algorithms. The results obtained for each configuration of the experiment is presented in the next section.

6 Results and Discussion

The accuracy of the propose approach to classification of mobile activity recognition is presented here. The results of using WISDM dataset [16] for evaluation are shown in Tables 3, 4 and 5. The tables show the results for the three different reduced data samples (centroid, maximum and minimum) employed individually by the nearest neighbour to classify the test data. As indicated in Table 3, centroid data give its best accuracy of 81.46% in classifying test instances when nearest neighbour is set to 1 and the percentage of data retained is set to 50% of the original dataset. The accuracy of maximum data characteristic shown in Table 4 has the best accuracy of 80.44% when k=2 and data retained is either 80% or 90% of the original training dataset while the accuracy of minimum data shown in Table 5 for classification decision give its best accuracy of 81.73% when K=1 and data retained is 80%. However, when we combined the predictions from all the three representative data and used the majority voting scheme to select the final class of an instance, the best accuracy of 81.18% is obtained when K=2 with 80% data retention (Table 6).

These results indicate that there is a trade-off between accuracy and the amount of data retained for classification across each of the three data characteristics and their ensemble predictions. Thus, we can select the percentage of data retention based on the level of desired accuracy. For this dataset, we can adopt 50% as the optimal data retention given that the overall best accuracy of 81.46% is achieved with centroid characteristic at this point. Going beyond this percentage of data retention does not yield any high significant increase in accuracy across each of the training data representation and their combination. Therefore, going beyond 50% data retention is ineffectual considering that the corresponding accuracy improvement is not very significant. We also observed that the best accuracy were obtained for the WISDM dataset[16] within the range of nearest neighbours set between 1 and 5 for all the three reduced data and therefore the remaining results from other values of K are omitted.

Table 7 shows the comparison of the accuracy of KNN which utilized all the dataset and the three data representations with their best accuracies obtained at the corresponding data retention level. We can see that the accuracy of basic KNN is lower than the centroid, minimum and the ensemble prediction accuracies. As indicated further in Figure 1, the accuracy of using centroid data for prediction is the best when K=1 and the percentage of data retention is set to 50%. This is followed closely by the ensemble prediction which utilized the combined predictions of the three data representations. Although, the ensemble

Table 3. Centroid Data Accuracy with Varying Percentage of Data Retained on WISDM Dataset

Centroid Data									
K	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**
1	**76.85**	**79.80**	**80.07**	**80.54**	**81.46**	**80.81**	**80.90**	**80.63**	81.00
2	76.01	77.77	78.78	79.80	79.70	80.26	80.63	**81.09**	80.81
3	75.83	78.32	78.78	78.32	79.43	79.24	80.17	79.70	78.60
4	76.20	77.31	77.95	78.60	79.24	78.14	79.34	79.61	79.61
5	75.37	76.57	77.58	78.04	78.51	78.14	79.34	79.34	79.43

Table 4. Maximum Data Accuracy with Varying Percentage of Data Retained on WISDM Dataset

Maximum Data									
K	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**
1	62.55	65.31	**73.15**	**75.28**	77.31	78.87	79.15	79.34	79.98
2	62.92	65.22	72.05	75.00	**77.40**	**79.06**	**80.07**	**80.44**	**80.44**
3	**64.11**	66.42	72.05	73.80	76.11	77.40	77.77	79.34	78.78
4	63.10	65.87	71.77	74.54	77.03	77.95	79.52	78.60	79.43
5	63.10	**66.61**	71.31	73.71	76.38	76.75	79.06	78.51	79.43

Table 5. Minimum Data Accuracy with Varying Percentage of Data Retained on WISDM Dataset

Minimum Data									
K	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**
1	70.48	71.59	75.09	78.04	**77.95**	**79.89**	**79.98**	**81.73**	80.44
2	69.10	71.49	75.37	77.03	77.58	79.15	79.52	80.72	**80.90**
3	69.37	73.43	75.83	76.57	77.68	78.97	78.78	79.80	79.61
4	69.83	73.62	75.55	77.31	77.86	78.23	79.61	80.72	80.35
5	69.46	72.42	75.00	76.94	77.49	78.14	78.41	79.52	79.98

Table 6. Ensemble Prediction Accuracy on WISDM Dataset

Ensemble Prediction									
K	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**
1	**73.71**	**76.85**	**78.32**	**79.61**	**79.80**	**81.18**	**80.81**	81.09	80.35
2	69.74	72.79	76.01	78.14	78.23	79.34	80.07	**81.18**	**81.00**
3	70.20	73.80	75.74	76.66	77.58	78.14	79.52	80.26	79.24
4	70.39	73.43	75.55	76.66	77.68	77.58	79.98	80.26	79.89
5	69.46	72.88	75.18	76.38	77.58	77.49	79.06	79.06	79.80

prediction is not suitable for this dataset because it will require the retention of 60% each for the three data representation. However, the minimum data representation which utilized 80% of the data is also better than that of the KNN which utilized all the data samples as reference set in classifying new unseen instances. Maximum data representation is the only data representation that gives a less accuracy compared to KNN and the margin of difference is very low. In general, we can see the benefits of using reduced samples over the

entire samples for nearest neighbour classification. It is clear that with this approach, we can be able to transform a training set into reduced set using clustering and extract useful features from the clusters to serve as more informative reference sets for the nearest neighbour classification. It should be noted that the general low accuracy below 90% for this dataset can be attributed to the nature of the dataset in terms quality of features and the subjects used for data collection. The dataset contains data from 32 different users of varying characteristics in performing the designated activity. This produces many variations in the training and testing data. Nevertheless, the performance of our transformed data are good given the fact that they can use a reduced dataset for online recognition compare to KNN that requires the entire training instance to achieve good performance.

Table 7. Different Data Accuracy Compared with KNN Accuracy on WISDM Dataset

KNN	Centroid-50%	Max-80%	Min-80%	Ensemble Prediction-60%
80.90	**81.46**	79.34	**81.73**	**81.18**
80.99	79.70	**80.44**	80.72	79.34
79.89	79.43	79.34	79.80	78.14
80.07	79.24	78.60	80.72	77.58
79.98	78.51	78.51	79.52	77.49

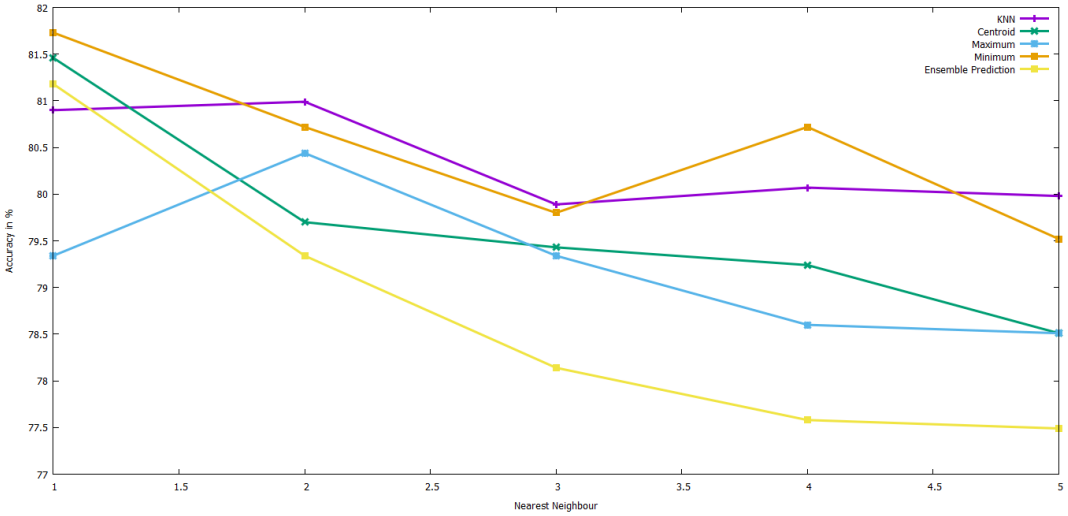

Fig. 1. Accuracy of Using Different Data Transformation and KNN Algorithm with WISDM Dataset

For the second dataset, the results obtained are presented in Tables 8, 9 and 10. The tables show the results for the three different representation (centroid, maximum and minimum) of the original dataset with varying amount of data retention employed to classify test instances. As indicated in Table 8, centroid data representation gives its best accuracy of 91.65% in classifying test data when nearest neighbour is set to 8 and the percentage of data retained is 40%. The best accuracy of using maximum data representation in place of the original dataset is 89.96%. This is obtained when K=6 and data retained is 80% as shown in Table 9.

Similarly, the best accuracy obtained for using minimum data representation is 90.87% when K=7 and data retained is at 80% as shown in Table 10. Moreover, the accuracy of ensemble predictions of all the three data representations yielded the best accuracy of 91.01% at the point when K=7 and a data retention of 60% as shown in Table 11.

Table 8. Centroid Data Accuracy with Varying Percentage of Data Retained on HARS Dataset

Centroid Data									
K	10%	20%	30%	40%	50%	60%	70%	80%	90%
1	89.45	88.77	89.58	88.77	89.38	87.85	88.16	87.58	88.16
2	87.65	88.29	89.07	88.39	88.33	87.00	86.94	86.80	86.94
3	90.33	89.85	90.70	89.75	90.02	88.97	89.35	89.41	89.35
4	90.50	90.60	90.80	90.30	90.13	89.07	89.62	89.45	89.62
5	90.50	90.40	90.63	90.46	90.70	89.89	89.99	90.13	89.99
6	90.77	**91.14**	**91.28**	91.35	90.97	90.23	90.23	90.26	90.23
7	90.74	90.74	90.77	91.21	90.91	90.74	90.46	89.92	90.46
8	91.14	91.11	90.91	**91.65**	**91.11**	90.94	90.23	90.23	90.23
9	90.40	90.40	90.46	91.11	90.60	90.87	**90.57**	**90.57**	**90.57**
10	**91.62**	90.70	90.84	91.52	91.01	**91.14**	90.40	90.50	90.40

Table 9. Maximum Data Accuracy with Varying Percentage of Data Retained on HARS Dataset

Maximum Data									
K	10%	20%	30%	40%	50%	60%	70%	80%	90%
1	66.03	79.00	82.46	83.58	85.58	86.46	87.17	87.61	87.17
2	62.71	73.19	78.96	80.86	83.24	84.66	85.44	85.88	85.44
3	71.46	80.12	85.31	85.82	87.51	88.26	88.67	89.18	88.67
4	69.19	78.62	84.32	85.07	87.00	87.65	88.19	88.94	88.19
5	72.18	80.56	85.04	86.26	87.72	89.21	89.45	89.79	89.45
6	71.16	79.88	85.24	86.19	87.55	88.70	89.01	**89.96**	89.01
7	72.82	81.37	85.37	**86.63**	88.19	**89.51**	88.94	89.62	88.94
8	72.58	**81.64**	85.61	86.56	88.09	88.84	89.24	89.18	89.24
9	**73.94**	81.37	**86.09**	86.46	**88.80**	89.11	**89.48**	89.75	**89.48**
10	73.29	81.34	85.95	86.09	88.67	89.11	89.35	89.75	89.35

Table 12 shows the comparison of the accuracy of KNN and the three data representations. The KNN utilized all the training dataset while the centroid, minimum and maximum data representation utilized a reduced and transformed form of the original dataset. We can see that the best accuracy of 90.77% is obtained for KNN when the nearest neighbour for deciding the final label of an instance is set to 8. This accuracy is less than the centroid data representation with K ranges between 1 and 10 and percentage of data retention between 10% and 60% level. This shows that a reduced set using centroid data is better than the KNN approach that uses all the training dataset. Since other characteristics utilized lower amount of data, their accuracy can be traded-off for the smaller amount of data required when compared to the basic KNN that utilized all the training data. Based on these results, we can select the percentage of data retention based on the level of desired accuracy and available computational resource on the mobile device. For this dataset, we can take the

Table 10. Minimum Data Accuracy with Varying Percentage of Data Retained on HARS Dataset

Minimum Data									
K	10%	20%	30%	40%	50%	60%	70%	80%	90%
1	74.69	79.50	83.92	85.31	86.73	87.17	87.44	87.72	87.44
2	70.41	74.55	80.56	82.05	84.26	85.04	85.92	86.02	85.92
3	77.13	82.22	86.73	86.87	88.53	89.24	89.58	89.72	89.58
4	74.89	79.54	85.14	85.71	87.51	88.50	89.04	89.58	89.04
5	77.20	82.97	86.22	87.78	89.18	89.99	90.36	90.53	90.36
6	76.42	81.81	86.60	86.77	88.97	89.68	90.23	90.33	90.23
7	77.50	84.39	87.31	87.58	89.38	**90.70**	**90.43**	**90.87**	**90.43**
8	76.89	83.61	**87.31**	87.34	88.90	89.92	90.06	90.46	90.06
9	**77.71**	**85.27**	87.17	**88.02**	**89.68**	90.23	89.99	90.40	89.99
10	76.93	84.56	87.21	87.82	89.04	90.26	89.99	90.53	89.99

Table 11. Ensemble Prediction Accuracy on HARS Dataset

Ensemble Predictions									
K	10%	20%	30%	40%	50%	60%	70%	80%	90%
1	78.18	84.73	87.95	87.72	88.50	88.16	87.99	87.92	87.99
2	73.97	79.71	84.32	85.17	86.09	86.16	86.53	86.66	86.53
3	80.18	86.63	**89.68**	89.14	90.16	89.85	90.02	89.99	90.02
4	78.69	84.83	88.46	88.56	89.68	89.58	89.41	89.75	89.41
5	79.64	86.53	88.43	89.62	**90.50**	90.91	**90.70**	**90.77**	90.70
6	79.67	85.88	88.94	89.38	90.43	90.53	90.33	90.74	90.33
7	80.22	86.60	89.24	**89.99**	90.33	**91.01**	90.70	90.70	**90.70**
8	80.32	87.04	89.11	89.96	89.92	90.57	90.13	90.63	90.13
9	80.25	87.11	89.07	89.62	90.16	90.67	90.33	90.50	90.33
10	**80.59**	**87.28**	88.87	89.68	89.79	90.97	90.09	90.57	90.09

Table 12. Different Data Accuracy Compared with KNN Accuracy on HARS Dataset

K	**KNN**	**Centroids-40%**	**Max-80%**	**Min-80%**	**Ensemble Prediction-60%**
1	88.19	88.77	87.61	87.72	88.16
2	86.29	88.39	85.88	86.02	86.16
3	89.35	89.75	89.18	89.72	89.85
4	89.11	90.30	88.94	89.58	89.58
5	89.99	90.46	89.79	90.53	90.91
6	89.99	91.35	**89.96**	90.33	90.53
7	90.50	91.21	89.62	**90.87**	**91.01**
8	**90.77**	**91.65**	89.18	90.46	90.57
9	90.53	91.11	89.75	90.40	90.67
10	90.40	91.52	89.75	90.53	90.97

10% data retention and the centroid data as the reference training data for classifying new insistences. Figure 2 shows the comparative accuracy of using each of the three reduced data as training set for nearest neighbour classification and their ensemble predictions. We can observe that the accuracy of minimum and centroid data are better than the KNN. This is

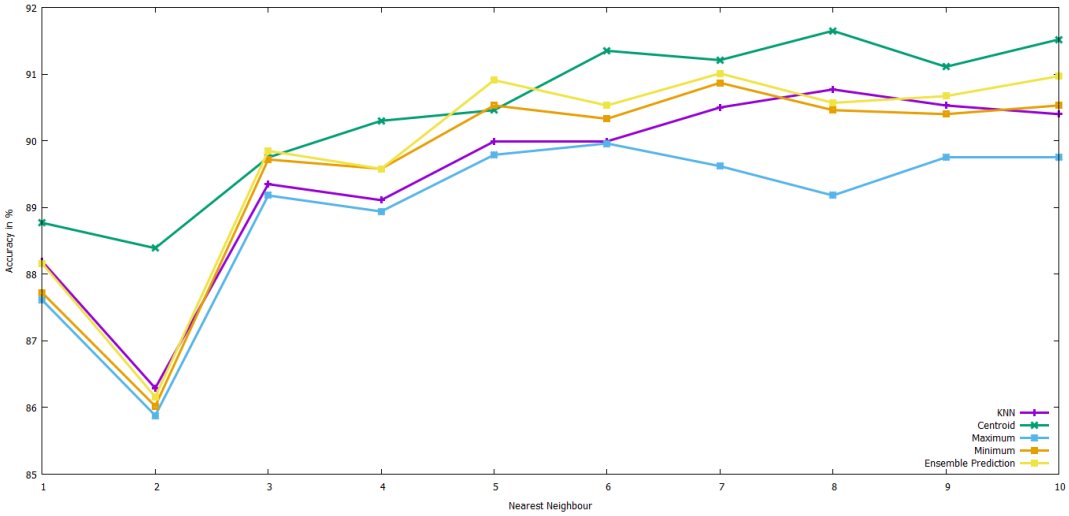

Fig. 2. Accuracy of Using Different Data Transformation and KNN Algorithm with HARS Dataset

more important since the two use a less amount of data for prediction. Also the accuracy of the ensemble prediction is also better when k=7 and the data reduction rate is 60%. It is only the maximum data that gives a lesser accuracy than KNN although the difference is very small. With this, we can conclude that our approach of reducing the dataset with different data transformation that is extracted from the clusters is better than using the KNN with all the training data for predictions.

7 Conclusion

This paper has identified the drawback of using KNN for online activity recognition on mobile phone. The drawback of KNN in terms of keeping all the large amount of training data available at recognition time is addressed by proposing a hybrid approach to classification. The algorithm is based on the concept of clustering and nearest neighbour classification. The novel approach employs bisecting k-Means algorithm to cluster the training instances. We introduce the concept of percentage data retention to direct the amount of clusters to create in each class present in the dataset. The set of clusters obtained and the other informative data extracted from each cluster serve as the reference set for the online phase of the nearest neighbour classification. This process retain proportional amount of data across each class of data in the original dataset. The evaluation of the approach shows that it performed better than basic KNN algorithm on a real mobile activity recognition dataset.

Acknowledgments

This research was supported by the National Information Technology Development Fund (NITDEF) under the auspices of the National Information Technology Development Agency Abuja, Nigeria.

References

[1] Muhammad Shoaib, Stephan Bosch, Ozlem Durmaz Incel, Hans Scholten, and Paul JM Havinga. A survey of online activity recognition using mobile phones. *Sensors*, 15(1):2059–2085, 2015.

[2] Gary M Weiss and Jeffrey W Lockhart. The impact of personalization on smartphone-based activity recognition. In *AAAI Workshop on Activity Context Representation: Techniques and Languages*, 2012.

[3] Sulaimon A Bashir, Daniel C Doolan, and Andrei Petrovski. The impact of feature vector length on activity recognition accuracy on mobile phone. *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2015, 1-3 July, 2015, London, U.K.*, 1:332–337, 2015.

[4] Sulaimon A Bashir, Daniel C Doolan, and Andrei Petrovski. The effect of window length on accuracy of smartphone-based activity recognition. *IAENG International Journal of Computer Science*, 23(1), 2016.

[5] Sian Lun Lau and K. David. Movement recognition using the accelerometer in smartphones. In *Future Network and Mobile Summit, 2010*, pages 1–9, June 2010.

[6] Zoltán Prekopcsák, Sugárka Soha, Tamás Henk, and Csaba Gáspár-Papanek. *Activity recognition for personal time management*. Springer, 2009.

[7] Nitin Bhatia et al. Survey of nearest neighbor techniques. *arXiv preprint arXiv:1007.0085*, 2010.

[8] T Hastie, R Tibshirani, and J Friedman. *The Elements of Statistical Learning Data Mining, Inference, and Prediction*. Springer Verlag, New York, 2nd edition edition, 2009.

[9] M Narasimha Murty and V Susheela Devi. *Pattern recognition: An algorithmic approach*. Springer Science & Business Media, 2011.

[10] Stephen J Preece, John Yannis Goulermas, Laurence PJ Kenney, and David Howard. A comparison of feature extraction methods for the classification of dynamic activities from accelerometer data. *IEEE Transactions on Biomedical Engineering*, 56(3):871–879, 2009.

[11] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[12] Media Anugerah Ayu, Siti Aisyah Ismail, Ahmad Faridi Abdul Matin, and Teddy Mantoro. A comparison study of classifier algorithms for mobile-phone's accelerometer based activity recognition. *Procedia Engineering*, 41:224–229, 2012.

[13] Mustafa Kose, Ozlem Durmaz Incel, and Cem Ersoy. Online human activity recognition on smart phones. In *Workshop on Mobile Sensing: From Smartphones and Wearables to Big Data*, pages 11–15, 2012.

[14] Zahraa Said Abdallah, Mohamed Medhat Gaber, Bala Srinivasan, and Shonali Krishnaswamy. Cbars: Cluster based classification for activity recognition systems. In *Advanced Machine Learning Technologies and Applications*, pages 82–91. Springer, 2012.

[15] Sulaimon A. Bashir, Daniel C Doolan, and Andrei Petrovski. Clusternn: A hybrid classification approach to mobile activity recognition. In Liming Chen, Matthias Steinbauer, Ismail Khalil, and Gabriele Kotsis, editors, *Proceedings of the 13th International Conference on Advances in Mobile Computing & Multimedia (MoMM2015)*, pages 263–267, 2015.

[16] Jennifer R Kwapisz, Gary M Weiss, and Samuel A Moore. Activity recognition using cell phone accelerometers. *ACM SigKDD Explorations Newsletter*, 12(2):74–82, 2011.

[17] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, and Jorge Luis Reyes-Ortiz. A public domain dataset for human activity recognition using smartphones. In *21th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2013*, 2013.