Comparison of Expectation Maximization and K-means Clustering Algorithms with Ensemble Classifier Model

M.N.SHAH ZAINUDIN[1,2], MD NASIR SULAIMAN[1], NORWATI MUSTAPHA[1], RAIHANI MOHAMED[1]
[1]Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
UPM Serdang, 43400 Serdang, Selangor
MALAYSIA

[2]Faculty of Electronics and Computer Engineering
Universiti Teknikal Malaysia Melaka
Hang Tuah Jaya, 76100 Durian Tunggal, Melaka
MALAYSIA
noorazlan@utem.edu.my, nasir@upm.edu.my, norwati@upm.edu.my, raihanim@gmail.com

Abstract: - In data mining, classification learning is broadly categorized into two categories; supervised and unsupervised. In the former category, the training example is learned and the hidden class is predicted to represent the appropriate class. The class is known, but it is hidden from the learning model. Unlike supervised, unsupervised directly build the learning model for unlabeled example. Clustering is one of the means in data mining of predicting the class based on separating the data categories from similar features. Expectation maximization (EM) is one of the representatives clustering algorithms which have broadly applied in solving classification problems by improving the density of data using the probability density function. Meanwhile, K-means clustering algorithm has also been reported has widely known for solving most unsupervised classification problems. Unlike EM, K-means performs the clustering by measuring the distance between the data centroid and the object within the same cluster. On top of that, random forest ensemble classifier model has reported successive perform in most classification and pattern recognition problems. The expanding of randomness layer in the traditional decision tree is able to increase the diversity of classification accuracy. However, the combination of clustering and classification algorithm might rarely be explored, particularly in the context of an ensemble classifier model. Furthermore, the classification using original attribute might not guarantee to achieve high accuracy. In such states, it could be possible some of the attributes might overlap or may redundant and also might incorrectly place in its particular cluster. Hence, this situation is believed in yielding of decreasing the classification accuracy. In this article, we present the exploration on the combination of the clustering based algorithm with an ensemble classification learning. EM and K-means clustering algorithms are used to cluster the multi-class classification attribute according to its relevance criteria and afterward, the clustered attributes are classified using an ensemble random forest classifier model. In our experimental analysis, ten widely used datasets from UCI Machine Learning Repository and additional two accelerometer human activity recognition datasets are utilized.

Key-Words: - Expectation maximization, K-means, random forest, clustering, classification

1 Introduction

Unsupervised learning such as clustering algorithm works by finding the similar data for unlabeled example by separating the data according to their similar nature [1]. Clustering means to divide the dataset into a plurality of the cluster's data by sharing some trait of each subset. In such states, the distance function method is used to measure the similarity or proximity of each data item [2]. There are several categories of the clustering algorithm, namely hierarchical-based, partitioned-based, density-based, grid-based and model-based [3]. The detailed explanation of each clustering category could be referred from [3]. Partitioned-based clustering algorithm has reported widely been applied by partitioned the object with some membership matrices. The object is determined according to a specific group using membership functions by observing the pertaining between the object and the group. The most commonly known clustering algorithm is K-means and expectation

maximization (EM) [1]. Both algorithms work as similar by allowing the model to refine the iterative process to find the best congestion. However, EM uses statistical methods to measure the relationship between two data items. Unlike EM, K-means utilize the distance functions such as Euclidian or Manhattan distance to calculate the distance between each of two data items.

In many situations, these clustering algorithms usually being applied for unlabeled example and might be rarely used in a labeled example. Even though clustering algorithm such K-means have proven to be successful in integrating with decision tree in solving human activity recognition [4], however, the combination of clustering algorithm with an ensemble classifier model are less reported. Moreover, the integration between clustering algorithm with other supervised classification algorithms, particularly in an ensemble decision tree model has reported infrequently [5]. An ensemble classifier model works by combining more than one learning model to maximize the accuracy performance. This classifier model has reported success in both unsupervised and supervised learning. In supervised, the assumption of the object is classified one time without relying on the dependencies of each object. However, it is hard to predict the unseen object with a very less number of examples. Contrasting with supervised, unsupervised considering the object relationship to predict the unseen object. For instance, a pair of objects with closely each other are more likely belonging to the same class than those are far apart from each other.

On the other hand, to predict the less number of attributes is believed as another challenge. In many cases, to predict the multi-class problems considered more difficult than two-class problems using very less number of attributes [6]. Hence, the transforming the attributes into a number of respective clusters become a solution. By proposing the attribute into the respected cluster is believed could help the learning model more able to effectively learn the class characteristic. This article consists of threefold. The EM and K-means clustering algorithm is proposed by transforming the original data into a clustered attribute according to their membership function criteria to simplify the problem complexity. In order to evaluate the ability of clustering algorithm with the respect of ensemble learning model, the clustered attributes are classified using random forest classifier. In addition, the proposed method has also been compared with several additional classifier models including J48, k-nearest neighbor (KNN), support vector machine (SVM) and random forest (RF). The experimental work is evaluated on several well-known datasets which are downloaded from UCI Machine Learning Repository. Moreover, we also experimented the proposed algorithm through two accelerometer activity datasets.

2 Materials and Methods

In this section, the detail explanation of the proposed algorithm is described. The deliberation begins with the description of the dataset used, the proposed hybrid clustering for simplifying the problem complexity in the context of an ensemble learning model.

2.1 Datasets

In this work, ten datasets are retrieved from UCI Machine Learning Repository. The dataset consisting of a distinct number of classes and containing varying types of attributes including nominal and numbering. Moreover, we also experimented our proposed hybrid algorithms throughout the accelerometer human activity dataset. Two accelerometer activity datasets recorded by [7] and [8] are utilized. In [7], single accelerometer sensor placed on the human front pant pocket is used, while in [8], four different accelerometer sensors including arm, belt, pocket and wrist are used. Total six different types of activities are collected from both datasets. The detailed explanation of each dataset could be referred from their articles respectively. In order to evaluate the overall accuracy of the ensemble classifier model, majority voting strategy is applied. In each experiment, 10-fold cross validation strategy is used to measure the subset performance. The detail description of the datasets used in the experimentation is tabulated in Table 1.

Table 1. Description of the dataset

Name	Attributes	Examples	Class
D1-Balance scale	4	625	3
D2-Car evaluation	6	1728	4
D3-Dermatology	34	366	6
D4-Ecoli	8	336	8
D5-Glass identification	10	214	7
D6-Hayes-Roth	5	160	3
D7-Iris	4	150	3
D8-Lenses	4	24	3
D9-Soybean	35	47	4

D10-Statlog	18	846	4
D11-WISDM [7]	36	21378	6
D12-Shoaib [8]	36	22946	6

2.2 Clustering algorithm

Clustering is an unsupervised learning which aims to find the structure of the example. For instance, K-means, affinity propagation, mean shift and spectral clustering has reported able to increase the accuracy while minimizing the computational complexity [9]. The collection of unlabeled data is explored to decide which group of the example belongs. Hence, a cluster is described as a collection of objects which are similar to each other and dissimilar with the object belonging to the other clusters [10]. However, the chosen of the best criterion of the clustering algorithm could be considered as an important aspect. The selection of the clustering algorithm not only constitutes good in maximizing the decision boundary but also needs outfits to the particular needs. As mentioned previously, clustering is broadly categorized in several categories but more characterized into two different types, namely distance-based clustering and conceptual-based clustering [2]. Distance-based clustering works by identifying the object within the same cluster based on the distance measurement. Conceptual-based clustering, on the other hand works by grouping the object based on the descriptive concept rather than according to the simple similarity measures. On top of that, there are numbers of clustering algorithm has been introduced. Four well-known clustering algorithms which are extensively reviewed: exclusive clustering, overlapping clustering, hierarchical clustering and probabilistic clustering. However, in this work, we elaborate the theory of the most commonly used clustering algorithms such as K-means and EM.

Exclusive clustering: if a certain object belongs to a definite cluster so that it could not be included in another cluster.

Overlapping cluster: cluster data is defined using fuzzy sets so that each point may belong to one or more clusters according to its membership values. Each object might receive different membership values.

Hierarchical cluster: the cluster is defined based on the union between two nearest clusters and the initial condition is determined. The final cluster is obtained after a few iterations done.

Probabilistic cluster: the cluster is defined using a probabilistic approach.

2.2.1 K-means clustering algorithm

K-means clustering algorithm works by partitioning a collection of data into a k number of clusters so that each data point is assigned to the cluster with the nearest mean. This clustering algorithm has also been reported in various applications, particularly in data mining [11], [12]. There are two main phases in K-means; the first phase is to calculate the k centroid while the second phase is to define each point to the cluster, which has the nearest centroid from the respective data point [11]. In order to define the nearest centroid, distance measurement function is used. Euclidean distance is one of the common distance function which is widely applied in K-means. After the grouping is completed, the new centroid of each cluster is recalculated. Euclidean distance is calculated between each center and each data point and the point in the cluster are assigned for the distance with minimum values. The centroid of each cluster is the point which the sum of the distances of all objects in the cluster. Hence, this algorithm will minimize the sum of the distances from each object with the centroid for all clusters [13]. The detailed algorithm of K-means [1] is illustrated in Table 2.

Table 2. K-means clustering algorithm

Input: number of clusters (k), number of observations (n)
Output: set of k-clusters
1. Choose k objects as initial cluster centers
2. Repeat
3. (Re) assign each object to the cluster, which has similar mean value of the object in the cluster
4. Update cluster means
5. Stop until there are no changes

2.2.2 Expectation maximization (EM) clustering algorithm

The EM clustering algorithm is introduced by the theory of Gaussian Mixture Model (GMM) which is used to improve the density of a given set of sample data. The probability density function of a single density estimation method with multiple Gaussian

probability density functions is used to model the data distribution [12], [14]. The EM clustering algorithm used to estimate the optimal model by estimating the parameter for each Gaussian blend component of the sample data set to maximize the log-likelihood. When the data is missing or incomplete, the EM algorithm uses a random variable to find the optimal parameter of the hidden distribution function of the given data. The detailed algorithm of EM [1] is illustrated in Table 3.

Table 3. EM clustering algorithm

Input: number of clusters (k), number of observations (n) **Output:** set of k-clusters with weight that maximize the log-likelihood
1. Expectation step: for each observation x, calculate membership probability values of x in each cluster h=1,…, k 2. Maximization step: update mixture model parameter (probability weight) 3. Stopping criteria: the algorithm stop if reaches to stopping criteria, otherwise j=j+1, repeat step 1

2.3 Ensemble learning model

An ensemble learning model works by combining more than one learning model to maximize the ability of the final predictive performance [15]. The weak classifier model is usually reinforced by the strongest classifier model and afterward, the probability is calculated to produce the final prediction results. Bagging is one of the ensemble methods that could reduce the data variance by creating the several subsets of the sample. This process is done by random sampling and also been called as a bootstrap. Meanwhile, boosting uses weight function to predict the performance. This method iteratively learns the model and all observations are given as an equal weight. The weight is increased for an incorrectly classified sample by observing the previous successive tree. Otherwise, the weight is decreased. Unlike boosting, bagging does not depend on the previously created tree. However, both of these methods share the same criteria where the majority voting is used to define the final prediction results [16]. On the other hand, due to the limitation of the ordinal decision tree which incapable to handle with missing values and impractical for a large number of attributes, random forest classifier is introduced [17]. In this method, several numbers of decision trees are combined by adding randomness layers to maximize the generalization ability than a single decision tree. The n-trees are randomly generated and each observation is predicted using each generated decision tree. The class who received the highest vote will be classified as final prediction results. Rather than effectively performed in various classification problems, a random forest also able to reduce the potential of overfitting.

3 Results and Discussion

In this section, we emphasize the detail experimental analysis and result regarding the proposed algorithm. For a fair comparison, we also compare the performance of our work with previously reported work. In order to evaluate the subset performance, 10-fold cross-validation is applied and average accuracy is used to measure the final prediction results. We also compare the performance of several state-of-the-art classification algorithms throughout of all ten datasets. In K-means, 500 iterations were selected and the number of k is initialized according to the number of distinct class values of each data set. On the other hand, default parameter values were used in EM. The maximum number of clusters in EM is initially defined by the cross-validation. So that, both clustering algorithms were evaluated using random forest classifier model. We generate 100 trees for each experimental analysis. Table 4 shows the experimental result of each data set using both clustering algorithms K-means and EM accordingly. The best results emphasize throughout **boldface**.

Table 4. Experimental results of K-means and EM clustering algorithm accordingly

		K-RF		EM-RF	
Data	**k**	**Clusters**	**Acc**	**Clusters**	**Acc**
D1	3	9	**0.998**	13	0.914
D2	4	16	0.968	16	**0.972**
D3	6	36	0.986	17	**0.989**
D4	8	49	0.848	16	**0.875**
D5	7	42	**0.818**	17	0.771
D6	3	9	0.803	13	**0.841**
D7	3	9	**0.953**	8	0.947
D8	3	9	0.792	4	**0.917**
D9	4	16	1.000	5	1.000
D10	4	16	**0.739**	36	0.726
Avg		21.1	0.891	14.5	0.895

From the table, it is clearly being seen that the average accuracy recorded from both clustering

algorithms (K-means and EM) has somewhat comparable performance. The classification accuracy of EM showed was 0.4% better than obtained by K-means. On average, about 68% number of clusters could be reduced from EM. Soybean has recorded similar accuracy, but the number of generating clusters using EM comparatively less than K-means. However, lenses are shown to greatly differ in terms of accuracy where the accuracy of EM (91.7%) has recorded somewhat higher than K-means (79.2%). Car evaluation and iris have shown reasonable performance for both algorithms even though an equal number of clusters were generated. On the other hand, we also compare the performance of work with several state-of-the-art classifier models. In such circumstances, we classify the data point without applying clustering algorithms. Table 5 shows the comparative performance of our work with several additional classifier models such as J48, KNN, SVM and RF.

Table 5. Comparative performance with additional classifier models

Data	J48	KNN	SVM	RF	K-RF	EM-RF
D1	0.766	0.848	0.898	0.814	**0.998**	0.914
D2	0.964	0.946	0.940	0.963	0.968	**0.972**
D3	0.959	0.954	0.940	0.970	0.986	**0.989**
D4	0.842	0.804	0.426	0.658	0.848	**0.875**
D5	0.659	0.706	0.692	0.799	**0.818**	0.771
D6	0.811	0.705	0.394	0.841	0.803	0.841
D7	0.960	0.953	**0.967**	0.953	0.953	0.947
D8	0.792	0.875	0.583	0.792	0.792	**0.917**
D9	0.979	1.000	1.000	1.000	1.000	1.000
D10	0.726	0.699	0.305	0.760	**0.739**	0.726
Avg	0.846	0.849	0.715	0.855	0.891	0.895

It is clearly being seen that the accuracy of clustered attributes is capable to achieve high accuracy. However, SVM is able to obtain the highest accuracy (96.7%) followed by J48 for iris. In addition, the accuracy of soybean showed somewhat comparable performance from all classifier models. The poorest performance was recorded from SVM where about 71% of accuracy was obtained on average. Though, an average accuracy recorded by J48, KNN and RF showed fairly comparable. In comparison, we also compare the performance of this work with previously reported work [18]. The author utilized regular random forest as a classifier and the evaluation was done through the same dataset. Table 6 shows the comparative performance of regular random forest with our work.

Table 6. Comparative performance of RF with K-RF and EM-RF

Data	RF	K-RF	EM-RF
D1	0.837	**0.998**	0.914
D2	0.742	0.968	**0.972**
D3	0.856	0.986	**0.989**
D4	0.845	0.848	**0.875**
D5	0.661	**0.818**	0.771
D6	0.634	0.803	**0.841**
D7	0.953	0.953	0.947
D8	0.717	0.792	**0.917**
D9	0.991	1.000	1.000
D10	0.712	**0.739**	0.726
Avg	0.795	**0.891**	**0.895**

Table 6 shows the classification with clustering algorithm were recorded higher than classification without clustering algorithm. The accuracy of iris has recorded by RF comparable similar with K-RF. Moreover, the highest accuracy recorded from soybean was 99.1, however, the result obtained by clustering with K-means and EM algorithms are able to overtake the accuracy up to 100%. Car evaluation and Hayes-Roth was shown the poorest performance when classifying the attribute without clustering. Hence, it could be summarized that the accuracy tends to increase when the clustering algorithm is fitted into a random forest classifier. About 10% of the increment of accuracy would be obtained when expanding the attribute into several numbers of clusters before it fed into classifier model. On the other hand, we also evaluate the proposed work through some additional datasets. In such states, we tested our hybrid algorithms for the accelerometer activity recognition dataset as described in section 2.1. Fig. 1 presents the classification results of the WISDM acceleration dataset.

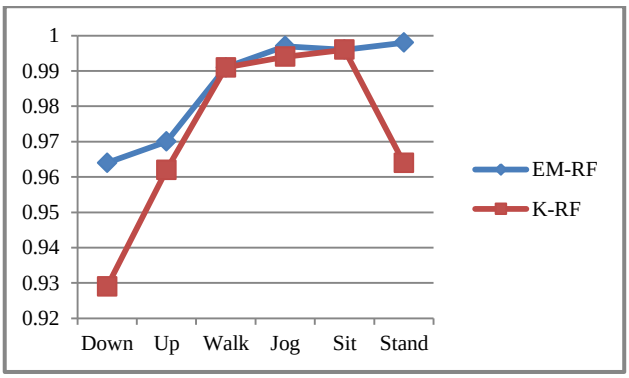

Fig. 1. Classification result of WISDM

It is clearly being seen that the accuracy received by EM outperformed than K-means, particularly in recognizing downstairs and standing. The accuracy of walking and sitting showed somewhat comparable using both clustering algorithms. Afterward, we also experimented our algorithm with [8] dataset. In comparison with WISDM, four different sensor placements are utilized. Hence, we compare the performance of our proposed algorithm for each sensor placement. Fig. 2 (a) – (d) show the classification result of each sensor placement; arm, belt, pocket and wrist accordingly.

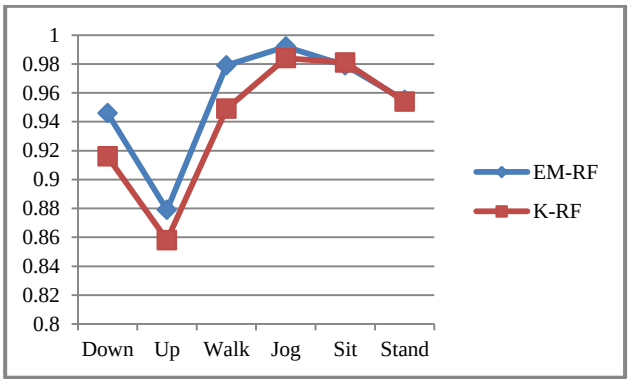

Fig. 2 (a). Classification result of arm placement

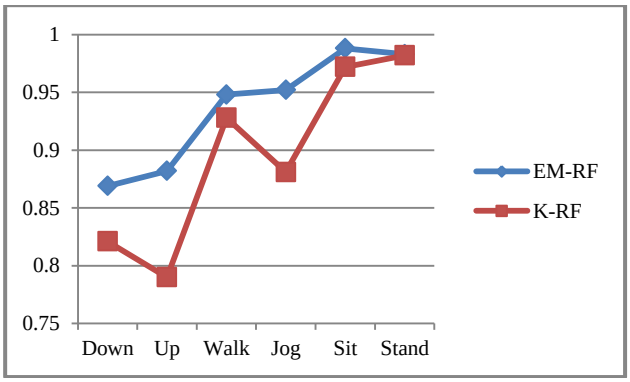

Fig. 2 (b). Classification result of belt placement

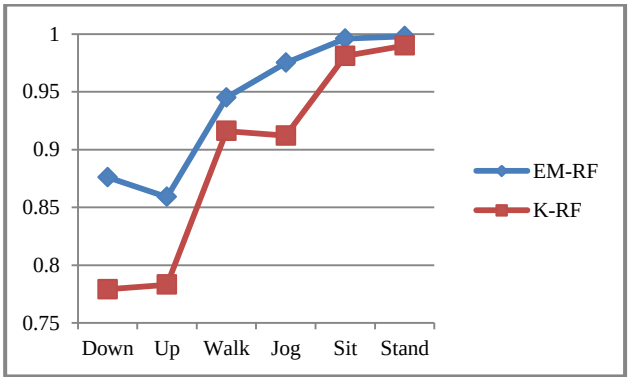

Fig. 2 (c). Classification result of pocket placement

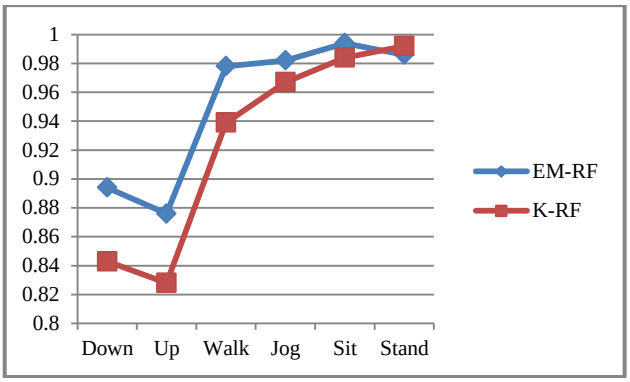

Fig. 2 (d). Classification result of wrist placement

The accuracy of EM has clearly outperformed than K-means even though the time of generating the cluster membership of EM was slightly longer than K-means. On average, EM is able to produce high accuracy in differentiating most of the activities. Yet, some activities such as sitting and standing shown somewhat similar with K-means. The overall time required (in seconds) to execute the cluster attribute is shown in Fig. 3. Hence, it could be summarized that by proposing our hybrid clustering algorithm with an ensemble classifier model is able to increase the accuracy of various types of datasets. The introducing of clustering attributes has also able to increase the diversity of differentiating numerous acceleration activities.

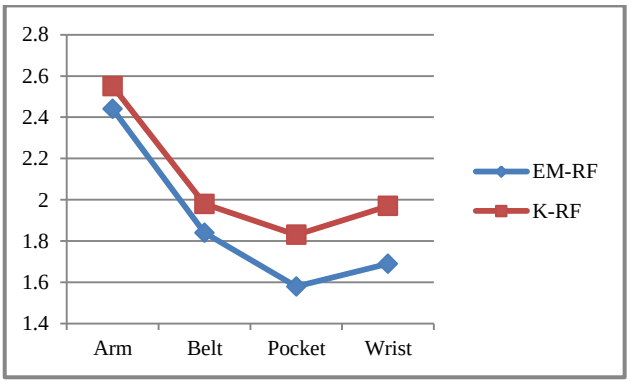

Fig. 3. Time required (in seconds) to execute the clustered attribute

4 Conclusion

In this article, two clustering algorithms, namely K-means and EM were applied to transform the original attribute into particular clusters. Both clustering algorithms were compared and evaluated using an ensemble random forest learning model. In order to evaluate the performance of the proposed hybrid algorithm, ten well-known datasets which are downloaded from UCI Machine Learning

Repository was utilized. The experimental results show that both clustering algorithms showed better performance in accuracy than previously reported work. The EM clustering algorithm performed somewhat better (89.5%) that K-means clustering algorithm (89.1%) on average when evaluating it using random forest learning model. We also evaluate our proposed hybrid algorithm through the accelerometer human activity dataset. The EM clustering algorithm is able to outperform the K-means of distinguishing various types of physical activities. Hence, the clustering algorithm is able to show more accurate classification as compared with the classification without clustering. As future work, we are planning to expand this work by using other clustering algorithms. In order to reduce the complexity, further optimizations should be carried out.

References:

[1] Y. G. Jung, M. S. Kang, and J. Heo, "Clustering performance comparison using *K* -means and expectation maximization algorithms," *Biotechnol. Biotechnol. Equip.*, vol. 28, no. sup1, pp. S44–S48, 2014.

[2] M. Singh, K. Kaur, and B. Singh, "Cluster Algorithm for Genetic Diversity," *World Acad. Sci. Eng. Technol. 18 2008*, pp. 453–457, 2008.

[3] S. Sharma, S. Kaur, and M. J. Kaur, "Hybrid Clustering and Classification," *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 5, no. 1, pp. 222–225, 2015.

[4] H. Qian, Y. Mao, W. Xiang, and Z. Wang, "Recognition of human activities using SVM multi-class classifier," *Pattern Recognit. Lett.*, vol. 31, pp. 100–111, 2010.

[5] T. Chakraborty, "EC3: Combining Clustering and Classification for Ensemble Learning," *J. Mach. Learn.*, vol. 13, no. 9, pp. 1–14, 2017.

[6] J. Fürnkranz, "Pairwise Classification as an Ensemble Technique," *Mach. Learn. ECML 2002*, vol. 2430, no. 2000, pp. 9–38, 2002.

[7] J. R. Kwapisz, G. M. Weiss, and S. a. Moore, "Activity recognition using cell phone accelerometers," *ACM SIGKDD Explor. Newsl.*, vol. 12, p. 74, 2011.

[8] M. Shoaib, H. Scholten, and P. J. M. Havinga, "Towards Physical Activity Recognition Using Smartphone Sensors," *2013 IEEE 10th Int. Conf. Ubiquitous Intell. Comput. 2013 IEEE 10th Int. Conf. Auton. Trust. Comput.*, pp. 80–87, 2013.

[9] I. P. Machado, A. Luisa Gomes, H. Gamboa, V. Paixao, and R. M. Costa, "Human activity data discovery from triaxial accelerometer sensor: Non-supervised learning sensitivity to feature extraction parametrization," *Inf. Process. Manag.*, vol. 51, no. 2, pp. 201–214, 2015.

[10] T. S. Madhulatha, "an Overview on Clustering Methods," *IOSR J. Eng.*, vol. 2, no. 4, pp. 719–725, 2012.

[11] H. K. Al-Mohair, J. Mohamad Saleh, and S. A. Suandi, "Hybrid Human Skin Detection Using Neural Network and K-Means Clustering Technique," *Appl. Soft Comput.*, vol. 33, pp. 337–347, 2015.

[12] F. Attal, S. Mohammed, M. Dedabrishvili, F. Chamroukhi, L. Oukhellou, and Y. Amirat, "Physical Human Activity Recognition Using Wearable Sensors," *Sensors*, vol. 15, no. 12, pp. 31314–31338, 2015.

[13] N. Dhanachandra, K. Manglem, and Y. J. Chanu, "Image Segmentation Using K-means Clustering Algorithm and Subtractive Clustering Algorithm," in *Procedia Computer Science*, 2015, vol. 54, pp. 764–771.

[14] D. R. Faria, C. Premebida, and U. Nunes, "A Probabilistic Approach for Human Everyday Activities Recognition using Body Motion from RGB-D Images," in *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, 2014, pp. 732–737.

[15] M. M. Jenghara and H. Ebrahimpour-komleh, "Rule Based Ensembles Using Pair Wise Neural Network Classifiers," *I.J. Intell. Syst. Appl.*, vol. 4, no. March, pp. 34–40, 2015.

[16] J. Bhatt, "A Survey on One Class Classification using Ensembles Method," *Int. J. Innov. Res. Sci. Technol.*, vol. 1, no. 7, pp. 19–23, 2014.

[17] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[18] M. N. Adnan and M. Z. Islam, "One-Vs-All Binarization Technique in the Context of Random Forest," in *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning.*, 2015, no. April, pp. 22–24.