Using Jaccard Distance Measure for Unsupervised Activity Recognition with Smartphone Accelerometers

Xianlong Wang[1], Yonggang Lu[1(✉)], Dachuan Wang[1], Li Liu[2(✉)], and Huiyu Zhou[3]

[1] School of Information Science and Engineering, Lanzhou University,
Lanzhou 730000, Gansu, China
`1781368972@qq.com, ylu@lzu.edu.cn, 910379526@qq.com`
[2] School of Software Engineering, Chongqing University, Chongqing 400044, China
`dcsliuli@cqu.edu.cn`
[3] Institute of Electronics, Communications and Information Technology,
Queens University of Belfast, Belfast BT3 9DT, UK
`h.zhou@ecit.qub.ac.uk`

Abstract. The rapid popularity of smartphones has led to a growing research interest in human activity recognition (HAR) with the mobile devices. Accelerometer is the most commonly used sensor of smartphone for HAR. Many supervised HAR methods have been developed. However, it is very difficult to collect the annotated or labeled training data for HAR. So, developing of effective unsupervised methods for HAR is very necessary. The accuracy of an unsupervised method, such as clustering, can be greatly affected by the similarity or distance measures, because the learning process of clustering method is completely depending on the similarity between objects. Although Euclidean distance measure is commonly used in unsupervised activity recognition, it is not suitable for measuring distance when the number of features is very large, which is usually the case in HAR. Jaccard distance is a distance measure based on mutual information theory and can better represent the differences between nonnegative feature vectors than Euclidean distance. It can also work well with a large number of features. In this work, the Jaccard distance measure is applied to HAR for the first time. In the experiments, the results of the Jaccard distance measure and the Euclidean distance measure are compared, using three different feature extraction methods which include time-domain, frequency-domain and mixed-domain feature extractions. To comprehensively analyze the experimental results, two different evaluation methods are used: (a) C-Index before clustering, (b) FM-index after using five different clustering methods which are Spectral Cluster, Single-Linkage, Ward-Linkage, Average-Linkage, and K-Medoids. Experiments show that, almost for every combination of the feature extraction methods and the evaluation methods, the Jaccard distance measure is consistently better than the Euclidean distance measure for unsupervised HAR.

Keywords: Activity recognition · Jaccard distance · Unsupervised HAR · Smartphone · Feature extraction

1   Introduction

With the improvement of health awareness of people and the continuous development of science technology, HAR is increasingly used in intelligent life, motion detection, and healthcare system [1–6]. The healthcare system can keep track of the health of user, and capture the activity of user for a long-term, which can help doctor to diagnose whether the user suffers from chronic diseases or not, and to urge the user to take proper exercise every day. The specialized medical equipment, wearable sensors or camera-based computer vision system have been used to recognize human activities, but they all require complex equipment and require camera to be placed in a fixed position, which is inconvenient for daily activity recognition.

Alternatively, smartphone is very popular now and people can carry it anytime, anywhere. Most smartphones are equipped with a rich set of embedded sensors. So smartphone has become an active field of research in the domain of perception and mobile computing. Within various sensors of smartphone, accelerometer is the most commonly used sensor for recording human activity signals. Khan et al. [7] have used a smartphone with a built-in triaxial accelerometer to collect five daily physical activities from five body positions, and the average classification accuracy of their method is about 96%. In the work of He and Li [8], three different sensors including accelerometer, gyroscope, and magnetic sensor embedded in a smartphone placed at the chest of a subject are used for HAR. The classification accuracy of their method is 95.03%. Their work also shows that within the three embedded sensors, the accelerometer is the most significant sensor for activity recognition.

Usually, raw time series data cannot be applied on HAR directly, feature extraction methods have to been used to produce a new data representation (called features) from the raw acceleration data before the classification. Popular features extracted from the acceleration data in the previous work can be divided into two types: time-domain features and frequency-domain features. Time-domain features include mean, variance, standard deviation, etc. [9], while frequency-domain features include frequency-domain entropy, discrete FFT coefficients, etc. [10]. In addition, mixed-domain features which include features from both time-domain and frequency-domain are also used in HAR [11].

In the traditional approaches to HAR, standard supervised classification methods, such as Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Decision Tree, etc., have been used after feature extraction. Sun et al. [12] have used accelerometer embedded mobile phones to monitor seven common physical activities of seven subjects and have boosted the overall accuracy from 91.5% to 93.1% by improving the SVM method. Anjum and IIyas [13] have collected a dataset of seven different activities, which include walking, running, ascending stairs, descending stairs, cycling, driving and remaining inactive using cell phone sensors and have evaluated a number of classification methods including Naive Bayes, Decision Tree, KNN and SVM. The Decision Tree classifier outperforms the other classifiers, and on average it produces a true positive rate of 95.2%.

All the above studies are using supervised classification methods to perform activity recognition. For the supervised methods, a large number of annotated data are required

to train the classifiers. However, the data annotation is a difficult task which usually requires lots of time and efforts. So the unsupervised methods which can use the raw data directly for HAR have special advantages. In our previous work [14], an unsupervised method is proposed for HAR with time-domain features extracted from the data collected using smartphone accelerometers. It is shown that the unsupervised method is also effective to distinguish several daily activities. The unsupervised methods, such as clustering methods, usually classify data based on the similarities between the data points. So the accuracy of the unsupervised classification methods can be greatly affected by the distance measure (or similarity measure) applied on the feature vectors. Although the Euclidean distance measure is the most commonly used method, it may fail to measure the differences between features effectively in the high-dimensional space due to the curse of dimensionality. In many cases, the normalization of features is also needed for using the Euclidean distance measure. The drawbacks of the Euclidean distance measure can be avoided by using the Jaccard distance measure instead. Jaccard distance measures the degree of the overlap between two sets of nonnegative feature values, so the normalization of features is not needed. Jaccard distance not only considers the differences between the feature vectors, but also considers the absolute values of the feature vectors, so it can better represent the differences between nonnegative feature vectors than Euclidean distance [15]. Furthermore, because Jaccard distance is based on mutual information theory, it can be applied to measure the similarities between objects in high-dimensional space. In this work, the Jaccard distance measure is applied to HAR for the first time as far as we know, and the superiority of the Jaccard distance to the Euclidean distance is shown by the experiments.

2   Related Work

Recently, position and orientation independent activity recognition using smartphone accelerometers has also been investigated. Fan et al. [16] use resultant acceleration to eliminate the effects of the phone orientations. A feature set including seven time-domain features and three frequency-domain features is used for classifying five activities: staying still, walking, jogging, ascending stairs and descending stairs. The classification accuracy produced by their method is 80.29%. Miao et al. [9] have used accelerometer, gyroscope, proximity sensor, light sensor and magnetic sensor of a smartphone for HAR. The magnitude of linear acceleration combined with signals collected from gyroscope sensor and magnetic sensor are used. Then six time-domain statistical features are extracted for recognizing five typical physical activities which include staying still, walking, jogging, ascending stairs and descending stairs. Their results show the possibility of position and orientation independent HAR using smartphones without firm attachment to the body. In this work, orientation-independent HAR is realized by using the resultant acceleration data from the smartphones too.

Many clustering methods have been applied on unsupervised HAR. In the work of Machado et al. [17], the activities of standing, sitting, walking, running, lying are studied with mixed-domain features. Four clustering methods, including K-Means, Spectral Clustering, Mean Shift and Affinity Propagation (AP) based on Euclidean distance are

used to distinguish different activities. In the subject-independent context, the accuracy rate of K-Means clustering reaches 88.75%, while the accuracy rate of other clustering algorithms ranges from 30% to 45%. In the case of subject-dependent context, the accuracy rate of K-Means clustering reaches 99.29% and the accuracy of other clustering methods ranges from 53% to 93.96%. In work of Gomes [18], K-Means, Mini Batch K-Means, AP, Mean Shift, DBSCAN, Spectral Clustering, and Ward-Linkage have been applied for HAR. They have found that the Spectral Clustering has produced the best results with an accuracy rate of $89.1 \pm 8.8\%$. Amitha and Rajakumari [19] have proposed a system using Naive Bayes classifier combined with a hierarchical agglomerative clustering algorithm for activity tracking. In this work, Spectral Clustering, K-Medoids, and three commonly used hierarchical clustering methods, including Single-Linkage, Ward-Linkage, and Average-Linkage are used for HAR.

3 Method

3.1 Data Collection

In the experiments, both an UCI dataset and a dataset collected by ourselves are used for HAR. The UCI dataset includes triaxial acceleration data of walking, running, ascending stairs, descending stairs, sitting and standing of 30 subjects [20]. However, the data of 13 subjects have missing values or invalid values. So only the data of the rest 17 subjects are used in our experiments.

The dataset collected by ourselves are from 6 healthy volunteers with ages from 22 to 28. These activities are walking, jogging, ascending stairs, descending stairs, sitting and standing. The acceleration data with a sampling frequency of 50 Hz are collected.

3.2 Feature Extraction

Before feature extraction, to reduce the bias caused by sensor sensitivity and noise, a sliding window with 50% overlap is employed to divide the time series into smaller time windows. For the UCI dataset the window of 2.56 s is used, while for our dataset, the window of 5.12 s is used. Similar as previous works [21], the magnitude of the acceleration, A_{3a}, is used for feature extraction, which is insensitive to the orientation and position of the devices. A_{3a} can be represented as:

$$A_{3a} = \sqrt{A_x^2 + A_y^2 + A_z^2}$$

The magnitudes of the acceleration data within each sliding window are then used for feature extraction. Three feature extraction methods based on frequency-domain, time-domain and mixed-domain are used in the experiments, which are explained in details below:

(a) **Time-domain Features:** The same set of time-domain features used in the work of Miao et al. [9] are extracted, which includes Mean, Standard Deviation, Median, Skewness, Kurtosis, and Inter-Quartile-Range.

(b) **Mixed-domain Features:** The same mixed-domain feature set as reported in the work of Figueira et al. [11] are extracted, which includes Root Mean Square (RMS), Median Absolute Deviation (MAD), Standard Deviation, Spectral Roll On, Mean Power Spectrum, Max power Spectrum.

(c) **Frequency-domain Features:** In the frequency-domain, to eliminate the effect of edge samples in each window, a sine window is first multiplied to the data in the window. The sine window is defined as:

$$win = \sin((i-1)\pi/(winLen-1))^2, \quad i \in [1:winLen]$$

where *winLen* represents the length of the window. Then FFT is applied to the data of each window. In order to improve the frequency resolution and reduce spectral leakage, the magnitudes of the Fourier coefficients are convolved with a hamming window of size 5. Then the final results of each window are used as feature vectors.

3.3   Jaccard Distance Measure

In this work, Jaccard distance [15] is used to calculate the similarity between feature vectors. If $X = (x_1, x_2, \ldots, x_n)$ and $Y = (y_1, y_2, \ldots, y_n)$ are two vectors and all x_i, y_i ($1 \le i \le n$) are non-negative, their Jaccard similarity coefficient is defined as:

$$J(X, Y) = \frac{\sum_{i=1}^{n} \min(x_i, y_i)}{\sum_{i=1}^{n} \max(x_i, y_i)},$$

and their Jaccard distance is defined as:

$$D_J(X, Y) = 1 - J(X, Y).$$

The value of the Jaccard distance ranges from 0 to 1, where 1 represents the identical sets, while 0 represents the disjoint sets. Because in our work, the feature values extracted using the three feature extraction methods are all greater than or equal to 0, Jaccard distance can be readily applied.

4   Experiments

4.1   Design of the Experiments

To compare the effectiveness of the Euclidean distance and the Jaccard distance in unsupervised HAR, the following procedures of the experiments are executed (as shown in Fig. 1): (a) Collect the acceleration data; (b) Divide the time series data into small time windows using a sliding window with 50% overlap; (c) Use the three feature extraction methods; (d) Calculate the distance; (e) Use C-index to compare Jaccard

distance and Euclidean distance; (f) Cluster the data based on the distances between feature vectors and then measure the clustering results using FM-index.
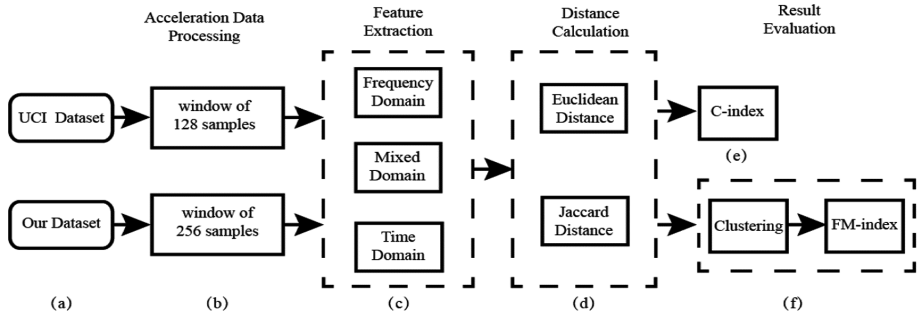
Fig. 1. Procedure of the experiments.

4.2 Evaluation Criterion

Both C-index [22] and FM-index [23] are used to evaluate the experimental results. C-index is used to measure the compatibility of the distance measure with the actual class labels. A smaller C-index value indicates a better compatibility where the data points have smaller distances within the same class and greater distances between different classes. The C-index is defined as:

$$C\_index = \frac{S - S_{\min}}{S_{\max} - S_{\min}}$$

where S is the sum of the distances over all m pairs of objects from the same class, S_{min} is the sum of the m smallest distances if all pairs of objects are considered. S_{max} is the sum of the m largest distances out of all the pairs. The interval of the C-index values is between 0 and 1. A smaller C-index value indicates a better compatibility of the distance measure with the actual class labels.

The FM-index is used to evaluate the clustering results produced by a clustering method. The maximum value of the FM-index is 1 that means that the clustering result are the same as the actual result, while the minimum value is 0 that means the clustering result and actual result are completely different.

4.3 Experimental Results

For the UCI dataset containing 17 subjects and our dataset containing 6 subjects, the C-index is used to measure the distances generated by three different feature extraction methods and the two different distance measures. The results are shown in Table 1. From Table 1, it can be seen that, for 60 out of all 69 cases, the C-index produced using the Jaccard distance measure are better than that produced using the Euclidean distance measure, regardless of which feature extraction method is used. The results show that

Jaccard distance can measure the distances between feature vectors better than Euclidean distance.

Table 1. Comparison of different feature extraction methods and distance measures on UCI dataset and our dataset using C-index.

Dataset	Frequency domain		Time domain		Mixed domain	
	Euclidean	Jaccard	Euclidean	Jaccard	Euclidean	Jaccard
UCI dataset	0.0645	**0.0298**	0.1723	**0.0621**	0.1127	**0.1087**
	0.0445	**0.031**	0.1105	**0.0375**	0.1058	**0.0889**
	0.0439	**0.0402**	0.1596	**0.0605**	0.0932	**0.0865**
	0.031	0.0892	0.2715	**0.1181**	0.1093	**0.1053**
	0.0617	**0.0554**	0.147	**0.065**	0.0931	**0.0941**
	0.2034	**0.0285**	0.1488	**0.0691**	0.1383	**0.0987**
	0.0555	**0.0404**	0.2497	**0.0712**	0.1632	**0.1416**
	0.0606	**0.0281**	0.1274	**0.0438**	0.0608	**0.0568**
	0.0531	**0.0373**	0.2003	**0.0694**	0.0973	**0.0849**
	0.0406	**0.031**	0.1504	**0.0412**	0.1072	**0.0809**
	0.0577	0.0644	0.0789	**0.0408**	**0.0409**	0.0432
	0.0885	**0.0116**	0.0927	**0.0594**	0.0583	**0.0421**
	0.1229	0.1316	0.2175	**0.0849**	0.181	**0.1707**
	0.0608	**0.0311**	0.077	**0.0256**	0.0256	**0.0245**
	0.0586	**0.0254**	0.185	**0.0674**	0.1688	**0.1379**
	0.0745	**0.0299**	0.1779	**0.0857**	0.1356	**0.1279**
	0.0395	**0.0287**	0.1367	**0.0444**	**0.0581**	0.0624
Our dataset	0.0294	**0.0282**	0.1734	**0.0049**	0.0525	**0.0335**
	0.0314	**0.0303**	0.0803	**0.012**	0.0378	**0.0272**
	0.0296	**0.0239**	0.0401	**0.0047**	0.0219	**0.0169**
	0.027	**0.0253**	0.129	**0.0097**	0.0702	**0.065**
	0.0201	0.021	0.1652	**0.0155**	0.0606	**0.0421**
	0.0286	**0.0273**	0.4472	**0.0082**	0.0227	**0.0197**

Based on the distances computed using the two distance measures, five different distance-matrix based clustering methods which include Spectral Cluster, Single-Linkage, Ward-Linkage, Average-Linkage, and K-Medoids are used to cluster the dataset. The FM-index is then used to measure the clustering results. Table 2 shows the FM-indices produced using different combinations of the three feature extraction methods, the five clustering methods and the two distance measures on the UCI dataset and our dataset, where the FM-indices of all the subjects in each dataset are averaged.

It can be seen from Table 2, regardless of which combination of the feature extraction methods and clustering methods is used, the FM-indices produced using the Jaccard distance measure are consistently better than the results produced using the Euclidean distance measure. The results also show the superiority of the Jaccard distance measure over the Euclidean distance measure for distinguishing different activities. It can also be seen in Table 2 that the FM-indices produced using the frequency-domain feature

Table 2. Comparison of different feature extraction methods, distance measures and clustering methods using FM-index.

Dataset	Clustering method	Frequency domain		Time domain		Mixed domain	
		Euclidean	Jaccard	Euclidean	Jaccard	Euclidean	Jaccard
UCI dataset	Spectral clustering	0.543	**0.583**	0.344	**0.378**	0.387	**0.433**
	Single Linkage	0.807	**0.851**	0.723	**0.771**	0.612	**0.68**
	Ward Linkage	0.77	**0.81**	0.763	**0.791**	0.635	**0.664**
	Average Linkage	0.79	**0.871**	0.757	**0.814**	0.618	**0.687**
	K-Medoids	0.653	**0.654**	0.633	**0.645**	0.589	**0.604**
Our dataset	Spectral clustering	0.674	**0.688**	0.543	**0.582**	0.387	**0.442**
	Single Linkage	0.845	**0.895**	0.878	**0.904**	0.62	**0.658**
	Ward Linkage	0.895	**0.91**	0.887	**0.909**	0.652	**0.701**
	Average Linkage	0.869	**0.886**	0.836	**0.87**	0.636	**0.695**
	K-Medoids	0.749	**0.729**	0.739	**0.726**	0.62	**0.624**

extraction method are all better than these produced using the other two feature extraction methods, except the result of using Single-Linkage on our dataset. This indicates that the frequency-domain features are more effective than the other two features for unsupervised HAR.

To further compare the Jaccard distance measure and the Euclidean distance measure for activity recognition, the FM-indices produced using the five clustering methods are averaged for each combination of the distance measure and the feature extraction
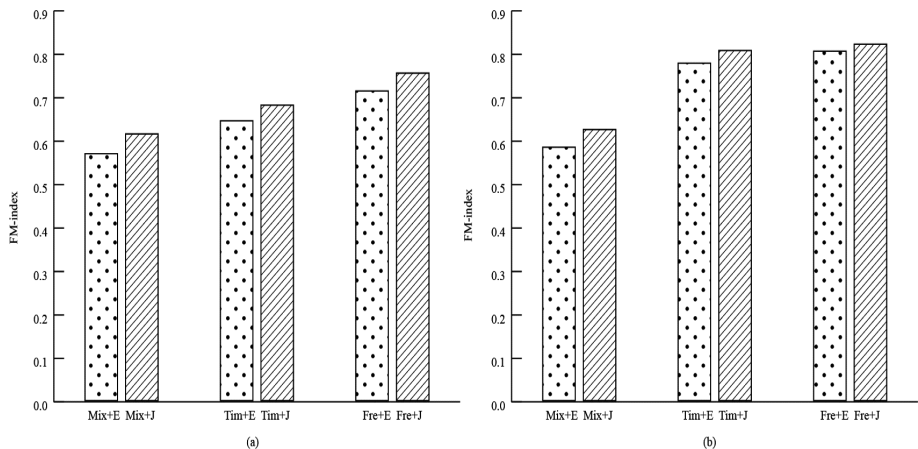
Fig. 2. Comparison of different combinations of the feature extraction methods and the distance measures using the average of FM-index of the five clustering methods on (a) UCI dataset, and (b) our dataset, where "Fre", "Tim", "Mix" represent frequency-domain, time-domain, mixed-domain feature extraction methods respectively, and "E", "J" represent Euclidean distance and Jaccard distance respectively.

methods. The results are shown in Fig. 2. It can be seen that the FM-index produced using Jaccard distance is better than that produced using Euclidean distance regardless of which feature extraction method is used, while the FM-index produced using the frequency-domain feature extraction method is better than that produced using the other two feature extraction methods.

5   Conclusion

In this paper, the Jaccard distance measure is proposed to replace the Euclidean distance measure for unsupervised HAR. Both the C-index before clustering and the FM-index after clustering show the superiority of the Jaccard distance measure over the Euclidean distance measure for unsupervised HAR. It is also found that the frequency-domain feature extraction method is better than the time-domain feature extraction method and the mixed-domain feature extraction method for unsupervised HAR using the proposed method. Future work includes applying the Jaccard distance measure on more datasets and improving the feature extraction methods.

References

1. Nie, L., Akbari, M., Li, T., Chua, TS.: A joint local-global approach for medical terminology assignment. In: Medical Information Retrieval Workshop at the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, Gold Coast, Australia, pp. 24–27 (2014)
2. Nie, L., Wang, M., Zhang, L., Yan, S., Bo, Z., Chua, T.S.: Wenzher: comprehensive vertical search for healthcare domain. In: The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, Gold Coast, Australia, pp. 1245–1246 (2014)
3. Nie, L., Wang, M., Zhang, L., Yan, S., Bo, Z., Chua, T.S.: Disease inference from health-related questions via sparse deep learning. IEEE Trans. Knowl. Data Eng. **27**(8), 2107–2119 (2015)
4. Liu, L., Peng, Y., Liu, M., Huang, Z.: Sensor-based human activity recognition system with a multilayered model using time series shapelets. Knowl. Based Syst. **90**, 138–152 (2015)
5. Nie, L., Zhang, L., Yang, Y., Wang, M., Hong, R., Chua, T.S.: Beyond doctors: future health prediction from multimedia and multimodal observations. In: The 23rd ACM International Conference on Multimedia, Brisbane, Australia, pp. 591–600 (2015)
6. Pitta, F., Takaki, M.Y., Oliveira, N.H., Sant'Anna, T.J.P., Fontana, A.D., Kovelis, D., et al.: Relationship between pulmonary function and physical activity in daily life in patients with COPD. Respir. Med. **102**(8), 1203–1207 (2008)
7. Khan, A.M., Lee, Y.K., Lee, S.Y., Kim, T.S.: Human activity recognition via an accelerometer-enabled-smartphone using kernel discriminant analysis. In: The 5th International Conference on Future Information Technology (FutureTech), Busan, Korea (South), pp. 1–6 (2010)

8. He, Y., Li, Y.: Physical activity recognition utilizing the built-in kinematic sensors of a smartphone. Int. J. Distrib. Sens. Netw. **9**(4), 481580 (2013)
9. Miao, F., He, Y., Liu, J., Li, Y., Ayoola, I.: Identify typical physical activity on smartphone with varying positions and orientations. Biomed. Eng. Online **14**(1), 32 (2015)
10. Huynh, T., Schiele, B.: Analyzing features for activity recognition. In: Proceedings of the 2005 Joint Conference on Smart Objects and Ambient Intelligence: Innovative Context-Aware Services: Usages and Technologies, Grenoble, France, pp. 159–163 (2005)
11. Figueira, C.R., Matias, R., Hugo, G.: Body location independent activity monitoring. In: The 9th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC), Rome, Italy, pp. 190–197 (2016)
12. Sun, L., Zhang, D., Li, B., Guo, B., Li, S.: Activity recognition on an accelerometer embedded mobile phone with varying positions and orientations. In: Yu, Z., Liscano, R., Chen, G., Zhang, D., Zhou, X. (eds.) UIC 2010. LNCS, vol. 6406, pp. 548–562. Springer, Heidelberg (2010). doi:10.1007/978-3-642-16355-5_42
13. Anjum, A., IIyas, M.U.: Activity recognition using smartphone sensors. In: 10th Consumer Communications and Networking Conference (CCNC), Las Vegas, USA, pp. 914–919 (2013)
14. Lu, Y., Wei, Y., Liu, L., Zhong, J., Sun, L., Liu, Y.: Towards unsupervised physical activity recognition using smartphone accelerometers. Multimedia Tools Appl. **76**, 1–19 (2016)
15. Levandowsky, M., Winter, D.: Distance between sets. Nature **234**(5), 34–35 (1971)
16. Fan, L., Wang, J., Wang, H.: Activity recognition based on position-independent smartphone. J. Comput. Inf. Syst. **10**(11), 4921–4928 (2014)
17. Machado, I.P., Gomes, A.L., Gamboa, H., Paixão, V., Costa, R.M.: Human activity data discovery from triaxial accelerometer sensor: non-supervised learning sensitivity to feature extraction parameterization. Inf. Process. Manag. **51**(2), 204–214 (2015)
18. Gomes, A.L.G.N.: Human activity recognition with accelerometry: novel time and frequency. Doctoral dissertation, Champalimaud Neuroscience Programme (2014)
19. Amitha, R., Rajakumari, K.: Hierarchical clustering based activity tracking system in a smart environment. Int. J. Comput. Sci. Tech. **3**, 2229–4333 (2012)
20. UCI dataset. https://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones
21. Wang, D., Liu, L., Wang, X., Lu, Y.: A novel feature extraction method on activity recognition using smartphone. In: Song, S., Tong, Y. (eds.) WAIM 2016. LNCS, vol. 9998, pp. 67–76. Springer, Cham (2016). doi:10.1007/978-3-319-47121-1_6
22. Hubert, L., Schultz, J.: Quadratic assignment as a general data-analysis strategy. Br. J. Math. Stat. Psychol. **29**(2), 190–241 (1976)
23. Fowlkes, E.B., Mallows, C.: A method for comparing two hierarchical clusterings. J. Am. Stat. Assoc. **78**(383), 553–569 (1983)