A Noble Feature Selection Method for Human Activity Recognition using Linearly Dependent Concept (LDC)

Win Win Myo
Artificial Intelligence Research Lab
Department of Computer Science
Faculty of Science
Prince of Songkla University
Hat Yai, Thailand
+66 831685868
winwinmyo76@gmail.com

Wiphada Wettayaprasit
Artificial Intelligence Research Lab
Department of Computer Science
Faculty of Science
Prince of Songkla University
Hat Yai, Thailand
+66 74289409
wiphada.w@psu.ac.th

Pattara Aiyarak
Artificial Intelligence Research Lab
Department of Computer Science
Faculty of Science
Prince of Songkla University
Hat Yai, Thailand
+66 74282815
pattara.a @psu.ac.th

ABSTRACT

Human physical activity recognition process using mobile phones is very complicated with many extracted features in which some features are irrelevant or redundant. Removing irrelevant or redundant features is not only reducing the dataset size but also saving the time consuming task. Hence, a reason to pick out the effective and useful features is our main study. We propose a noble feature selection technique using Linearly Dependent Concept (LDC). Our proposed work attempts a new feature selection method on UCI-HAR dataset. For classification, we use the feed forward neural network and compare the performance result with the original dataset. The goal of our study is not only to find an effective and useful features set from the original dataset but also to be better performance than original dataset. Finally, the experimental result of proposed method gives 2.7% more accuracy and reduces the relative error up to 2.67% of the original dataset.

CCS Concepts

• Computing methodologies → Feature selection

Keywords

Neural Network, Linearly Dependent, Feature Selection, Mobile Phone, Sensor, Human Activity Recognition.

1. INTRODUCTION

Human Activity Recognition (HAR) using mobile phone is a high-dimensional machine learning problem. It can be used to monitor the human daily routine and care for the individual human health. Moreover, it can provide many other applications such as health care applications, security applications, human survey system, fitness center, and so on. As today technology of mobile phone is very powerful performance day by day with the built-in sensors, the HAR using mobile phone with built-in sensors is a rapidly growing field in current research area. It is a reason of why we do this work and our study expects to support for better recognition performance of human activity recognition. In our study, we use a publicly available dataset: UCI-HAR dataset for recognition of human activity introduced by D.Anguita et al. [1]. The HAR process is the most important application among of the numerous physiological applications addressed by J.-L. Reyes-Ortiz et al. [2].

Today smartphones come in a variety of formats with built in sensors such as 3-dimentional accelerometer and gyroscope which are the most popular sensors on mobile phones to predict human activity recognition surveyed by Ó. D. Lara et al. [3]. These sensors can retrieve the signal data to get the useful information by extracting the available data. However, the human physical activity recognition technique is very complicated and complex to get accurate information. Although the researchers gave much more attentions to attitude their skills that predicted on human activities with their experiences, the complex activity recognition is still complex and challenging.

Due to the data of human activity recognition process is very complex with a variety of extracted features, removing the irrelevant or redundant features are more practical and necessary. The work attempted in this paper is focused on feature selection to pick out useful features from the UCI-HAR dataset. The goal of this study is to create more accurate recognition system by removing the redundant features from the original dataset. Although many researchers attempted to improve quantifying minute-by-minute what physical activities, some weak points still had. Being an active area in healthcare, the HAR using mobile phones with built in sensors is the increasing research area interested by many researchers.

As knowing the important of feature selection, L.Wang et al. proposed a powerful feature reduction method by GDA and showed that their GDA could sharply reduce the dimension of the feature space for large-scale datasets [4]. Then N. Díaz-Rodríguez et al. proposed a new tensor-based feature selection method : Tensor Manifold Discriminant Projections (TMDP) and demonstrated the evidence of their proposed method [5]. As the feature selection is also an important technique in the healthcare system, Z. Zhang et al. introduced a novel disease-specific feature selection method using the Dynamic Time Warping (DTW) for automatic heartbeat classification [6].

Before classification task, feature selection or reduction is an effective data reduction process. To perform feature and instance selection, J. Derrac et al. investigated an evolutionary model using a Co-operative Co-evolutionary algorithm for Instance and Feature Selection (IFS-CoCo) and approved this method in many computational problems [7]. In human activity recognition, the data dimensionality reduction is also an essential process because the

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
ICSCA 2018, February 8–10, 2018, Kuantan, Malaysia
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5414-1/18/02…$15.00
https://doi.org/10.1145/3185089.3185131

data captured by sensors in activity recognition is a large amount of data with redundant or incomplete information. Hence, Simao et al. introduced the Data Dimensionality Reduction (DDR) technique to reduce the incomplete data using re-sampling raw data and Principal Component Analysis (PCA) [8]. A new Expectation Maximization (EM) algorithm using simultaneous feature selection and Gaussian mixture model estimation for supervised classification problems were proposed by J. Kersten et al. [9]. And they showed that this approach was able to provide for better result. A brief description of the related review papers using feature selection or reduction techniques for the human activity recognition using mobile phones with built-in sensors is shown in Table 1.

Table 1. Description of related review papers on feature selection method

No	Cite No.	Year	Methods	Classifier	Accuracy %
1	[7]	2010	IFS-CoCo (a cooperative co-evolutionary algorithm for instance and feature selection).	Nearest Neighbor	87.69
2	[2]	2012	GDA (Generalized Discriminant Analysis) based on kernel function technique	RVM (Relevance Vector Machines)	99.20
3	[6]	2014	Disease-Specific Feature Selection Method	SVM	86.66
4	[9]	2014	EM (Expectation Maximization) model	SVM	88.50
5	[8]	2017	DDR (Data Dimensionality Reduction technique)	ANN, SVM	82.00

Despite many researchers dedicated how to recognize the human activity, it still remained the invisible things for all predicted works. To support the researchers' work, D.Anguita et al. published their data set as a publicly available dataset and acknowledged their result by exploiting a multiclass Support Vector Machine (SVM) [1]. Then M. Brown et al. used this dataset by analyzing with several methods for classification activities and created their raw data supporting to make a faster and more efficient classifier [10]. In fact, the extracted feature data are useful and important to get an accurate information in human activity recognition. For this point of view, G. Chetty et al. showed the ranking diffident features with different classifiers and discussed the most features which had the best accuracy [11]. To develop the human activity recognition, R.San-Segundo et al. proposed the segmentation human activity system with six different physical activities showing a best activity segmentation error rate [12]. A brief description of related reviewed papers using available public UCI-HAR dataset for human activity recognition is shown in Table 2.

Table 2. Description of related review papers using UCI-HAR dataset

No.	Cite No.	Year	Title	Classifier	Accuracy %
1	[1]	2013	A Public Domain Dataset for Human Activity Recognition Using Smartphones	SVM	96.00
2	[10]	2013	Activity Classification with Smartphone Data	Naïve Bayes	80.00
				GDA	96.00
				GDA+HMM	98.00
3	[11]	2015	Smart phone based data mining for human activity recognition	RF	96.30
				IBK	97.89
4	[12]	2016	Segmenting human activities based on HMMs using smartphone inertial sensors	HMMs	98.00

The rest of the paper is structured by the following ways. The state of the art of technology will be illustrated in Part 2. The evaluation of experiment result is described in Part 3. The discussion is addressed in Part 4. Finally, we discussed the conclusion of this paper in Part 5.

2. THE ART OF THE TECHNOLOGY

2.1 Human Activity Recognition Process

Human activity recognition is a sensor-based physical activity determination with the novel data mining and machine learning techniques to determine the personal human behaviors. The process of human activity recognition is shown in Figure 1. There are three levels to determine human activity. At the lowest level, the retrieved sensor data is located to the generation of some sample data using the mathematical concepts and language. At an intermediate level, this includes the testing hypotheses and the deriving estimates with deducing properties by analysis of data. The highest level is concerned to figure out the overall goal for the activity sequences.

2.2 UCI-HAR Data Set

This dataset is from human activity recognition process data using smart phones [1]. The total number of instances is 10,299 instances and the total number of features is 561 features. Our work will use this dataset to prove our proposed method for the removing irrelevant or redundant features.

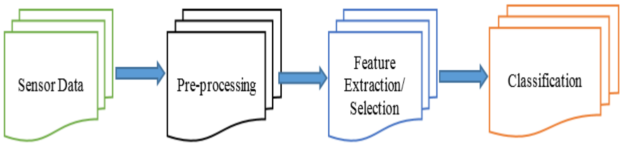

Figure 1. The process of Human Activity Recognition.

2.3 System Architecture

Basically, our study consists of two recognition processes: classification phase on original dataset and classification phase on new dataset selected by the proposed features selection method (using LDC). The schematic view of our system is shown in Figure 2. Then the study will compare the performances of classification results on these two datasets: original UCI-HAR dataset and new selected features dataset.

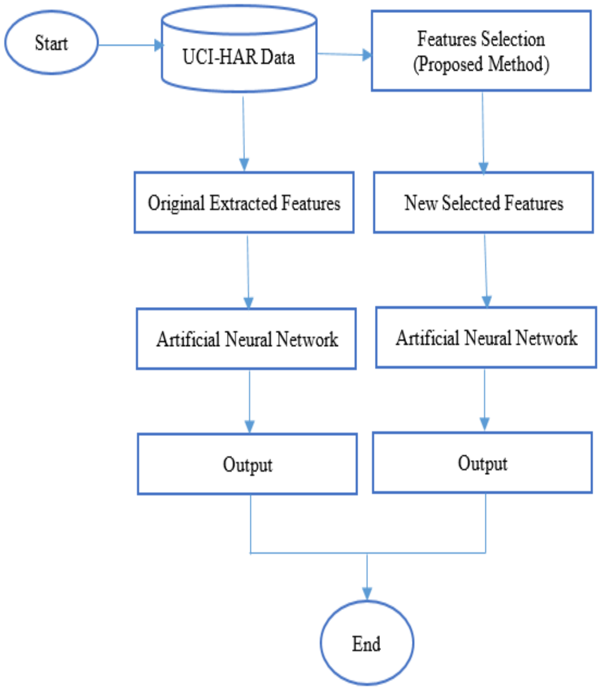
Start
UCI-HAR Data
Features Selection (Proposed Method)
Original Extracted Features
New Selected Features
Artificial Neural Network
Artificial Neural Network
Output
Output
End

Figure 2. System Architecture.

2.4 Proposed Feature Selection Method using linearly dependent concept (LDC)

This research is a feature selection based on the linearly dependent concept to remove the redundant features [13-15]. This method reduces data dimension size as well as helps to make human activity recognition process easier to recognize. The size of the data dimension and the useful features can be defined. Our study focuses on the UCI-HAR dataset to select the useful features. There are 561 features taken into consideration as vectors. We simply set $x_1, x_2, \ . \ . \ . \ , x_k \in V$ where V is a vector set of features. Let's consider for all vectors as a homogeneous linear combination system as below:

$$S = \sum_{i=1}^{k} c_i x_i \qquad (1)$$

where $c_i \in$ R and $x_i \in$ V.

And we have the following system,

$$c_1\begin{bmatrix}a_{11}\\a_{21}\\:\\.\\a_{p1}\end{bmatrix}+c_2\begin{bmatrix}a_{12}\\a_{22}\\:\\.\\a_{p2}\end{bmatrix}+.\ .\ .+c_k\begin{bmatrix}a_{1k}\\a_{2k}\\:\\.\\a_{pk}\end{bmatrix}=\begin{bmatrix}0\\0\\:\\.\\0\end{bmatrix}$$

Where p is total number of instances, k is total number of features, and a_{ij} are values of vectors, i= 1, 2, . . . p and j= 1, 2, . . ., k. In order to get the constant values, we use the Gaussian elimination method.

The architecture of our purposed feature selection technique is illustrated in the following Figure 3.

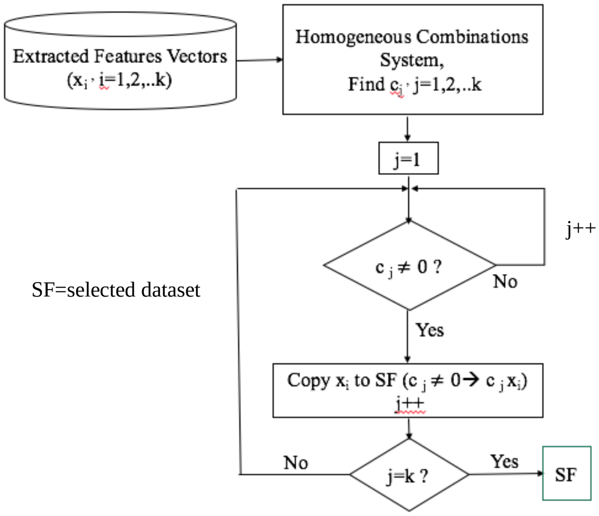
Extracted Features Vectors
(x_i , i=1,2,..k)
Homogeneous Combinations System,
Find c_j , j=1,2,..k
j=1
j++
SF=selected dataset
$c_j \neq 0$?
No
Yes
Copy x_i to SF ($c_j \neq 0 \rightarrow c_j x_i$)
j++
No
j=k ?
Yes
SF

Figure 3. Proposed LDC method.

By our proposed algorithm, we collect the useful features after checking the constant values. We apply this method to UCI-HAR dataset which is containing 561 features. After applying proposed method, 245 features are selected, and 316 features are discarded from original dataset.

2.5 Classifier

This system trains model A which is containing 561features without proposed method and model B which is containing 245 features with proposed method with Back-propagation feed forward neural network shown in Figure 4.

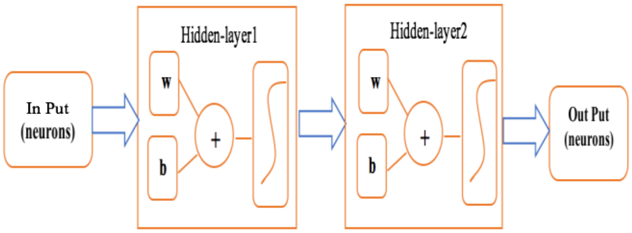

Figure 4. Back Propagation Neural network classifier with two hidden layers.

2.6 Dimension of Data Size

In model A, due to the hidden layer 1 is connecting with 561 input neurons, there are 5,777,739 (10299 x561) dimensional data size. Besides, the hidden layer 2 is also connecting with hidden later 1, there are 106,069,401 (10299 x 10299) dimensional data size.

Moreover, the output layer has 6 neurons and each neuron is fully connected to all neurons of the previous layer, there are 61,794 (10299x6) weights and connections. Therefore, there has a total of 111,908,934 dimensional data size model A.

In model B, being the hidden layer 1 with 245 features connections, there are 2,523,255(10299 x245) dimensional data size. On the other hand, the hidden layer 2 is also connecting with hidden later 1, there are 106,069,401 (10299x 10299) dimensional data size. While the output layer has 6 neurons and each neuron is fully connected to all neurons of the previous layer, there are 61,794 (10299x6) data size. Hence, there has a total of 108,654,450-dimensional data in model B.

Due to our new proposed model, we can discard extra 3,254,484 dimension of data size from model A which is described in Table 3.

Table 3. Dimension of data sizes on two models

Dataset	Input neurons	Hidden layer 1	Hidden layer 2	Output neurons	Total data size
Model A	561	10,299	10,299	6	111,908,934
Model B	245	10,299	10,299	6	108,654,450
Reduced dimensional data					3,254,484

3. EXPERIMENTAL RESULT

After applying the proposed method, we divide the dataset to evaluate the performance. We use 70% of data for training data, 15% of data for test data, and 15% of data for validation as shown in Table 4.

Table 4. Dividing the dataset to classify

Training Data	Validation Data	Testing Data
70%	15%	15%
6,625	1,837	1,837

The neural network classifier trains both models to recognize six activities (sitting, standing, laying, walking, walking upstairs, and walking downstairs). The achieved performance results on model A is shown in Table 5 and model B is described in Table 6.

Table 5. The performance result of model A

	Classification Result							
	Activity	Sitting	Standing	Laying	Walking	Walking upstairs	Walking downstairs	Accuracy %
Real Result	Sitting	302	0	0	0	0	0	100.00
	Standing	2	269	0	0	0	0	99.30
	Laying	0	1	233	0	0	0	99.60
	Walking	0	0	0	267	66	0	80.20
	Walking upstairs	0	0	0	1	334	0	99.70
	Walking downstairs	0	0	0	0	0	362	100.00
	Reliability	99.3	99.6	100	99.6	83.5	100	96.10

Table 6. The performance result on model B

	Classification Result							
		Sitting	Standing	Laying	Walking	Walking upstairs	Walking downstair	Accuracy %
Real Result	Sitting	302	0	0	0	0	0	100.00
	Standing	1	270	0	0	0	0	99.60
	Laying	0	2	232	0	0	0	99.15
	Walking	0	1	0	323	10	0	97.00
	Walking upstairs	0	0	0	7	328	0	97.80
	Walking downstairs	0	0	0	0	0	362	100.00
	Reliability	99.7	99.6	100	95.9	97.6	100	98.80

The compared performance result for model A and model B is described in Table 7 to compare their accuracies. The graph of accuracies on two models is shown in Figure 5.

Table 7. The compared performance result on model A vs. model B

Activities	Accuracy (%)	
	Model A	Model B
Sitting	100.00	100.00
Standing	99.30	99.60
Laying	99.60	99.15
Walking	80.20	97.00
Walking upstairs	99.70	97.80
Walking downstairs	100.00	100.00

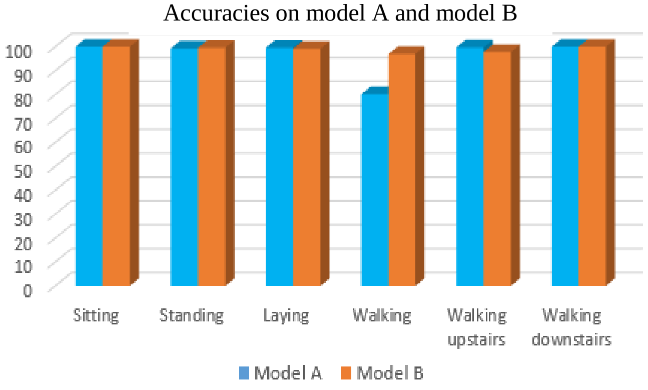

Figure 5. The graph of performance result on model A vs. model B

After the performance measures, we compute the relative error rates on two models with the following formulations. The results are shown in Table 8 and the graph of result error rates is illustrated in Figure 6.

True value = Approximation + True error

True error = True value – Approximation

$$Relative\ error = \frac{True\ error}{True\ value}$$

$$Percentive\ relative\ error = \frac{True\ eror}{True\ value}\ 100\%$$

Table 8. The error rates on model A vs. model B

Data	Sitting	Standing	Laying	Walking	Walking upstairs	Walking downstairs	Average error
Model A	0.00	0.74	0.43	19.82	0.30	0.00	3.81
Model B	0.00	0.37	0.86	3.29	2.09	0.00	1.14

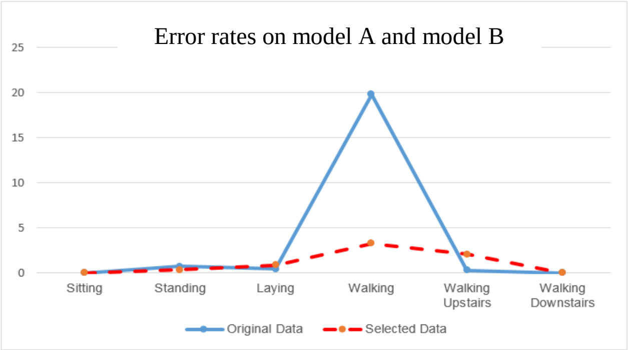

Figure 6. The error rates result on model A vs. model B

4. DISCUSSIONS

Our majority work is focused on finding a dataset with a minimum and useful features to be more effective on accuracy and to reduce the relative error rate. As mention above Tables and Figures, the accuracy result of this study is completely good. In both models, the performance results on sitting activity and walking downstairs activity are almost same. Moreover, there are nearly same accuracies at standing and laying activities. The result of standing activity of model B is 0.3% better than model A but the result of laying activity of model A is 0.45% better than model B. Furthermore, the walking upstairs result in model A is 1.9% better than model B. In that three cases, there has no the highest difference between them. However, the performance accuracy is the highest difference in walking activity between two models. The model B is showed 16.8% better accuracy than model A. After analyzing, the overall accuracy of model B is 2.7% better than model A. Besides, the study can reduce 2.67% error rate of model A.

5. CONCLUSIONS

A major focus of this study is to support for more effective determination of recognition on human activity. Although HAR using mobile phone sensors is very complicated process and data dimension is very high, the proposed model can reduce the data dimension size by leaving out the relevant and irrelevant features from original dataset. It helps the human activity recognition process to be easier recognition and save the time consuming for slow down the mining process. Finally, the experiment shows that the performance and error rate of our approach are better than the original dataset. Furthermore, it is also described the robustness of proposed method. As a result, we successfully apply this method to select effective features from human activity recognition data. As a future work, this technique will be used on every complicated and complex dataset of digital signals with very high dimensional data sizes.

6. REFERENCES

[1] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, “A Public Domain Dataset for Human Activity Recognition Using Smartphones,” Eur. Symp. Artif. Neural Networks, Comput. Intell. Mach. Learn., no. April, pp. 24–26, 2013.

[2] J.-L. Reyes-Ortiz, L. Oneto, A. Sama, X. Parra, and D. Anguita, “Transition-Aware Human Activity Recognition using smartphones.,” Neurocomputing An Int. J., vol. 171, pp. 754–767, 2016.

[3] Ó. D. Lara and M. A. Labrador, “A survey on human activity recognition using wearable sensors.,” IEEE Commun. Surv. Tutorials, vol. 15, no. 3, pp. 1192–1209, 2013.

[4] L. Wang, “Recognition of Human Activities with Wearable Sensors,” EURASIP J. Adv. Signal Process., vol. 2012, no. 108, pp. 1–13, 2012.

[5] N. Díaz-Rodríguez, O. L. Cadahía, M. P. Cuéllar, J. Lilius, and M. D. Calvo-Flores, “Handling real-world context awareness, uncertainty and vagueness in real-time human activity tracking and recognition with a fuzzy ontology-based hybrid method,” Sensors (Switzerland), vol. 14, no. 10, pp. 18131–18171, 2014.

[6] Z. Zhang, J. Dong, X. Luo, K. S. Choi, and X. Wu, “Heartbeat classification using disease-specific feature selection,” Comput. Biol. Med., vol. 46, no. 1, pp. 79–89, 2014.

[7] J. Derrac, S. Garcia^{b}, and F. Herrera, “IFS-CoCo: Instance and feature selection based on cooperative coevolution with nearest neighbor rule,” Pattern Recognit., vol. 43, no. 6, pp. 2082–2105, 2010.

[8] Simão, P. Neto, and O. Gibaru, “Using data dimensionality reduction for recognition of incomplete dynamic gestures,” Pattern Recognit. Lett., vol. 0, pp. 1–7, 2017.

[9] J. Kersten, “Simultaneous feature selection and Gaussian mixture model estimation for supervised classification problems,” Pattern Recognit., vol. 47, no. 8, pp. 2582–2595, 2014.

[10] M. Brown, T. Deitch, and L. O. Conor, “Activity Classification with Smartphone Data,” pp. 1–5, 2013.

[11] G. Chetty, M. White, and F. Akther, “Smart phone based data mining for human activity recognition,” Procedia Comput. Sci., vol. 46, no. Icict 2014, pp. 1181–1187, 2015.

[12] R. San-Segundo, “Segmenting human activities based on HMMs using smartphone inertial sensors,” Pervasive Mob. Comput., vol. 30, pp. 84–96, Aug. 2016.

[13] D. J. S. Robinson, “Linear algebra,” *World*, no. Book, Whole, 2011.

[14] T. Tao, “Lecture notes on linear algebra (Math 115A),” 2002.

[15] B. Textbook, “with Open Texts A First Course in an Open Text,” 2017.