Contemporary Engineering Sciences, Vol. 9, 2016, no. 21, 1031 - 1041
HIKARI Ltd, www.m-hikari.com
http://dx.doi.org/10.12988/ces.2016.67119

Mining Human Activity Using Dimensionality Reduction and Pattern Recognition

Ismail El Moudden

Laboratory of Mathematics, Computer Science & Applications
Faculty of Sciences, Mohammed V University in Rabat, Morocco

Mounir Ouzir

Laboratory of Biochemistry and Immunology, Department of Biology
Faculty of Sciences, Mohammed V University in Rabat, Morocco

Badreddine Benyacoub

Laboratory of Mathematics, Computer Science & Applications
Faculty of Sciences, Mohammed V University in Rabat, Morocco

Souad El Bernoussi

Laboratory of Mathematics, Computer Science & Applications
Faculty of Sciences, Mohammed V University in Rabat, Morocco

Abstract

Human activity recognition (HAR) is an emerging research topic in pattern recognition, especially in computer vision. The main objective of human activity recognition is to automatically detect and analyze human activities from the information acquired from different sensors. Human activity prediction using big data remains a challengingly open problem. Several approaches have recently been developed in order to find practical ways to solve high dimensionality of data problems. The aim of this study is to attempt, using data mining techniques, to deal with HAR modeling involving a significant number of variables in order to identify relevant parameters from data and thus to maximize the classification accuracy while minimizing the number of features. The proposed framework has been tested on a publicly HAR available dataset and the results have been interpreted and discussed.

Keywords: Human Activity Recognition, Dimensionality Reduction, Machine Learning

1 Introduction

Human activity recognition (HAR) is an important component in various scientific research contexts such as surveillance, healthcare and human computer interaction [1]. Using a set of sensors and intelligent detection algorithms, the main objective of HAR is to automatically detect and analyze different activities performed by humans [2]. In the last decade, HAR has gained increased attention due to the arrival of minimally invasive mobile sensing platforms such as smartphones, smartwatches and fitness bracelets. HAR using smartphones has become the most common approach to recognize physical activities, allowing easy collection of data which may be analyzed with machine learning methods for an eventually activity classification [3]. In fact, considerable previous works using different feature extraction techniques and classification algorithms showed that smartphones can address efficiently the problem of activity recognition [4, 5, 6, 7].

Successful identification of human activities may be useful to detect dangerous events or to monitor elderly physically or mentally disabled people, and children. Recently, a study performed on a smartphone based system has provided relevant information on human movement activities for both able-bodied and stroke populations, which may valuably contribute in clinical decision-making [8]. In the mentioned study, the classification algorithm performance using decision tree shows that an increase in activity classification complexity leads to a decrease in HAR performance with a stroke population [8]. One of the human activity recognition problems is high dimensionality. Many high dimensional classification techniques that allow considerable feature reduction through automatic selection of the most informative features have been developed for human activity recognition problems [4, 9, 10]. Dimensionality reduction can be applied to remove the irrelevant (or less relevant) features and reduce the computational complexity and increase the performance of the activity recognition process [9].

This study aims to create a framework for HAR high dimensionality in order to maximize the classification accuracy and to minimize the number of features using feature extraction and classification algorithms tested on a publicly available dataset.

2 Background

Human activity recognition is a challenging research topic in the field of computer vision and pattern recognition. Computer vision is an interdisciplinary field that deals with computer systems analysis and interpretation of visual capabilities of a close contents scene to those of human vision. One of the major objectives in computer vision is to recognize and understand human mobility, in order particularly to define the classification of human activities [2].

The pattern recognition process is a procedure that tells us the difference between objects, phenomena or events. Computational algorithms can be used to automate this process: they can be used to find dissimilarities, similarities and regularities automatically from the signals, generalize detection model based on these and, finally, give a recognition result [11]. The activity recognition problem is a classical pattern recognition problem with the aim to recognize automatically common human activities in real life settings [12]. The number of machine learning models that have been used for activity recognition and analysis varies almost as greatly as the types of recognized activities and sensor data used.

2.1 Dimension Reduction

Dimension reduction is the process of reducing the random number of variables under consideration. In recent years, the high dimensionality of modern massive datasets has provided a considerable challenge to efficient algorithmic solutions design. Dimensionality reduction techniques aims at finding the meaningful low dimensional data structures hidden in their high-dimensional observations which allow the user to better analyze the complex data sets. Feature extraction and feature selection are two popular methods for dimensionality reduction. Dimensionality reduction can be defined by assuming that we have dataset represented in a$n \times p$ matrix 'y' consisting of n data vectors $y_{i_{(i \in \{1,2,\ldots,n\})}}$with dimensionality '$p$'. Assume further that this dataset has intrinsic dimensionality 'k' (where$k < p$, and often $k \ll p$). In mathematical terms, intrinsic dimensionality means that the points in dataset 'y' are lying on or near a manifold with dimensionality 'k' that is embedded in the p-dimensional space. Dimensionality reduction techniques transform dataset 'y' with dimensionality 'p' into a new dataset 'x' with dimensionality 'k', while retaining data geometry as much as possible [13]. In general, neither data geometry manifold, nor the intrinsic dimensionality 'k' of the dataset 'y' are known. Therefore, dimensionality reduction is an attention-demanding problem that can only be solved by assuming certain data properties (such as its intrinsic dimensionality).

Dimension reduction techniques may be divided into two classes: linear methods [e.g. the Principal Component Analysis (PCA), linear discriminant analysis (LDA)] and nonlinear algorithms like Kernel PCA and Kernel LDA.

2.1.1 Principal Components Analysis (PCA)

Principal Components Analysis (PCA) is a linear technique that performs dimensionality reduction by embedding the data into a linear subspace of lower dimensionality [14]. The basic idea behind the PCA is to reduce the dimensionality of a data set, while retaining as much as possible the variation in the original variables. This is achieved by transforming the p original variable y=[y1,…,yp], to a new set of 'k' predictor variables $f = [f_1, f_2, \ldots, f_k]$, which are linear combinations of the original variables. More formally, PCA sequentially maximizes the variance of a linear combination of the original variables:

$$a_k = \underset{a^T a=1}{\operatorname{argmax}}(ya)$$

Subjected to the constraint $a_i^T \psi_y a_j$, for all $(1 \le i \le j)$, where ψ_y is covariance matrix of the original data. The orthogonal constraint ensures that the linear combinations are uncorrelated. Linear combinations $f_i = ya_i$ are known as the principal components. These linear combinations represent the selection of a new coordinate system obtained by rotating the original system. The new axes represent the directions with maximum variability, and are ordered in the amount terms of original data variation they account for. The first principal component accounts for as much of the variability in the original data as possible, and each succeeding component accounts for as much of the remaining variability as possible. Computations of the weighting vectors ‘a’ involves the calculation of the eigenvalue decomposition of ‘a’ data covariance matrix ψ_y:

$$\psi_y a_i = \lambda_i a_i$$

Whereas λ_i is the ith eigenvalue in the descending order for $(i = 1, \dots, k)$ and a_i is the corresponding eigenvector. The eigenvalue λ_i measures the variance of the ith principal component and the eigenvector a_i provides the loadings for the linear transformation. The number of components ‘k’ is specified on the basis of researcher’s prior knowledge or determined using dedicated procedures. Class prediction using standard methods can then be carried out in the reduced space by using the constructed principal components.

2.2 Class Prediction

The high dimension of ‘p’ is then reduced to a lower dimension ‘k’ after dimensionality reduction. The original data matrix is adapted by a matrix of factors (n×k, where k<n), constructed by PCA, as described in the previous section. Once the k-factors are composed, prediction of the response classes using K-Nearest Neighbors (KNN) and C5.0 is taken into consideration.

2.2.1 K-Nearest Neighbors (KNN)

Nearest neighbor (KNN) is a widely used classifier for activity recognition introduced first by Fix and Hodges in 1951. KNN is very simple, highly efficient and effective algorithm for pattern recognition that classifies new observation into the class, to which the majority of its k nearest neighbors at training data set belong [15]. In other words, k most similar training data observations to unclassified observation are searched [16]. KNN involves large storage requirement and intensive computation, and the value of k also needs to be determined properly. In fact, the rating for the k-nearest neighbor based on the similarity is calculated using Euclidean distance which is defined as follows [17]:

$$D(x, y) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$$

The k-nearest neighbor algorithm can be written as follows:

- Let k be number of nearest neighbors and D be the set of training samples y_i
- For each test sample x_i do compute using Euclidean distance for every sample y_i of D:
- ▪ Select the k close set training samples to test sample x_i
- ▪ Classify the sample x_i based on major class among its nearest neighbors.

2.2.2 C5.0 Decision trees

C5.0 algorithm, a commercial system from RuleQuest Research, is a new generation of decision trees algorithms. It has been developed as an improved version of well-known and widely used C4.5 classifier with many features like [18]:

- C5.0 algorithm can respond on noise and missing data.
- C5.0 provides boosting.
- A large decision tree may be difficult to read and comprehend.
- C5.0 provides the option of viewing the large decision tree as a set of rules which is easy to understand.
- Overfitting is solved by the C5.0 and Reduce error pruning technique.
- C5.0 can also predict which attributes are relevant in classification and which are not. This technique, known as Winnowing is especially useful while dealing with high dimensional datasets.

The use of this methods recurred that the root node at the top of the tree considers all samples and passes them through to the second node called “branch node”. The branch node generates rules for a group of samples based on an entropy measure. In this stage, C5.0 constructs a very large tree by considering all attribute values and finalizes the decision rule by pruning. It uses a heuristic approach for pruning based on splits statistical significance. After fixing the best rule, the branch nodes send the final class value in the last node, called the “leaf node”.

3 Methods

This section describes the dataset, the preprocessing conditions and the framework steps.

3.1 Dataset and Preprocessing

The established algorithms are applied to a publicly available dataset ‘Human Activity Recognition Using Smartphones Data Set’. HAR-database built from the recordings of 30 subjects performing activities of daily living (ADL) while carrying a waist-mounted smartphone with embedded inertial sensors. The dataset contains 561 attributes with 10299 instances. The experiments have been carried out with a group of 30 volunteers within an age bracket of 19-48 years. Each person performed six different activities (walking, walking upstairs, walking downstairs, sitting, standing, laying) wearing a smartphone on the waist. They used an accelerometer and a gyroscope in a Samsung Galaxy S2 phone to show that their model can distinguish these six basic activities. Detailed information about this dataset is reported by Anguita et al. [5]. The authors proposed a hardware-friendly multi-class SVM model that adapted a support vector machine for a smartphone activity recognition application that is used in healthcare with a performance of 89%.

A general framework is suggested for the structural (HAR) problem in which we focus on feature extraction and classification (Figure 1). The proposed method for human activity recognition is based on feature extraction using PCA and classification using KNN and C5.0.

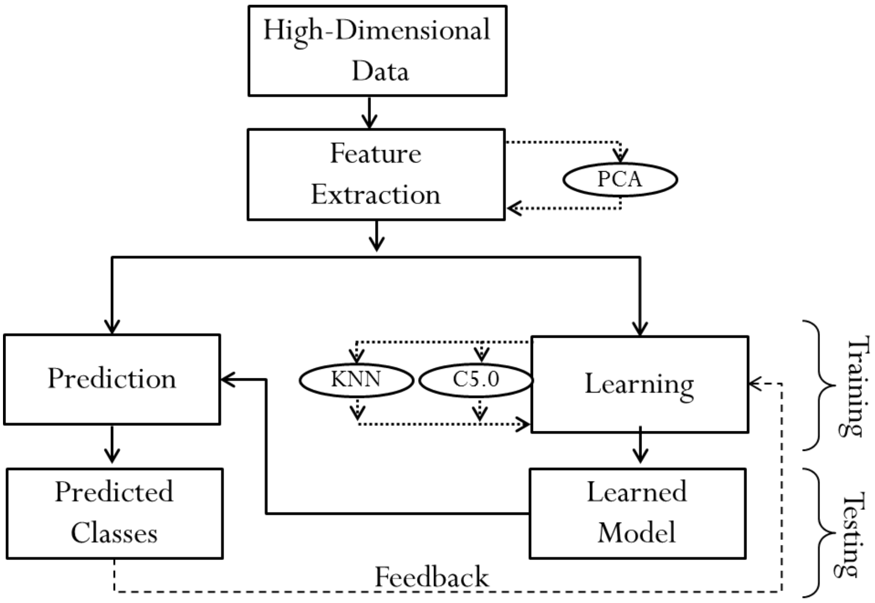

Figure 1: A dimensionality reduction framework (HAR)

3.2 Feature Extraction

Although the procedure described here can handle a large number of attributes, it may still be too large for practical use. While, the model assessment procedure requires fitting the data many times, it is very time-consuming process due to cross-validation and re-randomization. In addition, a considerable percentage of features do not show differential expression between the groups and only a subset of features is interest worthy. The feature extraction step by PCA is used to reduce large input sensor data to a smaller set of features (feature vector), which preserves information contained in the original data.

3.3 Class Prediction

The high dimension ‘p’ is then reduced to a lower dimension ‘k’ after dimension reduction. The original data matrix is constructed by a matrix of factors ($n\times k$, where $k< n$), constructed by (PCA), as described in the previous section. Once the k-factors are composed, prediction of the response classes using KNN and C5.0 is then made.

3.4 Performance Evaluation

Each model developed has the performance that has been measured in terms of the average accuracy, which means the number of correctly classifying cases under the total number cases in a testing set. The dataset is divided into a training set and a testing set. Comparing the classification performance of the three models, namely PCA-KNN, PCA-C5.0, can be realized by accuracy rate, which is the direct criteria to evaluate the classification models. It can be quantitatively evaluated by the following expression:

$$Accuracy = \frac{(The\ number\ correctly\ classified\ cases)}{(The\ total\ number\ of\ cases)}$$

After data preprocessing, the proposed performance evaluation procedure on the dataset is applied. The performance of each developed model was measured in terms of the average accuracy. Accuracy is the number of correctly classifying cases under the total number of cases in a testing set. The experiments are performed with training data and test data. The size is chosen differently and dependently on the available dataset in order to provide a reliable estimate and validate the developed models.

4 Results and discussion

The interest of dimensionality reduction by considering applications for the class prediction of HAR-data has been illustrated. We compare the results obtained by our approach with the performance of the direct classification approach.

4.1 Application to HAR-Dataset

After data preprocessing, the proposed performance evaluation procedure on the HAR-dataset is applied. We choose 8000 instances randomly to train model and the 2299 remainder instances are used to test the model. Implementation of this framework provide us the estimation of factors 'k' by dimensionality reduction and the classification accuracy performances.

4.2 Results

Table 1 represents the results obtained by applying the four models: PCA-KNN, PCA-C5.0, KNN and C5.0 on the experimental dataset. 99.19% and 99.22% have been obtained within the KNN and C5.0 respectively without dimension reduction. In addition, the time spent for the execution of the analysis is short for the C5.0 model compared to KNN model (Approximately 20% timeless). The PCA-KNN accuracy ranged from 94.83% for k=26, to 91.36% for k=8. On the other hand, the PCA-C5.0 accuracy results ranged from 94.89% for k=26, to 90.37% for k=8. It appears that PCA-C5.0 and PCA-KNN models accuracy is the same for k=26. However, PCA-KNN accuracy for k=8 is higher than PCA-C5.0.

In term of time of execution, KNN and C5.0 without dimensionality reduction requested 148080 seconds and 7594 seconds respectively. The PCA-C5.0 model seems to be very interesting; the execution time ranged from 7 seconds for k=26 to

2 seconds for k=8. Contrarily, PCA- KNN time of execution ranged from 171 seconds for k=26 to 30 seconds for k=8.

Table 1: Framework Results

(HAR) Attributes (p)	Dimensionality Reduction Model	Factors (k)	Pattern Recognition Models	Classification Accuracy %	Time of execution seconds
561	Pristine	561	KNN	99.19	148080
			C5.0	99.22	7594
	PCA	26	PCA-KNN	94.83	171
		15		93.32	58
		8		91.36	30
		26	PCA-C5.0	94.89	7
		15		93.28	4
		8		90.37	2

4.3 Discussion

In this paper, we proposed a linear method for recognition of daily human activities based on feature dimensionality reduction using PCA to low dimensionality feature space followed by two class prediction models (KNN and C5.0) as decision functions for classification of human activities.

Concerning dimensionality reduction task, three different criteria have been used to estimate the number of factors. The first criterion is the higher eigenvalue to one which provides us 8 factors. In the second criterion, the cumulative variance equal to 80% which produces 26 factors. After dimensionality reduction, the results of class prediction show that, for k=26 the accuracy is the same for the two class prediction models (PCA-KNN and PCA-C5.0) and for k=8 the highest accuracy is obtained from PCA- KNN model followed by PCA-C5.0.

In the case of third criterion, maximizing the classification accuracy while minimizing the number of factors is attempted. In this context, the factors minimum is k=15 representing almost half of second criterion (k=26) and related to classification accuracy of PCA-KNN and PCA-C5.0 (93.28% and 93.32% respectively) which is a near classification accuracy of the second criterion (94.89% and 94.89% for PCA-KNN and PCA-C5.0 respectively).

Moreover, the execution time of the third criterion with PCA-KNN model is equal to one third of the first criterion while it’s about half for PCA-C5.0 model. Though, it is evident that PCA-C5.0 model accomplished the very interesting time of execution in comparison with PCA-KNN.

5 Conclusion

This work reports an efficient dimensionality reduction approach for the class prediction related to the human activity recognition data composed by several activeties, such as lying down, walking, sitting, standing, ascending and descending stairs.

Primarily, the main advantage of this approach is to reduce the number of features from p=561 to k=26 through PCA, which provides a compact representation by projecting the data onto a feature space that captures the most representative features. After dimension reduction, classification techniques showed that, for k=26, the best accuracy is obtained by PCA-C5.0 model (94.89% in 7 seconds), followed by PCA-KNN model (94.83 % in 171 seconds).

Interestingly, when a small number of features is used in the classification (e.g. k=15), PCA-C5.0 model still gives better results (93.28%) related to time of execution (4 seconds) than PCA-KNN (93.32% in 58 seconds).

These results are competitive with prior activity recognition results of Anguita et al. [5] who used Support Vector Machine (SVM) classifier for the same dataset where 70% of the data was used for training and the rest for testing and reported that classification result of the standard floating-point Multiclass SVM (MC-SVM) and the Hardware-Friendly SVM (MC-HF-SVM) with small among of memory (k=8 bits) were 89.3% and 89.0% respectively.

A recent study using the same dataset, has shown that the best accuracy was obtained by J48 Decision tree, Random Forests, Random Committee and Lazy IBk classifiers with more than 90% for 128 features and 256 features and around 60% for 8 features [19]. However, Naïve Bayes and K-means classifiers performed poorly.

In terms of building time, IBk classifier performed the best building time of 0 seconds and 0.1 seconds for 128 features and 256 features respectively. By contrast, the time taken to build the first three models (J48 Decision tree, Random Forests, Random Committee) ranged from 20.1 seconds to 64.6 seconds [19].

Another study using Pareto Ensemble Pruning (PEP) approach, showed that PEP achieves an accuracy of 90.4% [20]

Overall, the PCA dimensionality reduction method can improve the performance of the k-nearest neighbor (KNN) and C5.0 classifiers for HAR high-dimensional data sets. The suggested method is capable of addressing the important dimensionality issues as well. The application of the proposed method on the experimental dataset results in the achievement of the best performance for dimensionality reduction (in terms of least time consumption and CPU-expenditure).

References

[1] O.C. Ann, L.B. Theng, Human activity recognition: a review, *Proceedings of the IEEE International Conference on Control System, Computing and Engineering (ICCSCE '14)*, Batu Ferringhi, Malaysia, IEEE, (2014), 389-393. http://dx.doi.org/10.1109/iccsce.2014.7072750

[2] J. Aggarwal and L. Xia, Human activity recognition from 3d data: A review, *Pattern Recognition Letters*, **48** (2014), 70-80.
http://dx.doi.org/10.1016/j.patrec.2014.04.011

[3] F. Attal, S. Mohammed, M. Dedabrishvili, F. Chamroukhi, L. Oukhellou, Y. Amirat, Physical Human Activity Recognition Using Wearable Sensors, *Sensors*, **15** (2015), 31314-31338. http://dx.doi.org/10.3390/s151229858

[4] K. Altun, B. Barshan, O. Tunçel, Comparative study on classifying human activities with miniature inertial and magnetic sensors, *Pattern Recognition*, **43** (2010), 3605-3620. http://dx.doi.org/10.1016/j.patcog.2010.04.019

[5] D. Anguita, A. Ghio, L. Oneto, X. Parra, J.L. Reyes-Ortiz, Human Activity Recognition on Smartphones Using a Multiclass Hardware-Friendly Support Vector Machine, Chapter in *Ambient Assisted Living and Home Care*, Springer Berlin Heidelberg, Germany, 2012, 216-223.
http://dx.doi.org/10.1007/978-3-642-35395-6_30

[6] A. Bayat, M. Pomplun, D. A. Tran, A study on human activity recognition using accelerometer data from smartphones, *Procedia Comput. Science*, **34** (2014), 450-457. http://dx.doi.org/10.1016/j.procs.2014.07.009

[7] L. Bedogni, M. Di Felice, L. Bononi, By train or by car? Detecting the user's motion type through smartphone sensors data, *In Proceedings of the 2012 IFIP Wireless Days* (WD), (2012), 1-6. http://dx.doi.org/10.1109/wd.2012.6402818

[8] N. A. Capela, E. D. Lemaire, N. Baddour, M. Rudolf, N. Goljar, H. Burger, Evaluation of a smartphone human activity recognition application with able-bodied and stroke participants, *Journal of NeuroEngineering and Rehabilitation*, **13** (2016), no. 5. http://dx.doi.org/10.1186/s12984-016-0114-0

[9] R. Damaševičius, M. Vasiljevas, J. Šalkevičius, M. Woźniak, Human Activity Recognition in AAL Environments Using Random Projections, *Comput. Math. Methods in Med.*, **2016** (2016), http://dx.doi.org/10.1155/2016/4073584

[10] S. Ghose, J. Mitra, M. Karunanithi, J. Dowling, Human Activity Recognition from Smart-Phone Sensor Data using a Multi-Class Ensemble Learning in Home Monitoring, Chapter in Driving Reform: Digital Health is Everyone’s Business, IOS Press, 214, 2015, 62-67. http://dx.doi.org/10.3233/978-1-61499-558-6-62

[11] P. Siirtola, *Recognizing Human Activities Based on Wearable Inertial Measurements- Methods and Applications*, Doctoral Dissertation, Department of Computer Science and Engineering, University of Oulu, (Acta Univ Oul C 524), 2015.

[12] E. Kim, S. Helal, D. Cook, Human activity recognition and pattern discovery, *IEEE Pervas. Comput*., **9** (2010), no. 1, 48-53. http://dx.doi.org/10.1109/mprv.2010.7

[13] I. El Moudden, S. El Bernoussi, B. Benyacoub, A dimensionality reduction framework for automatic speech recognition, *Proceedings of the 26th International Business Information Management Association Conference-Innovation Management and Sustainable Economic Competitive Advantage: From Regional Development to Global Growth*, IBIMA (2015), 2602-2608.

[14] I. T. Jolliffe, *Principal Component Analysis*, Springer, New York, 2002. http://dx.doi.org/10.1007/b98835

[15] E. Fix, J. L. Hodges, *Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties*, US Air Force School of Aviation Medicine Technical Report, **4** (1951), 477.

[16] D. J. Hand, H. Mannila, P. Smyth, *Principles of Data Mining*, MIT Press, Cambridge, USA, 2001.

[17] A. Karegowda, M. Jayaram, A. Manjunath, Cascading k-means clustering and k-nearest neighbor classifier for categorization of diabetic patients, *International Journal of Engineering and Advanced Technology*, **1** (2012), 147-151.

[18] A. S. Galathiya, A. P. Ganatra and C. K. Bhensdadia. Improved Decision Tree Induction Algorithm with Feature Selection, Cross Validation, Model Complexity and Reduced Error Pruning, *International Journal of Computer Science and Information Technologies*, **3** (2012), no. 2, 3427-3431.

[19] G. Chetty, M. White, F. Akther, Smart phone based data mining for human activity recognition, *Procedia Comput. Sci.*, **46** (2015), 1181-1187. http://dx.doi.org/10.1016/j.procs.2015.01.031

[20] C. Qian, Y. Yu and Z.-H. Zhou, Pareto ensemble pruning, *In Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI'15)*, 2935-2941, Austin, TX, 2015.

Received: July 18, 2016; Published: August 31, 2016