Expert Systems with Applications 42 (2015) 5800–5810

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Activity discovering and modelling with labelled and unlabelled data in smart environments

Jiahui Wen *, Mingyang Zhong

School of Information Technology and Electrical Engineering, The University of Queensland, QLD 4072, Australia

ARTICLE INFO

Article history:
Available online 11 April 2015

Keywords:
Data mining
Machine learning
Activity recognition
Similarity measurement
Labelled and unlabelled data
Smart environments

ABSTRACT

In the past decades, activity recognition had aroused great interest for the community of context-awareness computing and human behaviours monitoring. However, most of the previous works focus on supervised methods in which the data labelling is known to be time-consuming and sometimes error-prone. In addition, due to the randomness and erratic nature of human behaviours in realistic environments, supervised models trained with data from certain subject might not be scaled to others. Further more, unsupervised methods, with little knowledge about the activities to be recognised, might result in poor performance and high clustering overhead. To this end, we propose an activity recognition model with labelled and unlabelled data in smart environments. With small amount of labelled data, we discover activity patterns from unlabelled data based on proposed similarity measurement algorithm. Our system does not require large amount of data to be labelled while the proposed similarity measurement method is effective to discover length-varying, disordered and discontinuous activity patterns in smart environments. Therefore, our methods yield comparable performance with much less labelled data when compared with traditional supervised activity recognition, and achieve higher accuracy with lower clustering overhead compared with unsupervised methods. The experiments based on real datasets from the smart environments demonstrate the effectiveness of our method, being able to discover more than 90% of original activities from the unlabelled data, and the comparative experiments show that our methods are capable of providing a better trade-off, regarding the accuracy, overhead and labelling efforts, between the supervised and unsupervised methods.

1. Introduction

The application of machine learning and data mining techniques in pervasive computing, and the increasingly growing demand of health-care monitoring for elderly and cognitive impaired people (Chernbumroong, Cang, Atkins, & Yu, 2013) have offered unprecedented opportunities to research regarding activity recognition in smart environments.

Human activities are usually characterised by the signals from external sensors, and then can be recognised by analysing these signals. In the area of data-driven sensor-based activity recognition, generally, there are two methods to gather the activity information based on the types of sensors. One is to attach inertial sensors such as accelerometers onto human body to capture physical signals such as acceleration and angular velocity, and the activities that are to be recognised are usually physical motions such as sitting, standing and running. By contrast, the other one gathers information from the sensors deployed around the environments, and the information is usually the sensor events triggered during the interaction between human and the environments. The argument for this method is that high-level activities usually share common sets of physical actions, and are difficult to differentiate relying solely on physical signals. Furthermore, people have various ways to perform the high-level activities, and even the same individual may perform the same activity in different ways. However, these kinds of high-level activities could be characterised by the objects used by people, people's location and the time they perform the activities, all of which could be obtained from sensors such as electrical ID tags deployed in the environments.

Once the sensor readings are obtained and annotated, traditional activity recognition models are trained with the data and used to recognise the activities from the further sensor readings. Specifically, the classification model creates a map between the sensor readings and the annotated activities during the learning phase, and the map is interpreted by the parameters of the models, such as the conditional probability in Naive Bayesian, emission probability in Hidden Markov Models (HMM) and decisive nodes

* Corresponding author.
E-mail addresses: j.wen@uq.edu.au (J. Wen), m.zhong1@uq.edu.au (M. Zhong).

http://dx.doi.org/10.1016/j.eswa.2015.04.005
0957-4174/

in the Decision Tree. However, these kinds of supervised learning methods have their own drawbacks. First of all, data annotating is time-consuming and sometimes error-prone, as sensor readings of the datasets from smart environments usually last for several months and contain millions of sensor event logs. Further, due to the randomness and the erratic nature of human behaviours in realistic situation, different people perform the same activities in various ways, and even the same individual shows an evolution of performing the same activity. Therefore, statically trained models often sacrifice the accuracy when they are transferred to other users. Finally, supervised models only focus on predefined activities, and overlook the fact that hidden behaviours are also beneficial for us. Sometimes the accuracy of the models would be impacted, if only the predefined activities are modelled. For example, Cook, Krishnan, and Rashidi (2013) find that unlabelled data usually accounts for more than 50% of the dataset and is treated as *other class* activity. As a result, the accuracy of the supervised models are affected simply because instances of predefined activities are sometimes classified as *other class* activity.

Considering the limitations of supervised models, people turn to semi-supervised methods for help, in which a small set of labelled data are used to create the initial model, and the test instances classified with high confidence are selected to refine the model. However, most of them focus on physical activities, while the semi-supervised learning in home setting is still open for exploiting, such as the similarity measurement between feature vectors. Further, activity classes are predefined in all the previous work, which does not meet the requirement of smart environment that involves numerous hidden activities and even meaningless ones such as transitional sensor events between activities. Unsupervised methods (Cook et al., 2013) are also proposed to discover activities from unlabelled data, in which data mining techniques such as clustering is used to cluster the discovered patterns into the specified number of centroids (Rashidi, Cook, Holder, & Schmitter-Edgecombe, 2011). However, with limited knowledge about the activities to be recognised, they result in poor performance. Further more, the clustering overhead of unsupervised methods has never been studied before. The overhead cannot be overlooked because of the huge clustering space of smart home datasets.

In order to tackle the problems stated above, we propose a model to recognise activities in smart environments with labelled and unlabelled data. With small amount of labelled data and data mining techniques such as similarity measuring and clustering, we are able to discover activities from the unlabelled data, as well as hidden frequent patterns that might be beneficial for us. After that, training dataset is created with the discovered activities to train traditional classifiers that recognise activities from further sensor readings.

The contributions of this paper can be concluded as follows:

1. We develop an activity recognition model with labelled and unlabelled data, which is able to provide a trade-off regarding the accuracy, overhead and labelling efforts between the supervised and unsupervised method. Compared with supervised methods and unsupervised methods, we are able to obtain comparable recognition accuracy with using much less labelled data, and achieve higher accuracy with lower clustering overhead respectively.
2. We propose a similarity measurement method for discovering predefined activity patterns from unlabelled data. The proposed similarity measurement is invulnerable to the length of the window used to segment the sensor event sequences. In addition, it is also robust to the characteristics (such as length-varying, disordered and repeating) of the sensor event sequences in smart environments, compared with traditional string similarity algorithms.
3. We validate our model with publicly available datasets, and analyse its effectiveness through comprehensive experimental and comparison studies. Specifically, we compare our methods with traditional supervised and unsupervised techniques for activity recognition in terms of accuracy, clustering overhead and the amount of labelled data, and compare the proposed similarity measurement method with traditional string similarity algorithms with regard to the ability of discovering activity from unlabelled data.

2. Related work

Generally, the models recognising human activities can be classified into two categories: knowledge-driven models and data-driven models. In knowledge-driven models, the activities are usually represented in the form of rules specified with common sense, and the models have an advantage in being reused among different environments. However, the limitation of the statically and strictly defined rules makes the models being unable to deal with noises and uncertain information in sensor readings (Gu, Chen, Tao, & Lu, 2010). By contrast, data-driven models, which are trained with realistic data, are more powerful when facing the characteristics of randomness and erratic nature of human behaviours. To name a few, they include Naive Bayesian used in Bao and Intille (2004) and Tapia, Intille, and Larson (2004), HMM used in Patterson, Fox, Kautz, and Philipose (2005) and Van Kasteren, Noulas, Englebienne, and Kröse (2008), Support Vector Machine (SVM) in Cook et al. (2013), Brdiczka, Crowley, and Reignier (2009) and Zhan, Faux, and Ramos (2014), Decision Trees in Bao and Intille (2004) and Hevesi, Wille, Pirkl, Wehn, and Lukowicz (2014), KNN in Sundholm, Cheng, Zhou, Sethi, and Lukowicz (2014) and Hevesi et al. (2014) and Conditional Random Fields (CRF) in Vail, Veloso, and Lafferty (2007) and Zhan et al. (2014).

The aforementioned works leverage supervised methods to obtain high accuracy, but labelling the activity data is expensive and time-consuming. A growing number of recent researches focus on semi-supervised and unsupervised approaches to minimise data labelling efforts. For example, Lee and Cho (2014) apply global–local co-training algorithm with both labelled and unlabelled data, and the test instances classified with high confidence are fed into model in order to improve the accuracy. Stikic, Van Laerhoven, and Schiele (2008), Stikic, Larlus, and Schiele (2009) and Stikic and Schiele (2009) explore semi-supervised methods such as self-training, co-training, active learning and multi-instance learning for activity recognition with less labelled data. Maekawa and Watanabe (2011) introduce an unsupervised method to recognise the target user's activity with the model created by the data from source users that present similar physical characteristics to the target user. While others (Huynh, Fritz, & Schiele, 2008; Seitr, Chiu, Fritz, Amft, & Troster, in press; Sun, Yeh, Cheng, Kuo, & Griss, 2014) use the topic model to discover frequent patterns from the daily life. However, all the mentioned learning techniques have two limitations. On one hand, the activities they are to recognise are predefined and the hidden ones are ignored, hence those methods are lack of scalability in smart environments. On the other hand, most of the models focus on low-level physical activities recognition, in which the streaming data is segmented into windows and the features extracted from them are used to construct test instances (Fuentes, Gonzalez-Abril, Angulo, & Ortega, 2012), which is inapplicable to smart environments in which the sensor data is categorical rather than numerical, hence a new similarity measurement is required to evaluate the similarity of sensor event sequences.

In addition, many previous researches propose unsupervised methods based on automatically mined common sense (Wyatt, Philipose, & Choudhury, 2005). Gu et al. (2010) propose an