Pattern Recognition Letters 000 (2015) 1–9

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Factored four way conditional restricted Boltzmann machines for activity recognition[☆]

Decebal Constantin Mocanu[a,*], Haitham Bou Ammar[b], Dietwig Lowet[c], Kurt Driessens[d], Antonio Liotta[a], Gerhard Weiss[d], Karl Tuyls[e]

[a] Eindhoven University of Technology, Department of Electrical Engineering, Den Dolech 2, Eindhoven, 5612 AZ, The Netherlands
[b] University of Pennsylvania, GRASP Laboratory, 3330 Walnut Street, Philadelphia, PA 19104-6228, USA
[c] Philips Research, Human Interaction and Experiences, High Tech Campus 34, Eindhoven, 5656 AE, The Netherlands
[d] Maastricht University, Department of Knowledge Engineering, Bouillonstraat 8-10, Maastricht, 6211 LH, The Netherlands
[e] University of Liverpool, Department of Computer Science, Ashton Street, Liverpool, L69 3BX, United Kingdom

ARTICLE INFO

Article history:
Received 24 February 2014
Available online xxxx

Keywords:
Activity recognition
Deep learning
Restricted Boltzmann machines

ABSTRACT

This paper introduces a new learning algorithm for human activity recognition capable of simultaneous regression and classification. Building upon conditional restricted Boltzmann machines (CRBMs), Factored four way conditional restricted Boltzmann machines (FFW-CRBMs) incorporate a new label layer and four-way interactions among the neurons from the different layers. The additional layer gives the classification nodes a similar strong multiplicative effect compared to the other layers, and avoids that the classification neurons are overwhelmed by the (much larger set of) other neurons. This makes FFW-CRBMs capable of performing activity recognition, prediction and self auto evaluation of classification within one unified framework. As a second contribution, sequential Markov chain contrastive divergence (SMcCD) is introduced. SMcCD modifies Contrastive Divergence to compensate for the extra complexity of FFW-CRBMs during training. Two sets of experiments one on benchmark datasets and one a robotic platform for smart companions show the effectiveness of FFW-CRBMs.

1. Introduction

Robotic support for elderly people requires (possibly among others) capabilities such as monitoring and coaching [1,2], e.g., emergency detection and medication reminders, and accurate activity detection is vital for such services. On the monitoring side, a system that recognises human activity patterns allows for automated health guidance, as well as providing an objective measure for medical staff. Specifically, the fashion in which these daily activities are executed (e.g., speed, fluency) can serve as an important early indicator of possible problems. Accurate activity recognition is made difficult by the continuous nature of typical activity scenarios, which makes the task highly similar to time series prediction.

Much research has been aimed at detecting human activities based on the output of a variety of low-power, low-bandwidth sensors, such as passive infrared (PIR) sensors, and power and pressure meters placed either around the home, or on-body (e.g. accelerometers [3,4]). The drawback of such an approach lies in the inability to capture sufficiently reliable data that allows to differentiate between subtly different activities. In principle, the most accurate and suited sensors for activity recognition would be video-cameras in combination with advanced computer vision algorithms to interpret the data, but this approach leads to significant privacy issues.

As an alternative, we make use of motion capture data. More exactly, we use a Kinect® sensor[1] to generate a 3D point cloud and to extract the human skeleton joints from it. This approach yields relatively easy data to process and, as we will show, sufficient information to accurately recognise human activities.

Literature provides other techniques that can do both classification (i.e. from a set of possible time series categories determine to which category a new observation belongs or, in our case, recognise the activity performed by a person during a specific moment of time) and time series prediction (i.e. starting from the near history observations forecast the next values for a specific time series or, in our case, forecast the human body's movements or poses in the near future), each of them with their advantages and disadvantages. Among them, Linear Dynamic Systems such as Autoregression and Kalman filters are well suited to model linear time series. Although

[☆] This paper has been recommended for acceptance by G. Sanniti di Baja.

[*] Corresponding author. Tel.: +31 685 342 994.
E-mail address: d.c.mocanu@tue.nl (D.C. Mocanu).

[1] http://en.wikipedia.org/wiki/Kinect, [Accessed 8th June 2014].

http://dx.doi.org/10.1016/j.patrec.2015.01.013
0167-8655/© 2015 Elsevier B.V. All rights reserved.

extensions for non-linear systems exist, they still have difficulties with high non-linearity. Another successful class of time series models are hidden Markov models (HMMs). HMM have reached a lot of success in speech recognition. However, HMM models are also less suited for highly non-linear data and become unwieldy when the state space is large.

Recent research has profiled deep learning (DL) methods [5] as a promising alternative for pattern recognition problems. DL makes small steps towards mimicking the behaviour of the human brain [6,7]. It has been successfully applied to, for example, multi-class classification [8], collaborative filtering [9] and information retrieval [10]. Due to the success of DL based on restricted Boltzman machines (RBMs) [11] in modelling static data, a number of extensions for modelling time series have been developed. A straightforward extension of restricted Boltzmann machines to model time series are Temporal RBMs (TRBM) as described in [12]. Conceptually, a TRBM consists of a succession of RBMs (one for each time frame) with directed connections between the nodes representing consecutive timeframes. However, a lack of an efficient training method limits their application to real-world problems. Conditional RBMs (CRBM) propose a different extension of RBMs for modelling time sequences where two separate visible layers represent (i) the values from N previous time frames and (ii) those of the current time frame [13]. A CRBM can be viewed as adding AutoRegression to RBMs and hence are especially suited for modelling linear time variations. They have been successfully applied to motion capture data. To enable also the modelling of non-linear time variations, the CRBM concept has been further extended by incorporating three-way neural nodes interactions that are connected by a 3-way weight tensor [13]. To overcome the computational complexity induced by the 3-way weight tensor, the tensor can be factored resulting in a Factored Conditional RBM (FCRBM) [14]. These FCRBMs have been shown to give excellent results in modelling and predicting motion capture data. They are able to predict different human motion styles and combine two different styles into a new one.

To our knowledge, FCRBMs represent the current state of the art for capturing and predicting human motion, and therefore, we chose them as a basis for our work on activity recognition. However, the FCRBM is still not optimally suited to classify human motion or activities. The reason for this is that the hidden neurons in the FCRBM are used to model how the next frame of coordinates depends on the historic frames. The most natural way to extend the FCRBM to include classification capabilities is by letting the hidden neurons gate the interactions between the label and the prediction neurons. This results in a model with four-way neuron interactions.[2]

Hence, in this paper we propose a novel model, namely *Factored Four Way Conditional Restricted Boltzmann Machine* (FFW-CRBM) capable of both classification and prediction of human activity in one unified framework. An emergent feature of FFW-CRBM, so called self auto evaluation of the classification performance, may be very useful in the context of smart companions. It allows the machine to autonomously recognise that an activity is undetected and to trigger a retraining procedure. Due to the complexity of the proposed machine, the standard training method for DL models is unsuited. As a second contribution, we introduce *Sequential Markov chain Contrastive Divergence* (SMcCD), an adaptation of contrastive divergence (CD) [16]. To illustrate the efficacy and effectiveness of the model, we present results from two sets of experiments using real world data originating from (i) our previous developed smart companion robotic platform [17] and (ii) a benchmark database for activity recognition [18].

The remaining of this paper is organised as follows. Section 2 presents the mathematical definition of the problem tackled in this article. Section 3 presents background knowledge on deep learning for the benefit of the non-specialist reader. Section 4 details the mathematical model for the unfactorised version of the proposed method. Section 5 describes the FFW-CRBM model including the mathematical modelling. Section 6 describes the experiments performed and depict the achieved results. Finally, Section 7 concludes and presents directions of future research.

2. Problem definition

In essence, in this paper, we aim at solving time series classification and prediction simultaneously in one unified framework. Let $i \in \mathbb{N}$ represent the index of available instances, $t \in \mathbb{N}$ to denote time, $\mathbb{R}^d$ a d-dimensional feature space, $t-N:t-1$ the temporal window of observations recorded in the N time steps before t, $\mathcal{C} = \{0, 1, \ldots, k\}$ the set of possible classes, and $\mathbf{\Theta}$ the parameters of a generic mathematical model. The targeted problem can then be written as:

Given a data set $\mathcal{D} = \{\mathbf{X}^{(i)}, \mathbf{y}^{(i)}\}$ for all instances i, where:

- $\mathbf{X}^{(i)} \in \mathbb{R}^{d\times(t-N:t-1)}$, is a real-valued input matrix consisting of d rows of features, and $N-1$ columns corresponding to the associated temporal window $t-N:t-1$,
- $\mathbf{y}_t^{(i)} \in \mathbb{R}^d \times \mathcal{C}$ is the corresponding multidimensional output vector consisting of the d-dimensional real-valued features at time t and an associated class label (e.g. a robotic companion that recognises an activity and predict the corresponding human poses to avoid collision).

Determine $p(\mathbf{Y}|\mathbf{\Gamma}; \mathbf{\Theta})$, with $\mathbf{Y} = \{\mathbf{y}^{(i)}\}\ \forall i$ and $\mathbf{\Gamma} = \{\mathbf{X}^{(i)}\}\ \forall i$ representing the concatenation of all outputs and inputs respectively, such that: $\text{KL}(p_{\text{model}}(\mathbf{Y}|\mathbf{\Gamma}; \mathbf{\Theta})||p_{\text{empirical}}(\mathbf{Y}|\mathbf{\Gamma}))$ is minimised. KL represents the Kullback Leibler divergence between the empirical and approximated (i.e., model) distributions. This is signified by $p_{\text{model}}(\mathbf{Y}|\mathbf{\Gamma}; \mathbf{\Theta})$, which defines a joint distribution over $\mathbb{R}^d \times \mathcal{C}$ space.

3. Background

This section provides background knowledge needed for the remainder of the paper. Firstly, restricted Boltzmann machines (RBMs), being at the basis of the proposed technique, are detailed. Secondly, contrastive divergence, the algorithm used to fit the RBM's hyperparameters is detailed. Finally, factored conditional restricted Boltzmann machines, constituting the main motivation behind this work, are explained.

3.1. Restricted Boltzmann machine

Restricted Boltzmann machines (RBM) [11] are energy-based models for unsupervised learning. These models are stochastic with stochastic nodes and layers, making them less vulnerable to local minima [14]. Further, due to their neural configurations, RBMs posses excellent generalisation capabilities [5].

Formally, an RBM consists of visible and hidden binary layers. The visible layer represents the data, while the hidden increases the learning capacity by enlarging the class of distributions that can be represented to an arbitrary complexity. This paper uses the following notation: i represents the indices of the visible layer, j those of the hidden layer, and w_{ij} denotes the weight connection between the ith visible and jth hidden unit. Further, v_i and h_j denote the state of the ith visible and jth hidden unit, respectively. Using to the above notation, the energy function of an RBM is given by:

$$E(v, h) = -\sum_{i=1}^{n_v}\sum_{j=1}^{n_h} v_i h_j w_{ij} - \sum_{i=1}^{n_v} v_i a_i - \sum_{j=1}^{n_h} h_j b_j \tag{1}$$

where, a_i and b_j represent the biases of the visible and hidden layers, respectively; n_v and n_h are the number of neurons in the visible and hidden layer, respectively. The joint probability of a state

[2] Four-way (and higher) interactions are also biologically plausible since they appear to be necessary to explain the workings of the human brain [15].

Please cite this article as: D. C. Mocanu, et al., Factored four way conditional restricted Boltzmann machines for activity recognition, Pattern Recognition Letters (2015), http://dx.doi.org/10.1016/j.patrec.2015.01.013