Knowledge-Based Systems xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Cross-correlation conditional restricted Boltzmann machines for modeling motion style

Chunzhi Xie[a], Jiancheng Lv[*,a], Yunxia Li[b], Yongsheng Sang[a]

[a] Machine Intelligence Laboratory, College of Computer Science, Sichuan University, Chengdu 610065, PR China
[b] School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu 610065, PR China

ARTICLE INFO

Keywords:
CRBM
Deep learning
Temporal dependency
Cross-correlation
Unsupervised learning

ABSTRACT

Temporal dependency plays a fundamental role in nonlinear generative models for capturing temporal features. One such model, the conditional restricted Boltzmann machine (CRBM), learns these temporal dependencies by considering the visible variables in the previous time slice as additional fixed inputs so that static and temporal features can be captured simultaneously to generate new human motions. However, the temporal dependencies in the CRBM fail to describe various common equilibrium postures in human motion. In this paper, we present cross-correlation as a new representation for modeling temporal dependencies by introducing the Pearson correlation coefficient. We also propose an approach to enhance the discrimination of the CRBM for various human motions by incorporating cross-correlation and temporal dependency features. The experimental results on benchmark databases demonstrate that the proposed method not only retains all the merits of the CRBM, such as exact inference and efficient learning, but also greatly improves the model's ability to blend motion styles and achieve smooth transitions between various motion segments.

1. Introduction

Modeling human motion has numerous applications, such as tracking, activity recognition, style and content separation, person identification, computer animation and synthesis of new motions [1–5]. Human motion is often represented in time sequential sensor data, with high dimensionality, much noise and extremely complex temporal dependencies. Traditional modeling methods for human motion include linear dynamical systems [6–8], hidden Markov models [9,10], auto-regression models [11,12], and spatio-temporal feature points [13,14]. Estimated parameters can be used as features for performing classification in these methods. However, traditional shallow methods, which only contain a small number of nonlinear operations, do not have the capacity to model complex, high-dimensional, and noisy real-world time-series data accurately [5].

Deep learning is leading to major advances in solving problems that have resisted the best attempts of the artificial intelligence community for many years [15]. Deep learning is good at capturing complex structures in high-dimensional data. For many sequence tasks (e.g., voice [16], video recognition [17,18], medical sequence data [19–21]), deep learning has been shown to be effective in determining good representations and classifiers [22–24]. Many algorithms for modeling time-series data, such as conditional restricted Boltzmann machines (CRBMs) [25,26], gated CRBMs [27], factored CRBMs (FCRBMs) [11], temporal restricted Boltzmann machines (TRBMs) [28], recurrent TRBMs (RTRBMs) (RTRBMs) [29], structured RTRBMs [30], recurrent neural networks [31–33], and temporal sigmoid belief networks [34], have been developed. Among these variants of restricted Boltzmann machine (RBM)-based methods, the CRBMs may be the first model for sequence modeling. It has been successfully applied in collaborative filtering, classification and motion modeling. Inspired by this model, many successors such as the FCRBM and RTRBM have been proposed.

The CRBM learns static features from sample frames. Moreover, it learns temporal features from these samples and then generates new sequences in which the new frames resemble the training frames. The core of the CRBM lies in temporal features, the essence of which is the structural features of the object in the frames. However, it fails to describe human structures in equilibrium, which is very common in human motion, such as walking, striding, running, and walking with running. In particular, in real-world data there are always several different motions in the same video in which the human body structure is in a state of equilibrium. However, it is difficult for the CRBM to distinguish various types of motion in equilibrium because the descriptions provided by the model are always the same or similar.

The traditional CRBM for time-series modeling only depends on the structure of the object, and does not consider correlations between

* Corresponding author.
E-mail addresses: lvjiancheng@scu.edu.cn, xcz_xihua@sina.com.cn (J. Lv).

https://doi.org/10.1016/j.knosys.2018.06.026
Received 2 August 2017; Received in revised form 23 June 2018; Accepted 28 June 2018

temporal objects. The Pearson correlation coefficient is used to quantitatively describe the degree of relatedness between two variables. It has been widely used in temporal data processing [35–37]. The Pearson correlation coefficient allows the temporal correlation between two Gaussian distributed variables to be measured.

In this paper, we first analyze the deficiencies of temporal dependencies in the CRBM, and then propose cross-correlation features based on the Pearson correlation coefficient. Furthermore, we obtain cross-correlation and temporal dependency (CCTD) features by combining cross-correlation features and temporal dependencies, and simulate the behavior of humans by applying the CCTD-based features to the CRBM model. Cross-correlation can describe joints which allow vigorous styles in motions, such as the raised foot in walking or the bent knee in jumping. This idea is consistent with spatio-temporal interest points (STIPs)-based methods [14]. However, STIPs-based methods only use manually designed STIPs to extract underlying features rather than learning spatio-temporal features automatically. Moreover, compared with the traditional CRBM, the CCTD-based CRBM (CCTD-CRBM) enhances temporal representations using the intensity of actions to capture temporal dependency features from different motions. Our experiments show that the proposed algorithm may help the CRBM to be a better generation model.

The remainder of this paper is organized as follows: In Section 2, we state some preliminaries and notations. Section 3, we present our motivation. In Section 4, we introduce CCTD features for temporal modeling. In Section 5, we describe the experiments conducted to examine the effectiveness of the proposed method. Finally, in Section 6,we conclude this paper.

2. Notations and preliminaries

An RBM [38–41] is an undirected graph that consists of visible units $v = (v_j)_{j\in M}$, hidden units $h = (h_i)_{i\in N}$, and weights $W = (W_{ji})_{j\in M, i\in N}$ that connect visible and hidden units, where M is the number of visible units and N is the number of hidden units. In an RBM, the joint distribution of the configuration (v, h) is given by

$$p(v, h) = \frac{e^{-\varepsilon(v,h)}}{\sum_v \sum_h e^{-\varepsilon(v,h)}}, \tag{1}$$

where $\varepsilon(v, h)$ is the energy of configuration (v, h), defined as

$$\varepsilon(v, h) = -\sum_{j=1}^{M} b_j v_j - \sum_{i=1}^{N} c_i h_i - \sum_{j=1}^{M}\sum_{i=1}^{N} w_{ji} h_i v_j, \tag{2}$$

where v_j and b_j are the status and bias of visible unit j, respectively, and h_i and c_i are the status and bias of hidden unit i, respectively.

A CRBM is a modified RBM for time-series data. The status of the visible and hidden layers in a CRBM at time step t correspond to visible units $v^{(t)}$ and hidden units $h^{(t)}$ respectively. There are also undirected weights that connect the visible and hidden layers; however, there are no connections within a layer, as in an RBM. Thus, a CRBM can capture static features at time step t. More importantly, there are S previously visible layers, which indicate the temporal relationships between the previous samples and current sample. Connection $B^{(t-s)}(s \in [1, ...S])$ between the previously visible $t-s$ layer and hidden layer is used to map the temporal features of the $t-s$ time step, where S is the number of previous frames (objects) that we expect the CRBM to learn. Connection $A^{(t-s)}$ between the previously visible $t-s$ layer and visible layer is used to add individual features so as to generate continuous individual frames (Fig. 1 (a)). Because the frame data are continuous, units in the CRBM should be linear and real-valued with noise. In CRBMs, the joint distribution of any configuration between the hidden and visible layers at time step t is defined by the log-likelihood:

$$\begin{aligned} logp^{(t)}(v, h) = & -\sum_i \frac{(v_i^{(t)} + VCon_i + c_i)^2}{2\sigma_i^2} \\ & + \sum_j (b_j + HCon_j) h_j^{(t)} \\ & + \sum_{i,j} \frac{v_i^{(t)}}{\sigma_i} h_j^{(t)} w_{ij} + const, \end{aligned} \tag{3}$$

where $VCon_i$ and $HCon_j$ are the impact of the temporal features on visible unit i and hidden unit j, respectively; c_i and b_j are the bias of visible unit i and hidden unit j, respectively; w_{ij} is the symmetrical weight; and σ_i is the standard deviation of Gaussian noise of the visible units. The conditional distribution can be obtained as follows:

$$p(h_j^{(t)} = 1|v) = f\left(b_j + HCon_j + \sum_i v_i^{(t)} w_{ij}\right), \tag{4}$$

$$p(v_i^{(t)}|h) = \mathcal{N}\left(c_i + \mathcal{V}Con_i + \sum_j h_j^{(t)} w_{ij}, 1\right), \tag{5}$$

where $f(.)$ is the sigmoid function, and $\mathcal{N}(\mu, \mathcal{V})$ is a Gaussian distribution. Gradient descent is also used in CRBM learning, and the contrastive divergence (CD) method is used for gradient approximation.

In the CRBM, the impacts of the temporal features $HCon_j$ and $VCon_i$

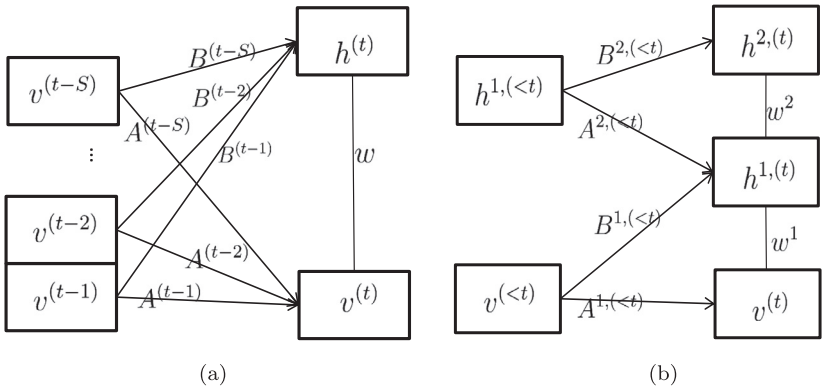

Fig. 1. Architecture of conditional restricted Boltzmann machine (CRBM). (a) One-level model. (b) Two-level model.