受限波尔兹曼机简介

张春霞[1]，姬楠楠[1]，王冠伟[2]

[1] 西安交通大学数学与统计学院，西安 710049

[2] 西安交通大学机械学院，西安 710049

摘要： 受限波尔兹曼机 (Restricted Boltzmann Machines, RBM) 是一类具有两层结构、对称连接且无自反馈的随机神经网络模型, 层间全连接, 层内无连接。近年来, 随着 RBM 的快速学习算法-对比散度的出现, 机器学习界掀起了研究 RBM 理论及应用的热潮。实践表明, RBM 是一种有效的特征提取方法, 用于初始化前馈神经网络可明显提高泛化能力, 堆叠多个 RBM 组成的深度信念网络能提取更抽象的特征。鉴于 RBM 的优点及其广泛应用, 本文对 RBM 的基本模型、学习算法、参数设置、评估方法、变形算法等进行了详细介绍, 最后探讨了 RBM 在未来值得研究的方向。

关键词： 机器学习; 深度学习; 受限波尔兹曼机; 对比散度; Gibbs采样.

中图分类号： TP181(自动推理、机器学习), O235(模式识别理论)

Introduction of Restricted Boltzmann Machines

ZHANG Chun-Xia[1], JI Nan-Nan[1], WANG Guan-Wei[2]

[1] School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710049

[2] School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an 710049

Abstract: A Restricted Boltzmann Machine (RBM) is a particular type of random neural network model which has two-layer architecture, symmetric connections and no self-feedback. The two layers in an RBM are fully connected but there are no connections within the same layer. Recently, with the advent of a fast learning algorithm for RBM (i.e., contrastive divergence), the machine learning community set off a surge to study the theory and applications of RBM since it has many advantages. For example, RBM is an effective method to detect features. When a feed-forward neural network is initialized with an RBM, its generalization capability can be significantly improved. A deep belief network composed of several RBMs can detect more abstract features. Due to the advantages and wide applications of RBM, this paper attempts to provide a started guide for novice. It presents a detailed introduction of basic RBM model, its representative learning algorithm, parametric settings, evaluation methods, its variants and etc. Finally, some research directions of RBM that are deserved to be further studied are discussed.

Key words: Machine learning; Deep learning; Restricted Boltzmann machine; Contrastive

基金项目： 高等学校博士学科点专项科研基金(20100201120048), 数学天元青年基金(11126277), 国家自然科学基金(11201367), 中央高校基本科研业务费专项基金.

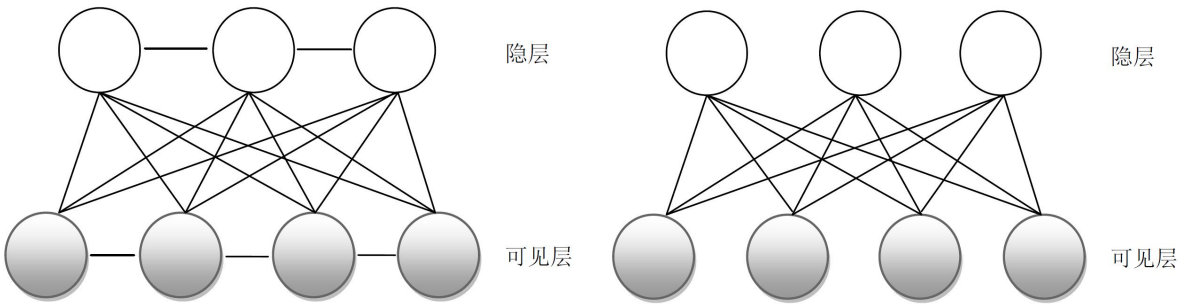
作者简介： 张春霞(1980-), 女, 讲师, 主要研究方向：模式识别、集成学习. 姬楠楠(1985-), 女, 博士生, 主要研究方向: 深度学习、模式识别. 王冠伟(1979-), 男, 博士生, 主要研究方向: 特征提取、聚类分析等。通信作者: 张春霞(基金负责人), E-Mail: cxzhang@mail.xjtu.edu.cn,

divergence; Gibbs sampling.

0 引言

机器学习研究的主要任务是设计和开发计算机可以智能地根据实际数据进行“学习”的算法, 从而使这些算法可以自动地发现隐藏在数据中的模式和规律。目前, 各种机器学习算法在科学研究、工业应用、医学、金融等诸多领域都扮演着非常重要的角色。人工神经网络 (Artificial Neural Network, ANN) [1, 2] 作为一种通过模仿生物神经网络的结构和功能而建立起来的计算模型, 是很具有代表性的一类机器学习方法。ANN 因其自学习、自组织、较好的容错性和优良的非线性逼近能力等优点, 而受到众多领域学者的广泛关注。

在诸多人工神经网络模型中, 波尔兹曼机 (Boltzmann Machine, BM) [3]是 Hinton 和 Sejnowski 于 1986 年提出的一种根植于统计力学的随机神经网络。这种网络中的神经元是随机神经元, 神经元的输出只有两种状态(未激活、激活), 一般用二进制的0和1表示, 状态的取值根据概率统计法则决定。从功能上讲, BM 是由随机神经元全连接组成的反馈神经网络, 且对称连接, 无自反馈, 包含一个可见层和一个隐层的 BM 模型如图1(a)所示。



(a) BM模型的结构. (b) RBM模型的结构.

图 1: BM和RBM模型的结构比较.

BM 具有强大的无监督学习能力, 能够学习数据中复杂的规则。但是, 拥有这种学习能力的代价是其训练(学习)时间非常长。此外, 不仅无法确切地计算 BM 所表示的分布, 甚至得到服从 BM 所表示分布的随机样本也很困难。为克服这一问题, Smolensky [4] 引入了一种限制的波尔兹曼机 (Restricted Boltzmann Machine, RBM)。RBM 具有一个可见层, 一个隐层, 层内无连接, 其结构如图1(b)所示。RBM 具有很好的性质 [5]: 在给定可见层单元状态(输入数据)时, 各隐单元的激活条件独立; 反之, 在给定隐单元状态时, 可见层单元的激活亦条件独立。这样一来, 尽管 RBM 所表示的分布仍无法有效计算, 但通过 Gibbs 采样 (Gibbs sampling) 可以得到服从RBM所表示分布的随机样本。此外, Roux 和 Bengio [6] 从理论上证明, 只要隐单元的数目足够多, RBM能够拟合任意离散分布。自 Hinton [7]于 2002 年提出了 RBM 的快速学习算法-对比散度 (Contrastive Divergence, CD)之后, 机器学习界掀起了一轮研究 RBM、CD 算法的理论及

应用的热潮。理论方面, RBM 的 CD 快速学习算法促进了研究者们对随机近似理论、基于能量的模型、未归一化的统计模型的研究 [8]。应用方面, RBM 目前已被成功地应用于不同的机器学习问题 [9–14], 如分类、回归、降维、高维时间序列建模、图像特征提取、协同过滤等等。

2006年, Hinton 等人 [15] 提出了一种深度信念网络 (Deep Belief Nets, DBN), 并给出了该模型的一个高效学习算法。这个算法成为了其后至今深度学习算法的主要框架。在该算法中, 一个 DBN 模型被视为由若干个RBM堆叠在一起, 训练时可通过由低到高逐层训练这些RBM来实现: (1) 底部 RBM 以原始输入数据训练; (2) 将底部 RBM 抽取的特征作为顶部 RBM 的输入训练; (3) 过程(1)和(2)可以重复来训练所需要的尽可能多的层数。由于 RBM 可以通过 CD 快速训练, 这一框架绕过了直接从整体上训练 DBN 的高复杂度, 从而将其化简为对多个 RBM 的训练问题。Hinton 建议, 经过这种方式训练后, 可以再通过传统的全局学习算法(如反向传播算法)对网络进行微调, 从而使模型收敛到局部最优点。这种学习算法, 本质上等同于先通过逐层 RBM 训练将模型的参数初始化为较优的值, 再通过少量的传统学习算法进一步训练。这样一来, 不仅解决了模型训练速度慢的问题, 大量试验结果也表明, 这种方式能够产生非常好的参数初始值, 从而大大提升了模型的建模能力。自此, 机器学习领域又产生了一个新的研究方向-深度学习 (Deep learning) [16–18], 明确提出了面向人工智能的机器学习算法的设计目标。

当前, 以 RBM 为基本构成模块的 DBN 模型被认为是最有效的深度学习算法之一。鉴于 RBM 在深度学习领域中占据的核心位置以及其本身的良好性质, 为了给 RBM 的初学者提供入门指导, 同时为设计与之相关的新算法提供参考, 本文将对 RBM 进行较为系统的介绍, 详细阐述其基本模型、具有代表性的快速学习算法、参数设置、评估方法及其变形算法, 最后对 RBM 在未来值得研究的方向进行探讨。

本文后续内容安排如下: 第1节介绍受限波尔兹曼机 RBM 的基本模型, 第2节详细阐述当前训练 RBM 的快速学习算法, 第3节讨论 RBM 的参数设置, 第4节给出评价 RBM 优劣的方法, 第5节简单介绍几种具有代表性的 RBM 变形算法, 第6节是总结与展望, 主要探讨 RBM 在未来值得研究的方向。

1 受限波尔兹曼机RBM的基本模型

RBM 也可以被视为一个无向图 (undirected graph) 模型, 如图2所示。$\mathbf{v}$ 为可见层, 用于表示观测数据, $\mathbf{h}$ 为隐层, 可视为一些特征提取器 (feature detectors), W 为两层之间的连接权重。Welling [19]指出, RBM 中的隐单元和可见单元可以为任意的指数族单元(即给定隐单元(可见单元), 可见单元(隐单元)的分布可以为任意的指数族分布), 如 softmax 单元、高斯单元、泊松单元等等。这里, 为了讨论方便起见, 我们假设所有的可见单元和隐单元均为二值变量, 即 $\forall i, j, v_i \in \{0, 1\}, h_j \in \{0, 1\}$。

如果一个 RBM 有n个可见单元和m个隐单元, 用向量$\mathbf{v}$和$\mathbf{h}$分别表示可见单元和隐单元的状态。其中, v_i表示第i个可见单元的状态, h_j表示第j个隐单元的状态。那么, 对于一组给定的状

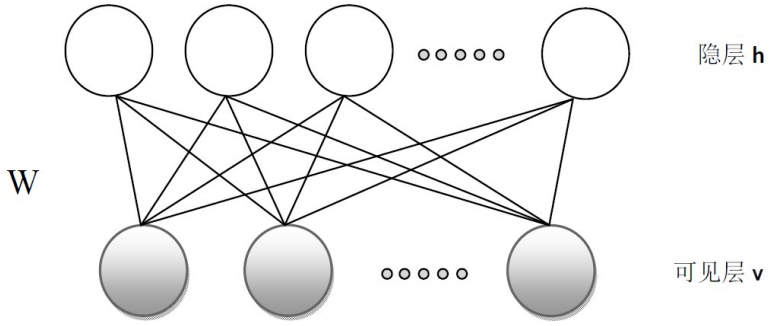


图 2: RBM的图模型表示, 层内单元之间无连接.

态$(\mathbf{v},\mathbf{h})$, RBM作为一个系统所具备的能量定义为

$$E(\mathbf{v},\mathbf{h}|\boldsymbol{\theta}) = -\sum_{i=1}^{n} a_i v_i - \sum_{j=1}^{m} b_j h_j - \sum_{i=1}^{n}\sum_{j=1}^{m} v_i W_{ij} h_j, \tag{1}$$

上式中, $\boldsymbol{\theta} = \{W_{ij}, a_i, b_j\}$是 RBM 的参数, 它们均为实数。其中, W_{ij}表示可见单元i与隐单元j之间的连接权重, a_i表示可见单元i的偏置(bias), b_j表示隐单元j的偏置。当参数确定时, 基于该能量函数, 我们可以得到$(\mathbf{v},\mathbf{h})$的联合概率分布,

$$P(\mathbf{v},\mathbf{h}|\boldsymbol{\theta}) = \frac{e^{-E(\mathbf{v},\mathbf{h}|\boldsymbol{\theta})}}{Z(\boldsymbol{\theta})}, \quad Z(\boldsymbol{\theta}) = \sum_{\mathbf{v},\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h}|\boldsymbol{\theta})} \tag{2}$$

其中$Z(\boldsymbol{\theta})$为归一化因子(也称为配分函数, partition function)。

对于一个实际问题, 我们最关心的是由RBM所定义的关于观测数据$\mathbf{v}$的分布$P(\mathbf{v}|\boldsymbol{\theta})$, 即联合概率分布$P(\mathbf{v},\mathbf{h}|\boldsymbol{\theta})$的边际分布, 也称为似然函数(likelihood function),

$$P(\mathbf{v}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h}|\boldsymbol{\theta})}. \tag{3}$$

为了确定该分布, 需要计算归一化因子$Z(\boldsymbol{\theta})$, 这需要2^{n+m}次计算。因此, 即使通过训练可以得到模型的参数W_{ij}, a_i和b_j，我们仍旧无法有效地计算由这些参数所确定的分布。

但是, 由RBM的特殊结构(即层间有连接, 层内无连接)可知: 当给定可见单元的状态时, 各隐单元的激活状态之间是条件独立的。此时, 第j个隐单元的激活概率为

$$P(h_j = 1|\mathbf{v},\boldsymbol{\theta}) = \sigma(b_j + \sum_i v_i W_{ij}). \tag{4}$$

其中, $\sigma(x) = \frac{1}{1+\exp(-x)}$为sigmoid激活函数。

由于RBM的结构是对称的, 当给定隐单元的状态时, 各可见单元的激活状态之间也是条件独立的, 即第i个可见单元的激活概率为

$$P(v_i = 1|\mathbf{h},\boldsymbol{\theta}) = \sigma(a_i + \sum_j W_{ij} h_j). \tag{5}$$

2 基于对比散度的RBM快速学习算法

学习RBM的任务是求出参数$\boldsymbol{\theta}$的值, 以拟合给定的训练数据。参数$\boldsymbol{\theta}$可以通过最大化RBM在训练集(假设包含T个样本)上的对数似然函数学习得到, 即

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \sum_{t=1}^{T} \log P(\mathbf{v}^{(t)}|\boldsymbol{\theta}). \tag{6}$$

为了获得最优参数$\boldsymbol{\theta}^*$, 我们可以使用随机梯度上升法(stochastic gradient ascent)求$\mathcal{L}(\boldsymbol{\theta}) = \sum_{t=1}^{T} \log P(\mathbf{v}^{(t)}|\boldsymbol{\theta})$的最大值。其中, 关键步骤是计算$\log P(\mathbf{v}^{(t)}|\boldsymbol{\theta})$关于各个模型参数的偏导数。

由于

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &= \sum_{t=1}^{T} \log P(\mathbf{v}^{(t)}|\boldsymbol{\theta}) = \sum_{t=1}^{T} \log \sum_{\mathbf{h}} P(\mathbf{v}^{(t)}, \mathbf{h}|\boldsymbol{\theta}) \\
&= \sum_{t=1}^{T} \log \frac{\sum_{\mathbf{h}} \exp[-E(\mathbf{v}^{(t)}, \mathbf{h}|\boldsymbol{\theta})]}{\sum_{\mathbf{v}} \sum_{\mathbf{h}} \exp[-E(\mathbf{v}, \mathbf{h})|\boldsymbol{\theta}]} \\
&= \sum_{t=1}^{T} \left(\log \sum_{\mathbf{h}} \exp[-E(\mathbf{v}^{(t)}, \mathbf{h}|\boldsymbol{\theta})] - \log \sum_{\mathbf{v}} \sum_{\mathbf{h}} \exp[-E(\mathbf{v}, \mathbf{h})|\boldsymbol{\theta}] \right),
\end{aligned} \tag{7}$$

令θ表示$\boldsymbol{\theta}$中的某一个参数, 则对数似然函数关于θ的梯度为

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \theta} &= \sum_{t=1}^{T} \frac{\partial}{\partial \theta} \left(\log \sum_{\mathbf{h}} \exp[-E(\mathbf{v}^{(t)}, \mathbf{h}|\boldsymbol{\theta})] - \log \sum_{\mathbf{v}} \sum_{\mathbf{h}} \exp[-E(\mathbf{v}, \mathbf{h}|\boldsymbol{\theta})] \right) \\
&= \sum_{t=1}^{T} \left(\sum_{\mathbf{h}} \frac{\exp[-E(\mathbf{v}^{(t)}, \mathbf{h}|\boldsymbol{\theta})]}{\sum_{\mathbf{h}} \exp[-E(\mathbf{v}^{(t)}, \mathbf{h}|\boldsymbol{\theta})]} \times \frac{\partial(-E(\mathbf{v}^{(t)}, \mathbf{h}|\boldsymbol{\theta}))}{\partial \theta} \right. \\
&\quad \left. - \sum_{\mathbf{v}} \sum_{\mathbf{h}} \frac{\exp[-E(\mathbf{v}, \mathbf{h}|\boldsymbol{\theta})]}{\sum_{\mathbf{v}} \sum_{\mathbf{h}} \exp[-E(\mathbf{v}, \mathbf{h}|\boldsymbol{\theta})]} \times \frac{\partial(-E(\mathbf{v}, \mathbf{h}|\boldsymbol{\theta}))}{\partial \theta} \right) \\
&= \sum_{t=1}^{T} \left(\left\langle \frac{\partial(-E(\mathbf{v}^{(t)}, \mathbf{h}|\boldsymbol{\theta}))}{\partial \theta} \right\rangle_{P(\mathbf{h}|\mathbf{v}^{(t)}, \boldsymbol{\theta})} - \left\langle \frac{\partial(-E(\mathbf{v}, \mathbf{h}|\boldsymbol{\theta}))}{\partial \theta} \right\rangle_{P(\mathbf{v}, \mathbf{h}|\boldsymbol{\theta})} \right)
\end{aligned} \tag{8}$$

其中, $\langle \cdot \rangle_P$ 表示求关于分布 P 的数学期望。$P(\mathbf{h}|\mathbf{v}^{(t)}, \boldsymbol{\theta})$ 表示在可见单元限定为已知的训练样本 $\mathbf{v}^{(t)}$ 时, 隐层的概率分布, 故式(8)中的前一项比较容易计算。$P(\mathbf{v}, \mathbf{h}|\boldsymbol{\theta})$ 表示可见单元与隐单元的联合分布, 由于归一化因子 $Z(\boldsymbol{\theta})$ 的存在, 该分布很难获取, 导致我们无法直接计算式(8)中的第二项, 只能通过一些采样方法(如 Gibbs 采样)获取其近似值。值得指出的是, 在最大化似然函数的过程中, 为了加快计算速度, 上述偏导数在每一迭代步中的计算一般只基于部分而非所有的训练样本进行, 关于这部分内容我们将在后面讨论 RBM 的参数设置时详细阐述。

下面, 假设只有一个训练样本, 我们分别用 “data” 和 “model” 来简记 $P(\mathbf{h}|\mathbf{v}^{(t)}, \boldsymbol{\theta})$ 和 $P(\mathbf{v}, \mathbf{h}|\boldsymbol{\theta})$ 这两个概率分布, 则对数似然函数关于连接权重 W_{ij}、可见层单元的偏置 a_i 和隐层单

元的偏置 b_j 的偏导数分别为

$$\frac{\partial \log P(\mathbf{v}|\boldsymbol{\theta})}{\partial W_{ij}} = \langle v_i h_j \rangle_{\text{data}} - \langle v_i h_j \rangle_{\text{model}},$$
$$\frac{\partial \log P(\mathbf{v}|\boldsymbol{\theta})}{\partial a_i} = \langle v_i \rangle_{\text{data}} - \langle v_i \rangle_{\text{model}},$$
$$\frac{\partial \log P(\mathbf{v}|\boldsymbol{\theta})}{\partial b_j} = \langle h_j \rangle_{\text{data}} - \langle h_j \rangle_{\text{model}}.$$

2.1 RBM中的Gibbs采样

Gibbs采样(Gibbs sampling) [20]是一种基于马尔可夫链蒙特卡罗(Markov Chain Monte Carlo, MCMC)策略的采样方法。对于一个K维随机向量$\mathbf{X} = (X_1, X_2, \cdots, X_K)$, 假设我们无法求得关于$\mathbf{X}$的联合分布$P(\mathbf{X})$, 但我们知道给定$\mathbf{X}$的其他分量时, 其第$k$个分量$X_k$的条件分布, 即$P(X_k|X_{k^-})$, $X_{k^-} = (X_1, X_2, \cdots, X_{k-1}, X_{k+1}, \cdots, X_K)$。那么, 我们可以从$\mathbf{X}$的一个任意状态(比如$[x_1(0), x_2(0), \cdots, x_K(0)]$)开始, 利用上述条件分布, 迭代地对其分量依次采样, 随着采样次数的增加, 随机变量$[x_1(n), x_2(n), \cdots, x_K(n)]$的概率分布将以$n$的几何级数的速度收敛于$\mathbf{X}$的联合概率分布$P(\mathbf{X})$。换句话说, 我们可以在未知联合概率分布$P(\mathbf{X})$的条件下对其进行采样。

基于RBM模型的对称结构, 以及其中神经元状态的条件独立性, 我们可以使用Gibbs采样方法得到服从RBM定义的分布的随机样本。在RBM中进行k步吉布斯采样的具体算法为: 用一个训练样本(或可见层的任何随机化状态)初始化可见层的状态$\mathbf{v}_0$, 交替进行如下采样:

$$\begin{aligned} \mathbf{h}_0 &\sim P(\mathbf{h}|\mathbf{v}_0), \qquad & \mathbf{v}_1 &\sim P(\mathbf{v}|\mathbf{h}_0), \\ \mathbf{h}_1 &\sim P(\mathbf{h}|\mathbf{v}_1), \qquad & \mathbf{v}_2 &\sim P(\mathbf{v}|\mathbf{h}_1), \\ &\cdots\cdots, \qquad & \mathbf{v}_{k+1} &\sim P(\mathbf{v}|\mathbf{h}_k). \end{aligned}$$

在采样步数k足够大的情况下, 我们可以得到服从RBM所定义的分布的样本。此外, 使用Gibbs采样我们也可以得到式(8)中第二项的一个近似。

2.2 基于对比散度的快速学习算法

尽管利用吉布斯采样我们可以得到对数似然函数关于未知参数梯度的近似, 但通常情况下需要使用较大的采样步数, 这使得RBM的训练效率仍旧不高, 尤其是当观测数据的特征维数较高时。

2002年, Hinton [7]提出了RBM的一个快速学习算法, 即对比散度(Contrastive Divergence, CD)。与吉布斯采样不同, Hinton指出当使用训练数据初始化$\mathbf{v}_0$时, 我们仅需要使用k(通常$k = 1$)步吉布斯采样便可以得到足够好的近似。在CD算法一开始, 可见单元的状态被设置成一个训练样本, 并利用式(4)计算所有隐层单元的二值状态。在所有隐层单元的状态确定之后, 根据式(5)来确定第i个可见单元v_i取值为1的概率, 进而产生可见层的一个重构(reconstruction)。

这样, 在使用随机梯度上升法最大化对数似然函数在训练数据上的值时, 各参数的更新准则为

$$\Delta W_{ij} = \epsilon(\langle v_i h_j \rangle_{\text{data}} - \langle v_i h_j \rangle_{\text{recon}}),$$
$$\Delta a_i = \epsilon(\langle v_i \rangle_{\text{data}} - \langle v_i \rangle_{\text{recon}}),$$
$$\Delta b_j = \epsilon(\langle h_j \rangle_{\text{data}} - \langle h_j \rangle_{\text{recon}}),$$

这里, ϵ是学习率(learning rate), $\langle \cdot \rangle_{\text{recon}}$表示一步重构后模型定义的分布。

在RBM中, 可见单元数一般等于训练数据的特征维数, 而隐单元数需要事先给定。为了与前文记号一致, 假设可见单元数和隐单元数分别为n和m。令W表示可见层与隐层间的连接权重矩阵($m \times n$阶), $\mathbf{a}$(n维列向量)和$\mathbf{b}$(m维列向量)分别表示可见层与隐层的偏置向量。RBM的基于CD的快速学习算法主要步骤可描述如下:

- **输入**: 一个训练样本$\mathbf{x}_0$; 隐层单元个数m; 学习率ϵ; 最大训练周期T.
- **输出**: 连接权重矩阵W、可见层的偏置向量$\mathbf{a}$、隐层的偏置向量$\mathbf{b}$.
- **训练阶段**:
  初始化: 令可见层单元的初始状态$\mathbf{v}_1 = \mathbf{x}_0$; W、$\mathbf{a}$和$\mathbf{b}$为随机的较小数值。
  For $t = 1, 2, \cdots, T$

  For $j = 1, 2, \cdots, m$(对所有隐单元)

  计算$P(\mathbf{h}_{1j} = 1|\mathbf{v}_1)$, 即$P(\mathbf{h}_{1j} = 1|\mathbf{v}_1) = \sigma(b_j + \sum_i v_{1i} W_{ij})$;
  从条件分布$P(\mathbf{h}_{1j}|\mathbf{v}_1)$中抽取$\mathbf{h}_{1j} \in \{0, 1\}$.

  EndFor

  For $i = 1, 2, \cdots, n$(对所有可见单元)

  计算$P(\mathbf{v}_{2i} = 1|\mathbf{h}_1)$, 即$P(\mathbf{v}_{2i} = 1|\mathbf{h}_1) = \sigma(a_i + \sum_j W_{ij} h_{1j})$;
  从条件分布$P(\mathbf{v}_{2i}|\mathbf{h}_1)$中抽取$\mathbf{v}_{2i} \in \{0, 1\}$.

  EndFor

  For $j = 1, 2, \cdots, m$(对所有隐单元)

  计算$P(\mathbf{h}_{2j} = 1|\mathbf{v}_2)$, 即$P(\mathbf{h}_{2j} = 1|\mathbf{v}_2) = \sigma(b_j + \sum_i v_{2i} W_{ij})$;

  EndFor

  按下式更新各个参数

  - $W \leftarrow W + \epsilon(P(\mathbf{h}_{1\cdot} = 1|\mathbf{v}_1)\mathbf{v}_1^T - P(\mathbf{h}_{2\cdot} = 1|\mathbf{v}_2)\mathbf{v}_2^T)$;
  - $\mathbf{a} \leftarrow \mathbf{a} + \epsilon(\mathbf{v}_1 - \mathbf{v}_2)$;
  - $\mathbf{b} \leftarrow \mathbf{b} + \epsilon(P(\mathbf{h}_{1\cdot} = 1|\mathbf{v}_1) - P(\mathbf{h}_{2\cdot} = 1)|\mathbf{v}_2)$;

EndFor

算法1. RBM的基于CD的快速学习算法主要步骤.

在上述算法中, 记号$P(\mathbf{h}_{k\cdot}=1|\mathbf{v}_k)$ $(k=1,2)$是m维列向量, 其第j个元素为$P(\mathbf{h}_{kj}=1|\mathbf{v}_k)$。

尽管上述基于CD的学习算法是针对RBM的可见单元和隐层单元均为二值变量的情形提出的, 但很容易推广到可见层单元为高斯变量、可见层和隐层单元均为高斯变量等其他情形, 关于这方面的研究具体可参见 [21–25]。

此外, 还有一些研究者在CD算法的基础上, 对其作了进一步改进。例如, Tieleman [26]提出了持续对比散度(Persistent Contrastive Divergence, PCD)算法, 该算法与CD的区别在于: 首先, PCD不再使用训练数据初始化CD算法中的Gibbs采样的马氏链; 其次, PCD算法中的学习率较小且不断衰减。根据随机近似理论, 尽管每次更新参数后模型都发生了改变(每次对于W, $\mathbf{a}$和$\mathbf{b}$的更新, RBM定义的分布都会发生改变), 但由于学习率较小且不断衰减, 则可认为那条马氏链产生的负样本是由当前RBM定义的分布的一个近似分布采样而来。Tieleman和Hinton [27]进一步改进了PCD算法, 他们通过引入一组辅助参数以加快PCD中的马氏链的混合率, 提出了快速持续对比散度(Fast Persistent Contrastive Divergence, FPCD)算法。关于RBM的学习算法, 除了上述提到的基于CD的一些方法之外, 还有最大化拟似然函数(maximum pseudo-likelihood)、比率匹配方法(ratio matching)等, 有兴趣的读者可参阅 [28]查找关于RBM学习算法比较详细的阐述。

3 RBM的参数设置

RBM 的训练通常是基于 CD 的方法(即算法1)进行的, 但如何设置其中的一些参数(如隐单元个数、学习率、参数的初始值等), 是需要有一定经验的。近来, 已有部分研究结果 [29, 30]表明: 对于特定的数据集和 RBM 结构, 如果参数设置不合适, RBM 将很难对真正的数据分布正确建模。因此, 对实际使用者(尤其是初学者)来说, 了解 RBM 中参数设置的一般规则是非常重要的。根据 Hinton [23] 提供的建议以及我们进行数值试验所获部分经验, 对 RBM 中的参数设置可参考以下规则。

小批量数据及其容量 对于连接权重、可见层和隐层偏置的更新, 虽然可以基于一个训练样本进行(类似于在线学习的方式), 但计算量将很大。将训练集事先分成包含几十或几百个样本的小批量数据 (mini-batches) 进行计算将更高效, 这主要是可以利用图形处理器 GPU(Graphic Processing Unit) 或 Matlab 中矩阵之间相乘运算的优势。同时, 为了避免在小批量数据的样本容量发生改变时, 学习率也必须做相应的修改, 通常的做法是在参数的更新过程中, 使用参数的平均梯度(即总梯度除以数据容量), 即

$$\theta^{(t+1)}=\theta^{(t)}+\epsilon\left(\frac{1}{B}\sum_{t'=Bt+1}^{B(t+1)}\frac{\partial\log P(\mathbf{v}^{(t')}|\boldsymbol{\theta})}{\partial\theta}\right),$$

这里, B 表示小批量数据的容量, 其值不应设得太大。$B=1$ 表示参数更新以在线学习的方式进行, 而$B=T$则表示传统的批处理方式。一般而言, 若训练集是包含来自不同类(具有同等概

率)的样本, 理想的 B 应为总类数, 使得每批数据中都包含来自每个类的一个样本, 以减小梯度估计的抽样误差。对于其他数据集, 则可先随机化训练样本的次序, 再将其分为容量为10的倍数的小批量数据。

学习率 学习率若过大, 将导致重构误差急剧增加, 权重也会变得异常大。设置学习率的一般做法是先做权重更新和权重的直方图, 令权重更新量为权重的 10^{-3} 倍左右。如果有一个单元的输入值很大, 则权重更新应再小一些, 因为同一方向上较多小的波动很容易改变梯度的符号。相反地, 对于偏置, 其权重更新可以大一些。

权重和偏置的初始值 一般地, 连接权重 W_{ij} 可初始化为来自正态分布 $N(0, 0.01)$ 的随机数, 隐单元的偏置 b_j 初始化为0。对于第 i 个可见单元, 其偏置 a_i 通常初始化为 $\log[p_i/(1-p_i)]$, 其中 p_i 表示训练样本中第 i 个特征处于激活状态所占的比率。如果不这样做, 在学习的早期阶段, RBM 会利用隐单元使得第 i 个特征以概率 p_i 处于激活状态。

动量学习率 学习率ϵ的选择至关重要, ϵ大收敛速度快, 但过大可能引起算法不稳定; ϵ小可避免不稳定情况的出现, 但收敛速度较慢。为克服这一矛盾, 一种具有代表性的思想是在参数更新式中增加动量项(momentum), 使本次参数值修改的方向不完全由当前样本下的似然函数梯度方向决定, 而采用上一次参数值修改方向与本次梯度方向的组合。在某些情况下, 这可以避免算法过早地收敛到局部最优点。以连接权重参数W_{ij}为例, 其更新公式为

$$W_{ij}^{(t+1)} = kW_{ij}^{(t)} + \epsilon \frac{\partial \mathcal{L}}{\partial W_{ij}^{(t)}},$$

其中k为动量项学习率。开始时, k可设为0.5, 在重构误差处于平稳增加状态时, k可取为0.9。

权衰减 使用权衰减(weight-decay)策略的主要目的是避免学习过程出现过拟合 (overfitting)现象, 一般做法是在正常的梯度项后额外增加一项, 以对较大的参数值作出惩罚。最简单的罚函数是 L_2 函数 $(\lambda/2)\sum_i \sum_j W_{ij}^2$, 即所有权重参数的平方和的1/2再乘上一个正则化系数 λ, λ在RBM中又称为权损失 (weight-cost)。重要的是, 惩罚项关于权重参数的梯度必须乘上学习率, 否则, 学习率的改变将导致优化的目标函数也发生改变。在 RBM 中, 若使用 L_2 罚函数, 则权损失系数的取值可以取介于0.01与0.0001之间的任意值。值得指出的是, 权衰减策略只需应用于连接权重参数 W_{ij} 上, 可见层和隐层偏置不需使用, 因为它们不大可能导致过拟合。并且在某些情况下, 偏置的值还必须较大才行。

隐单元个数 如果我们关心的主要目标是避免过拟合而不是计算复杂度, 则可以先估算一下用一个好的模型描述一个数据所需的比特数, 用其乘上训练集容量。基于所得的数, 选择比其低一个数量级的值作为隐元个数。如果训练数据是高度冗余的(比如数据集容量非常大), 则可以使用更少一些的隐元。

以上讨论的是RBM中的一些常用的参数设置, 针对一个实际问题, 应使用什么类型的可见单元和隐单元, 在其中如何加入稀疏性使得隐单元只在少数情况下处于激活状态等问题的讨论, 可参见文 [23, 31]。

4 RBM的评估算法

对于一个已经学习得到或正在学习中的 RBM, 应通过何种指标评价其优劣呢?

显然, 最简单的指标就是该RBM在训练数据上的似然度$\mathcal{L}(\boldsymbol{\theta}) = \sum_{t=1}^{T} \log P(\mathbf{v}^{(t)}|\boldsymbol{\theta})$。但是, $\mathcal{L}(\boldsymbol{\theta})$的计算涉及到归一化常数$Z(\boldsymbol{\theta})$, 而这个值是无法通过数学方法直接解析得到的, 但我们又不可能枚举RBM的所有状态。因此, 只能采用近似方法对RBM进行评估。

4.1 重构误差

所谓"重构误差"(reconstruction error), 就是以训练数据作为初始状态, 根据RBM的分布进行一次Gibbs采样后所获样本与原数据的差异(一般用一范数或二范数来评估)。

$Error = 0$ %初始化误差

for all $\mathbf{v}^{(t)},\ t \in \{1, 2, \cdots, T\}$ **do** %对每个训练样本$\mathbf{v}^{(t)}$进行以下计算

  $\mathbf{h} \sim P(\cdot|\mathbf{v}^{(t)})$ %对隐层采样

  $\mathbf{v} \sim P(\cdot|\mathbf{h})$ %对可见层采样

  $Error = Error + \| \mathbf{v} - \mathbf{v}^{(t)} \|$ %累计当前误差

end for

return $Error$ %返回总误差

算法2. 重构误差的计算.

重构误差能够在一定程度上反映 RBM 对训练数据的似然度, 不过并不完全可靠 [23]。但总的来说, 重构误差的计算十分简单, 因此在实践中非常有用。

4.2 退火式重要性采样

"退火式重要性采样"(Annealed Importance Sampling, AIS) [32]是目前比较主流的RBM评估方法。它的想法非常直接, 就是利用蒙特卡罗方法估计RBM对数据的似然度。只不过没有使用MCMC, 而是通过一种叫做"重要性采样"(Importance Sampling) [20]的算法进行逼近。这种算法的优点在于: 当目标分布十分陡峭时, 不直接对其进行采样, 而是引入另一个简单的分布, 在这个简单的分布上采样。然后, 利用采样所获样本和两个分布之间的关系对原分布上的均值进行估算。

"重要性抽样"的基本思想如下: 假设我们要计算某个分布$P_A(x)$的归一化常数Z_A, 那么, 我们可以引入另一个状态空间相同, 但更容易采样的分布$P_B(x)$, 并且事先知道它的归一化常数Z_B。这时, 只要能计算出Z_A/Z_B的值, 我们就可以算出原分布的归一化常数Z_A。假

设$Z_A = \sum_x f(x)$, $Z_B = \sum_x g(x)$, 考虑它们的比例

$$\frac{Z_A}{Z_B} = \frac{\sum_x f(x)}{\sum_x g(x)} = \sum_x \frac{g(x)}{\sum_x g(x)} \frac{f(x)}{g(x)} = \left\langle \frac{f(x)}{g(x)} \right\rangle_{P_B},$$

上式表明Z_A/Z_B最终等同于函数$f(x)/g(x)$在引入的辅助分布$P_B(x)$上的均值。由于辅助分布上采样较容易, 这就绕过了传统 MCMC 可能面临的多模式问题。不过, 如果两个分布的差别很大, 这个估计值的偏差就会很高, 导致估算的结果很不准确。AIS 的想法是, 在两个分布中间进一步引入大量的中间分布, 使得相邻的两个分布总是十分相似的, 如此一来, 就克服了分布差别过大时造成的高偏差问题。

在评估RBM时, 我们可以引入一个非常简单的RBM, 使其归一化常数可以直接计算出来。然后, 利用 AIS, 估算两个 RBM 的归一化常数之比, 最后将这个比例乘上简单 RBM 的归一化常数, 即得到被评估 RBM 的归一化常数。从而, RBM 对训练数据的似然度就可以顺利算出了。

5 基本RBM模型的变形算法

自 RBM 的基本模型提出以来, 尤其是 Hinton 提出基于 CD 的快速学习算法之后, 研究者们针对 RBM 已发展了很多变形算法(如稀疏RBM, 稀疏组RBM, 分类RBM, 条件RBM等)[8, 12, 13, 17, 28, 33–36]。由于文章篇幅限制, 本节将对几种具有代表性的算法作一简单介绍。

稀疏受限波尔兹曼机(Sparse Restricted Boltzmann Machine, SRBM) 由于稀疏表示(sparse representation)模型符合生物视觉系统特性, 且能够提取图像的高级特征, 近年来在机器学习、图像处理、压缩感知等研究领域都得到了广泛关注。一般而言, RBM学习到的特征表示是分布式、非稀疏的。在实际应用中, 隐单元只在少数情况下处于激活状态更容易解释(相应隐单元仅被用来表示很小一部分训练数据), 且判别性能在某些情况下还会得到改进。Lee [33]在对数似然函数的基础上, 引入了一个稀疏惩罚项, 以惩罚隐单元的平均激活概率偏离给定水平p所引起的损失, 提出了一种稀疏受限波尔兹曼机(Sparse Restricted Boltzmann Machines, SRBM)。给定训练数据$\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \cdots, \mathbf{v}^{(T)}$, SRBM的目标函数为

$$\underset{\{\omega_{ij}, a_i, b_j\}}{\text{minimize}} \quad -\sum_{t=1}^{T} \log \sum_{\mathbf{h}} P(\mathbf{v}^{(t)}, \mathbf{h}^{(t)}) + \lambda \sum_{j=1}^{m} \left| p - \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[h_j^{(t)} | \mathbf{v}^{(t)}] \right|^2,$$

这里, $\mathbb{E}[\cdot]$表示数据已知时的条件期望, λ是正则化系数, p是控制隐单元稀疏度的常数(需事先指定)。在学习过程中, 可先基于对比散度的学习算法给出对数似然函数的梯度近似, 再利用正则化项的梯度进行梯度下降, 直至算法收敛。在MNIST手写体数据集和自然图像上的试验结果表明, 稀疏RBM可以提取手写体的笔划特征及自然图像中类似于Gabor滤波的特征, 这与人脑$V1$区简单细胞感受野(receptive fields)十分相似。更重要地, 堆叠两个稀疏RBM可以提取更抽象的特征。文中试验结果表明, 对于自然图像, 两个稀疏RBM可以提取轮廓(contours)、拐角(angles)以及边缘合并(junctions of edges)等特征, 这些特征与人脑V2区细胞的感受野十分相

似。较之前的稀疏表示方法, 堆积稀疏RBM不但可以提取类似于V1区简单细胞的感受野, 而且能够提取类似于V2区细胞的感受野, 这促进了机器学习向人工智能的迈进。

稀疏组受限波尔兹曼机(Sparse Group Restricted Boltzmann Machine, SGRBM)在实际问题中, 特征之间往往显示很强的统计相关性, 并且很多特征经常成组地出现(或成组地不出现)。由于直接学习所有隐单元之间的统计相关性是困难的, 尤其是在高维问题建模时。为简化该问题, 罗恒 [13]将组稀疏(sparse group)方法应用到RBM中, 提出了稀疏组受限波尔兹曼机。其主要思想如下: 首先, 将隐单元划分到不重叠的组中, 只考虑组内隐单元状态的相关性; 其次, 不是去"学习"组内隐单元状态的相关性, 而是通过正则化方法惩罚组内隐单元的总体激活程度, 从而使组内隐单元在学习过超中不再是条件独立的。具体地, 给定训练数据, SGRBM 在似然函数中引入了一个惩罚项, 该项是关于隐单元激活概率的混合范数(L_1/L_2范数)。假设一个RBM中有m个隐单元, $\mathcal{H}$表示隐单元下标组成的指标集, 即$\mathcal{H}=\{1,2,\cdots,m\}$。将这些隐单元分成$K$组, 假设所有的组大小相同, 并且互不重叠, 记第k组的指标集为$\mathcal{G}_k$, $\mathcal{G}_k\subset\mathcal{H}$, $k=1,2,\cdots,K$。给定分组$\mathcal{G}=\{\mathcal{G}_1,\mathcal{G}_2,\cdots,\mathcal{G}_K\}$和观测数据$\mathbf{v}^{(l)}$, 第$k$组隐单元激活概率的$L_2$范数定义为

$$N_k(\mathbf{v}^{(l)})=\left(\sum_{m\in\mathcal{G}_k}P(h_m=1|\mathbf{v}^{(l)})^2\right)^{\frac{1}{2}},$$

该范数可解释为第k组隐单元在给定观测数据$\mathbf{v}^{(l)}$时的总体激活程度。给定所有隐单元组的范数, 相应的L_1/L_2范数定义为

$$\sum_{k=1}^{K}|N_k(\mathbf{v}^{(l)})|=\sum_{k=1}^{K}\left(\sum_{m\in\mathcal{G}_k}P(h_m=1|\mathbf{v}^{(l)})^2\right)^{\frac{1}{2}}.$$

于是, SGRBM的目标函数

$$\underset{\{\omega_{ij},a_i,b_j\}}{\text{minimize}}\quad -\sum_{t=1}^{T}\log\sum_{\mathbf{h}}P(\mathbf{v}^{(t)})+\lambda\sum_{k=1}^{K}|N_k(\mathbf{v}^{(l)})|,$$

其中, λ是正则化常数。为了学习得到模型参数, 罗恒建议采用如下迭代算法: 基于给定训练数据, 首先应用对比散度更新模型参数一次, 再使用正则化项的梯度更新参数一次, 直至算法收敛。

L_1/L_2正则化的作用可以从组间和组内两个层面来解释。在组间(L_1范数), 为了最小化混合范数, 给定训练数据, 正则化项会鼓励隐单元的激活概率形成一种组稀疏的表示, 也就是大量隐单元组的L_2范数为0。由于隐单元的激活概率是非负的, 隐单元组的L_2范数为0便意味着组内所有隐单元的激活概率均为0。在组内(L_2范数), 通常认为各分量均会受到同等程度的惩罚, 故而不会产生组内的稀疏表示。但是由于RBM 隐单元的激活概率的函数形式, 仍旧会产生一种组内的稀疏表示。与SRBM相比, SGRBM可以学习到更局部化的特征。此外, 将SGRBM用于初始化深层神经网络, 在MNIST和OCR英文字母数据集上也达到了更高的识别率。

分类受限波尔兹曼机(Classification Restricted Boltzmann Machine, ClassRBM)当使用RBM解决分类任务时, 最常见的做法是将RBM视为一个特征提取器(feature detector): 使用观测数

据(忽略类标签)训练RBM, 然后以原训练数据在训练好的RBM的隐单元激活概率以及原有的类标签组成新的训练集, 进而使用其他常用的分类算法训练分类器。由于RBM是采用无监督学习的方式训练的, 学习到的特征并不完全适合分类任务。Larochelle等人 [34, 35]指出RBM可直接用于解决有监督学习任务, 并提出了分类受限波尔兹曼机, 其主要思想是利用包含二值随机变量的隐单元来拟合输入特征与类标签的联合分布。Larochelle等人提出, ClassRBM可以使用三种不同的训练目标进行学习, 其中的参数仍然可以基于对比散度进行训练。ClassRBM使得分类过程得以简化(不需要再训练另外的分类器), 保证了学习到的特征的判别能力, 并且可以以在线学习的方式进行训练, 可实时监测其学习到的特征表示的判别性能。传统RBM的另一个用处是初始化深层神经网络, ClassRBM与其相比, 省去了第二个训练阶段。

6 总结与展望

对比散度较好地解决了 RBM 的学习效率问题, 使得近些年 RBM 在许多领域都得到了广泛研究和应用。本文对 RBM 的基本模型、基于对比散度的快速学习算法、参数设置、评估方法及其具有代表性的几种变形算法作了较为详细的介绍。RBM 为人们解决智能问题提供了一种强有力的工具, 并为其他领域的研究提供了新技术和新思路, 研究前景广阔。尤其是随着深度神经网络的兴起, 借助 RBM 来学习深层网络逐渐成为深度学习研究中的主流, 也使得 RBM 在深度学习领域中逐渐占据核心地位。然而, 在 RBM 及其相关理论和学习算法的研究中, 仍有许多问题值得我们作进一步探讨。例如, 如何提高 RBM 在无监督学习场景下所提取特征的辨别能力？在不增加隐单元个数的情况下, 只利用 RBM 能量函数的非参数化形式能否提高其逼近性能？RBM 能否用于图像分割、高维数据的聚类、缺失数据的重构等更广泛的实际应用？这些问题的研究和探讨都将具有十分重要的理论和实际意义。

参考文献（References）

[1] 叶世伟, 史忠植. 神经网络原理[M]. 北京：机械工业出版社, 2006.

[2] Haykin S. 神经网络与机器学习(英文版, 第3版)[M]. 北京：机械工业出版社, 2009.

[3] Hinton G E, Sejnowski T J. Learning and relearning in Boltzmann machines [A]. In: Rumelhart D E, McClelland J L (eds.). *Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 1: Foundations*. Cambridge, MA: MIT Press, 1986.

[4] Smolensky P. Information processing in dynamical systems: Foundations of harmony theory [A]. In: Rumelhart D E, McClelland J L (eds.). *Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 1: Foundations*. Cambridge, MA: MIT Press, 1986.

[5] Freund Y, Haussler D. Unsupervised learning of distributions on binary vectors using two layer networks [R]. Santa Cruz: University of California, UCSC-CRL-94-25, 1994.

[6] Roux N L, Bengio Y. Representational power of restricted Boltzmann machines and deep belief networks [J]. *Neural Computation*, 2006, 20(6): 1631-1649.

[7] Hinton G E. Training products of experts by minimizing contrastive divergence [J]. *Neural Computation*, 2002, 14(8): 1771-1800.

[8] Cho K Y. Improved learning algorithms for restricted Boltzmann machines [D]. Espoo: Aalto University, 2011.

[9] Teh Y W, Hinton G E. Rate-coded restricted Boltzmann machines for face recognition [C]. *In: Advances in Neural Information Processing Systems 13 (NIPS'00)*, 2001, MIT Press, pp. 908-914.

[10] Salakhutdinov R, Mnih A, Hinton G E. Restricted Boltzmann machines for collaborative filtering [C]. *In: Proceedings of the 24th International Conference on Machine Learning*, Corvallis, OR, 2007, pp. 791-798.

[11] 吴证, 周越, 杜春华, 袁泉. 组合主成分分析的受限波尔兹曼机神经网络的降维方法[J]. 上海交通大学学报, 2008, 42(4): 559-563.

[12] 吴金龙. Netflix Prize中的协同过滤算法[D]. 北京: 北京大学, 2010.

[13] 罗恒. 基于协同过滤视角的受限波尔兹曼机研究[D]. 上海: 上海交通大学, 2011.

[14] 潘闻特, 申丽萍. 基于BM神经网络编码的生理信号情感识别[J]. 计算机工程与设计, 2012, 33(3): 1101-1106.

[15] Hinton G E, Osindero S, Teh Y. A fast learning algorithm for deep belief nets [J]. *Neural Computation*, 2006, 18(7): 1527-1554.

[16] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks [J]. *Science*, 2006, 313: 504-507.

[17] Bengio Y. Learning deep architectures for AI [J]. *Foundations and Trends in Machine Learning*, 2009, 2(1): 1-127.

[18] Arel I, Rose D C, Karnowski T P. Deep machine learning-A new frontier in artificial intelligence research [Research Frontier] [J]. *IEEE Computational Intelligence Magazine*, 2010, 5(4): 13-18.

[19] Welling M, Rosen-Zvi M, Hinton G E. Exponential family harmoniums with an application to information retrieval [C]. *In: Advances in Neural Information Processing Systems 17 (NIPS'04)*, Cambridge, MA: MIT Press, 2005, pp. 1481-1488.

[20] Liu J S. Monto Carlo strategies in scientific computing [M]. New Work: Springer-Verlag, 2001.

[21] Chen H, Murray A F. Continuous restricted Boltzmann machine with an implementable training algorithm [J]. *IEE Proceedings Vision, Image and Signal Processing*, 2003, 150(3): 153-158.

[22] Nair V, Hinton G E. Rectified linear units improve restricted Boltzmann Machines [C]. *In: Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel, 2010, pp. 807-814.

[23] Hinton G E. A practical guide to training restricted Boltzmann machines [R]. Montreal: Department of Computer Science, University of Toronto, 2010.

[24] Courville A, Bergstra J, Bengio Y. A spike and slab restricted Boltzmann Machine [J]. *Journal of Machine Learning Research-Proceedings Track*, 2011, 15: 233-241.

[25] Tran T, Phung D Q, Venkatesh S. Mixed-variate restricted Boltzmann machines [J]. *Journal of Machine Learning Research-Proceedings Track*, 2011, 20: 213-229.

[26] Tieleman T. Training restricted boltzmann machines using approximations to the likelihood gradient [C]. *In: Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, 2008, pp. 1064-1071.

[27] Tieleman T, Hinton G E. Using fast weights to improve persistent contrastive divergence [C]. *In: Proceedings of the 26th International Conference on Machine Learning*, Montreal, Canada, 2009, pp. 1033-1040.

[28] Bengio Y, Courville A, Bincent P. Unsupervised feature learning and deep learning: a review and new perspectives [R]. Montreal: Department of Computer Science and Operations Research, University of Montreal, 2012.

[29] Schulz H, Müller A, Behnke S. Investigating convergence of restricted boltzmann machine learning [C]. *In: NIPS 2010 Workshop on Deep Learning and Unsupervised Feature Learning*, Whistler, Canada, 2010, pp. 1-9.

[30] Fischer A, Igel C. Empirical analysis of the divergence of Gibbs sampling based learning algorithms for restricted Boltzmann machines [C]. *In: Proceedings of the 20th International Conference on Artificial Neural Networks, Part III, LNCS 6354*, Berlin, Springer-Verlag, 2010, pp. 208-217.

[31] Gengio Y. Practical recommendations for gradient-based training of deep architectures [R]. Technical Report, Department of Computer Science and Operations Research, University of Montreal, 2012.

[32] Neal R M. Annealed importance sampling [J]. *Statistics and Computing*, 2001, 11(2): 125-139.

[33] Lee H, Ekanadham C, Ng A Y. Sparse deep belief net model for visual area V2 [C]. *In: Advances in Neural Information Processing Systems 20 (NIPS'07)*, Vancouver, Canada: MIT Press, 2008, pp. 873-880.

[34] Larochelle H, Bengio Y. Classification using discriminative restricted Boltzmann machines [C]. *In: Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, 2008, pp. 536-543.

[35] Larochelle H, Mandel M, Pascanu R, Bengio Y. Learning algorithms for the classification restricted Boltzmann machine [J]. *Journal of Machine Learning Research*, 2012, 13: 643-669.

[36] Sohn K, Lee H. Learning invariant representations with local transformations [C]. *In: Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, 2012.