基于动量方法的受限玻尔兹曼机的一种有效算法

沈卉卉[1,2,3],李宏伟[1,3]
(1. 中国地质大学数理学院,湖北武汉 430074;2. 湖北经济学院信息管理与统计学院,湖北武汉 430205;
3. 中国地质大学(武汉)地球内部多尺度成像湖北省重点实验室,湖北武汉 430074)

摘 要: 深度学习给模式识别与机器学习带来了巨大的变化,已成功应用于语言处理、图像处理、信号处理、商业经济等方面. 受限玻尔兹曼机(Restricted Boltzmann Machine,RBM)是一个表示能力强、很好的生成模型,多个 RBM 堆叠而构成的深度信念网络模型(Deep Belief Nets,DBN)的学习时间会较长. 为加快整个 DBN 网络的学习时间和提高分类效果,本文提出基于动量方法 RBM 的一种有效算法. 该算法在 RBM 预训练阶段,结合梯度上升算法特点采取快速上升的动量方式;以及 BP 算法微调阶段,为了能精确的找到最优点,结合梯度下降算法特点,相应的引入缓慢下降式的动量项,即在梯度上升和梯度下降过程中都使用不同的动量方式. 本文算法在 MNIST 手写数字体和 CMU-PIE 人脸数据库上进行了实验,结果表明,提出的改进算法能够有效地增强图像特征的表达能力,提高图像的分类效果和实验效率.

关键词: 深度学习; 受限玻尔兹曼机; Kullback – Leibler (KL) 距离; 蒙特卡罗思想; 动量
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2019)01-0176-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2019.01.023

An Effective Algorithm of Restricted Boltzmann Machine Based on Momentum Method

SHEN Hui-hui[1,2,3], LI Hong-wei[1,3]
(1. *School of Mathematics and Physics, China University of Geosciences, Wuhan, Hubei* 430074, *China*;
2. *School of Statistics & Information Management, Hubei University of Economics, Wuhan, Hubei* 430205, *China*;
3. *Hubei Subsurface Multi-scale Imaging Key Laboratory, China University of Geosciences, Wuhan, Hubei* 430074, *China*)

Abstract: Deep learning is bringing revolution to pattern recognition and machine learning, which has been successfully applied to language processing, image processing, signal processing, business economy and so on. Restricted Boltzmann machine (RBM) is a strong representation and generative mod el, however, the learning time of deep belief nets (DBN), which consists of multiple stacking RBM, will be longer. In this paper, the improved momentum method is used not only in gradient ascent algorithm but also in gradient descent algorithm for both classification accuracy enhancement and training time decreasing. According to the characteristics of the gradient ascent algorithm, a rapidly ascending momentum method is used in the RBM pre-training phase, which greatly improves the speed of learning. According to the characteristics of the gradient descent algorithm, an improved slowly descending momentum term is also used in the fine-tuning stage to accurately find the best point. Through the recognition experiments on the MNIST dataset and CMU-PIE face dataset, the achieved results show that the improved momentum algorithm can effectively enhance the ability of image feature expression and improve both accuracy and computation efficiency.

Key words: deep learning; restricted Boltzmann machine; Kullback – Leibler (KL) divergence; Monte Carlo method; momentum

收稿日期:2018-01-23;修回日期:2018-05-03;责任编辑:蓝红杰
基金项目:湖北省教育厅科技处重点项目(No. 20182203);湖北省高等学校优秀中青年创新团队计划项目(No. T201516)

1 引言

2006 年以来,机器学习领域中“深度学习”开始受到学术界广泛关注,到今天已经成为互联网大数据和人工智能的一个热潮[1]. 深度学习有三种典型的模型方法[1]:卷积神经网络模型(Convolutional Neural Networks,CNN)、深度信念网络模型、堆栈自编码网络模型(Stacked Auto—encoders,SAE).

RBM 因表示力强、易于推理等优点被成功用于深度神经网络中[1]. DBN 由多个 RBM 堆叠进行逐层学习而构成,因此,学习时间会相应的加长.

为加速网络收敛且得到可靠的估计,已有文献对 RBM 算法的改进或优化有两类. 一类是加正则项[2~4]或随机"dropout"[5,6],以达到稀疏目的,给目标任务提高效率,但仍然存在训练缓慢等问题[2]. 另一类算法,如 Hinton 等提出的对比散度(Contrastive Divergence,CD)算法[7]、专家乘积方法[8]和贪婪算法[9]训练 DBN 取得了一定效果,但训练时间还是长. Yang 等[10]提出弱监督学习的去噪 RBM 特征提取算法,但 RBM 分类效果并不明显. Lopes 等[11]和 Zhang 等[12]分别提出 RBM 在多个处理器和 Hadoop 平台进行,时间有较大提高,但 RBM 分类效果仍不明显. 还有引入动量项[13~19]来优化网络提高学习效率,其中文献[17]将经典动量[14]和 Nesterov 动量[15]应用到 RBM 中,其分类效果提高也不明显. 文献[19]则从理论上分析这两种动量加速效果差的原因,提出网络权值衰减的动量算法,但仍然是以增大网络权值为代价,权值过大会导致网络的泛化性能降低[19].

基于以上研究,本文是在文献[17,20,9]的基础上进行改进. 首先,利用文献[20]的 KL 距离解释 RBM 模型原理,推导出 RBM 的其中一种参数更新方式,但本文各参数更新方式与文献[20]不一样,且采样技术也不一样,文献[20]用的是平行退火采样,本文用文献[9]的 Gibbs 采样;接着,构建两个隐层的 DBN 网络,为加速整个网络收敛和提高鲁棒性,本文提出在 RBM 预训练阶段和微调阶段分别引入不同于以上动量方式的一种改进算法. 在每个 RBM 中结合梯度上升算法,采取快速收敛的动量方法来训练;同时,用 BP 算法微调整个网络时,却引入缓慢下降式的不同动量项,从而减少整个网络的学习时间和提高网络泛化能力.

2 受限玻尔兹曼机模型

如图 1 所示,RBM 有 n 个可见单元和 m 个隐单元,v_i 表示第 i 个可见单元的状态,a_i 表示可见单元 i 的偏置;h_j 表示第 j 个隐单元的状态;b_j 表示隐单元 j 的偏置;w_{ij}表示可见单元 i 与隐单元 j 之间的连接权重。向量 $\boldsymbol{v}=(v_1,v_2,\cdots,v_n)^{\mathrm{T}}$ 和 $\boldsymbol{h}=(h_1,h_2,\cdots,h_m)^{\mathrm{T}}$ 分别表示可见单元和隐单元的状态向量;$\boldsymbol{a}=(a_1,a_2,\cdots,a_n)^{\mathrm{T}}$ 和 $\boldsymbol{b}=(b_1,b_2,\cdots,b_m)^{\mathrm{T}}$ 表示可见层和隐层的偏置向量;$\boldsymbol{W}=(w_{ij})_{m\times n}\in R^{m\times n}$表示链接权重的矩阵. 令 $\theta=\{w_{ij},i=1,2,\cdots,n,j=1,2,\cdots,m;a_i,i=1,2,\cdots,n;b_j,j=1,2,\cdots,m.\}$,RBM 的主要任务就是求出这些参数 θ 的更新方式,以拟合给定的训练数据的概率分布.

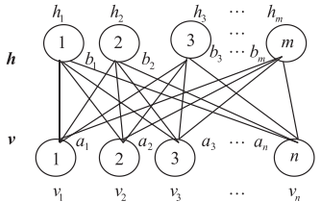


图 1　RBM 网络结构示意图

RBM 是一种随机网络,其层内无连接,因此,在给定可见层单元状态时,各隐层单元的激活条件独立;反之,给定隐层单元的状态时,可见层单元的激活也条件独立. RBM 中的神经元只有激活和不激活两种状态,也即取值是 0 或 1.

对于一组给定的状态$(\boldsymbol{v},\boldsymbol{h})$,RBM 作为一个系统所具备的能量函数定义为:

$$E(\boldsymbol{v},\boldsymbol{h}) = -\sum_{i=1}^{n}a_iv_i-\sum_{j=1}^{m}b_jh_j-\sum_{i=1}^{n}\sum_{j=1}^{m}v_iW_{ij}h_j,\quad \forall i,j,v_i,h_j\in\{0,1\} \tag{1}$$

利用该能量函数可得$(\boldsymbol{v},\boldsymbol{h})$的联合概率分布:

$$p(\boldsymbol{v},\boldsymbol{h})=\frac{e^{-E(\boldsymbol{v},\boldsymbol{h})}}{Z_\theta} \tag{2}$$

其中 $Z_\theta=\sum_{\boldsymbol{v},\boldsymbol{h}}e^{-E(\boldsymbol{v},\boldsymbol{h})}$,$Z_\theta$ 称为归一化因子.

2.1 最大化 RBM 网络表示的分布 $p(\boldsymbol{v})$

对于一个实际问题,学习 RBM 的目的是让 RBM 网络表示的可见层节点 $\boldsymbol{v}$ 的分布 $p(\boldsymbol{v})$最大可能的拟合输入样本所在样本空间的分布 $q(\boldsymbol{v})$. 从信息熵的角度理解:即使得 p 和 q 之间的 KL 距离最小,则两种分布越接近[20]:

$$\begin{aligned}KL(q||p) &= \sum_{\boldsymbol{v}\in\Omega}q(v)\ln\frac{q(\boldsymbol{v})}{p(\boldsymbol{v})}\\ &= \sum_{\boldsymbol{v}\in\Omega}q(\boldsymbol{v})\ln q(\boldsymbol{v})-\sum_{\boldsymbol{v}\in\Omega}q(\boldsymbol{v})\ln p(\boldsymbol{v})\end{aligned} \tag{3}$$

输入样本一旦确定,则 $q(\boldsymbol{v})$是确定的,即式(3)中第一项是确定的. 要使 KL 距离最小,则使式(3)第二项最大. 第二项的求和是对整个样本空间求和,因只有输入样本集 S,没有整个样本空间 Ω,于是利用蒙特卡罗方法[9]求其近似值.

假设有 T 个训练样本,$\boldsymbol{v}^t$ 表示第 t 个训练样本,则用这些样本均值逼近其期望:

$$\sum_{\boldsymbol{v}\in\Omega}q(\boldsymbol{v})\ln p(v)\approx\frac{1}{n}\sum_{\boldsymbol{v}\in S}\ln p(\boldsymbol{v})=\frac{1}{T}\sum_{t=1}^{T}\ln p(\boldsymbol{v}^t)$$

上式略等号左边是对整个样本空间 Ω 求和,右边求和是对采集样本 S,S 服从 $q(\boldsymbol{v})$ 分布.

因此,式(3)第二项最大化,就是求解 $\frac{1}{T}\sum_{t=1}^{T}\ln p(\boldsymbol{v}^t)$ 最大化.

2.2 RBM 问题的求解目标

我们最关心的是 RBM 网络表示的概率分布 $p(\boldsymbol{v},\boldsymbol{h})$ 的边缘分布 $p(\boldsymbol{v})$:

$$p(\boldsymbol{v})=\sum_{\boldsymbol{h}}p(\boldsymbol{v},\boldsymbol{h})=\sum_{\boldsymbol{h}}\frac{e^{-E(\boldsymbol{v},\boldsymbol{h})}}{Z_\theta}=\frac{1}{Z_\theta}\sum_{h}e^{-E(\boldsymbol{v},\boldsymbol{h})}$$

若训练样本的集合 S 有 T 个样本:

$$S=\{\boldsymbol{v}^1,\boldsymbol{v}^2,\cdots,\boldsymbol{v}^T\},\boldsymbol{v}^t=(v_1^t,v_2^t,\cdots,v_n^t)^T,t=1,2,\cdots,T$$

学习 RBM 的目标就是最大化如下似然函数:令

$$L(\theta)=\frac{1}{T}\sum_{t=1}^{T}\ln p(\boldsymbol{v}^t)$$

$$\theta^*=\arg\max_\theta L(\theta)=\arg\max_\theta\left[\frac{1}{T}\sum_{t=1}^{T}\ln p(\boldsymbol{v}^t)\right]$$

因此,最大化似然函数常用梯度上升法到最优参数 θ^*,对 θ 求导,使 θ 向增大的方向变化:

$$\theta:=\theta+\eta\frac{\partial L}{\partial\theta}$$

其中 $\eta>0$ 为学习率,关键是找梯度 $\frac{\partial L}{\partial\theta}$.

$$\begin{aligned}L(\theta)&=\frac{1}{T}\sum_{t=1}^{T}\ln p(\boldsymbol{v}^t)=\frac{1}{T}\sum_{t=1}^{T}\ln\sum_{h}p(\boldsymbol{v}^t,\boldsymbol{h})\\&=\frac{1}{T}\sum_{t=1}^{T}\left(\ln\sum_{\boldsymbol{h}}e^{-E(\boldsymbol{v}^t,\boldsymbol{h})}-\ln\sum_{\boldsymbol{v},\boldsymbol{h}}e^{-E(\boldsymbol{v},\boldsymbol{h})}\right)\end{aligned}$$

令 θ_i 是 $\theta=\{w_{ij},a_i,b_j\}$ 中的某一个参数,则对数似然函数关于 θ_i 的梯度为:

$$\begin{aligned}\frac{\partial L(\theta)}{\partial\theta_i}&=\frac{\partial\left(\frac{1}{T}\sum_{t=1}^{T}\left(\ln\sum_{\boldsymbol{h}}e^{-E(\boldsymbol{v}^t,\boldsymbol{h})}-\ln\sum_{\boldsymbol{v},\boldsymbol{h}}e^{-E(\boldsymbol{v},\boldsymbol{h})}\right)\right)}{\partial\theta_i}\\&=\frac{1}{T}\sum_{t=1}^{T}\frac{\partial}{\partial\theta_i}\left(\ln\sum_{\boldsymbol{h}}e^{-E(\boldsymbol{v}^t,\boldsymbol{h})}-\ln\sum_{\boldsymbol{v},\boldsymbol{h}}e^{-E(\boldsymbol{v},\boldsymbol{h})}\right)\\&=\frac{1}{T}\sum_{t=1}^{T}\left[E_{p(\boldsymbol{h}\mid\boldsymbol{v}^t)}\left(\frac{\partial(-E(\boldsymbol{v}^t,\boldsymbol{h}))}{\partial\theta_i}\right)-E_{p(\boldsymbol{v},\boldsymbol{h})}\left(\frac{\partial(-E(\boldsymbol{v},\boldsymbol{h}))}{\partial\theta_i}\right)\right]\end{aligned}\tag{4}$$

对式(4)右边的两项全部采用蒙特卡罗方法[9]近似求期望,当给定一个训练样本时,于是有关 w_{ij},a_i,b_j 三个参数的导数分别为:

$$\begin{aligned}\frac{\partial L(\theta)}{\partial w_{ij}}&=\sum_{\boldsymbol{h}}p(\boldsymbol{h}\mid\boldsymbol{v}^t)\times\frac{\partial(-E(\boldsymbol{v}^t,\boldsymbol{h}))}{\partial w_{ij}}-\sum_{\boldsymbol{v},\boldsymbol{h}}p(\boldsymbol{v},\boldsymbol{h})\times\frac{\partial(-E(\boldsymbol{v},\boldsymbol{h}))}{\partial w_{ij}}\\&=\sum_{\boldsymbol{h}}p(\boldsymbol{h}\mid\boldsymbol{v}^t)\times(\boldsymbol{v}_i^t\boldsymbol{h}_j)-\sum_{\boldsymbol{v},\boldsymbol{h}}p(\boldsymbol{v},\boldsymbol{h})\times(v_ih_j)\\&\approx v_i^{(0)}h_j^{(0)}-v_i^{(k)}p(h_j=1\mid\boldsymbol{v}^{(k)})\end{aligned}\tag{5}$$

$$\begin{aligned}\frac{\partial L(\theta)}{\partial a_i}&=\sum_{\boldsymbol{h}}p(\boldsymbol{h}\mid\boldsymbol{v}^t)\times\frac{\partial(-E(\boldsymbol{v}^t,\boldsymbol{h}))}{\partial a_i}-\sum_{\boldsymbol{v},\boldsymbol{h}}p(\boldsymbol{v},\boldsymbol{h})\times\frac{\partial(-E(\boldsymbol{v},\boldsymbol{h}))}{\partial a_i}\\&=E_{p(\boldsymbol{h}\mid\boldsymbol{v}^t)}(\boldsymbol{v}_i^t)-E_{p(\boldsymbol{v},\boldsymbol{h})}(v_i)\\&\approx v_i^{(0)}-v_i^{(k)}\end{aligned}\tag{6}$$

$$\begin{aligned}\frac{\partial L(\theta)}{\partial b_j}&=\sum_{\boldsymbol{h}}p(\boldsymbol{h}\mid\boldsymbol{v}^t)\times\frac{\partial(-E(\boldsymbol{v}^t,\boldsymbol{h}))}{\partial b_j}-\sum_{\boldsymbol{v},\boldsymbol{h}}p(\boldsymbol{v},\boldsymbol{h})\times\frac{\partial(-E(\boldsymbol{v},\boldsymbol{h}))}{\partial b_j}\\&=E_{p(\boldsymbol{h}\mid\boldsymbol{v}^t)}(h_j)-E_{p(\boldsymbol{v},\boldsymbol{h})}(h_j)\\&\approx h_j^{(0)}-p(h_j=1\mid\boldsymbol{v}^{(k)})\end{aligned}\tag{7}$$

其中 $h_j^{(0)}$ 为由 $\boldsymbol{v}^{(0)}$ 第一次采样到的隐层单元中第 j 个隐单元 h_j 的状态值.

为使算法稳定,引入学习率 η,各参数更新:

$$\Delta w_{ij}=\Delta w_{ij}+\eta[v_i^{(0)}\times h_j^{(0)}-v_i^{(k)}\times p(h_j=1\mid v^{(k)})]$$

$$\Delta a_i=\Delta a_i+\eta[v_i^{(0)}-v_i^{(k)}]$$

$$\Delta b_j=\Delta b_j+\eta[h_j^{(0)}-p(h_j=1\mid v^{(k)})]$$

3 改进的动量方法

动量方法在神经网络领域中,通常是用来加速和提高 BP 算法的. Rumelhart 等[13]提出动量项能加速整个 BP 网络的学习速度,也是改进算法稳定性的一种途径. 因此,本文借鉴动量方法的思想来提高网络收敛速度和改进算法的稳定性.

文献[3]用随机最大化似然函数算法,结合权值衰减的经典动量方法[3],手写体识别实验取得了 1.1% 的错误率. Sutskever 等[15]提出 Nesterov 动量方法. Nitanda[16]将 Nesterov 动量和随机梯度方差减少技术相结合,手写体识别实验,错误率在 4% 左右. Zareba 等[17]用经典动量和 Nesterov 动量两种方法做了比较,手写体识别实验中,得出 Nesterov 动量效果好,RBM 识别错误率是 2.04%. Yuan 等[18]分析了动量随机梯度法的收敛性和效果性能. 这些研究表明在随机梯度方法中,动量方法的确有加速 RBM 网络收敛和提高学习性能的效果. 以上动量方法要么只用于 BP 网络,或仅用于 RBM 预训练阶段,对 DBN 整个网络的训练和微调阶段,没有同时用动量方法.

基于以上研究,结合 RBM 结构原理,我们提出根据梯度上升和梯度下降算法特点,各自更新过程中分别引入不同方式的动量项,即在 RBM 预训练阶段和微调阶段都同时使用不同的动量方式,既可以提高识别效果,又可以减少网络学习时间. 参数的设置,最开始设 w_{ij},a_i,b_j 都为 0,学习率 η 设置较大时,收敛快,但会引起不稳定;η 设置较小时,可避免不稳定,但收敛慢. 为克服这一情况,引入动量项(momentum)m^*,使得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[13],这样可以避免算法过早的收敛到局部最优. 这种组合有不同的组合方式,经典动量方式[21,14]、Nesterov 动量方式[15-17]及权值衰减的动量方式[3,8].

经典的动量方式[14,17,21,22]为:

$$\theta_t = \theta_{t-1} + m^* \Delta\theta_{t-1} - \eta[\frac{\partial L(\theta)}{\partial \theta_i}] \tag{8}$$

式(8)表示的是上一次参数值修改方向与本次梯度的负方向的组合[22].

Nesterov 动量方式[15,17]为:

$$\theta_t = \theta_{t-1} + m^* \Delta\theta_{t-1} - \eta[\frac{\partial L(\theta + m^* \Delta\theta_{t-1})}{\partial \theta_i}] \tag{9}$$

Nesterov 动量和经典动量之间的区别体现在梯度计算上,Nesterov 动量中,梯度计算在当前速度之后. 因此,Nesterov 动量可看作在经典动量方法中添加了一校正因子[22].

文献[19]从理论上分析式(8)和式(9)这两种动量加速效果差的原因,进而提出网络权值衰减的动量算法,但仍然是以增大网络权值为代价[19]. 基于此,本文采取的动量方式不同于以上. 根据 RBM 训练采取的是梯度上升算法,则动量方式采取的是上一次参数值修改方向与本次梯度正方向相加,步长最大,在梯度方向上不停的加速,这样更加快速的收敛到最优. 因此,本文在 RBM 训练中采取参数更新方式如下:

$$\theta_t = \theta_{t-1} + m^* \Delta\theta_{t-1} + \eta[\frac{\partial L(\theta)}{\partial \theta_i}] \tag{10}$$

而在 DBN 网络整个微调阶段,采取的是梯度下降算法,则结合梯度下降的自身特点,为使梯度下降过程中能精确的找到最优点,微调时采用更新步长较小,便于不错过每个最低点. 因此,微调阶段 BP 网络中参数 $\boldsymbol{W}$ 的更新方式为:

$$nn.\boldsymbol{W}(n) = nn.\boldsymbol{W}(n-1) - m' \times nn.\Delta\boldsymbol{W}(n-1) - \eta' \times [\frac{1}{T}\Delta\boldsymbol{W}] \tag{11}$$

4　实验

实验采用 MNIST 数据集和 CMU-PIE 人脸数据库的数据图像作为实验对象,用算法 1 在两个数据集上做实验,来说明本文提出的动量算法能提高网络识别效果和减少网络训练及学习时间. 实验环境是在 Microsoft Windows 7 的操作系统下,硬件配置是:Intel(R) core(TM) i7-4770 HQ CPU @2.2GHz,16GB 内存.

算法 1　DBN 学习算法的主要步骤

1) 给定训练集 $S = \{\boldsymbol{v}^1, \boldsymbol{v}^2, \cdots, \boldsymbol{v}^T\}$,分批训练,每批有 $S = T$ 个样本,可见单元和隐单元分别设为 n 个和 m 个,每个 RBM 的学习率设 η,动量为 m^*,迭代次数为 J. BP 网络的学习率设为 η',动量为 m' 迭代次数为 J'.
2) 随机初始化 $\Delta w_{ij} = \Delta a_i = \Delta b_j = 0$ for $i = 1,2,\cdots,n; j = 1,2,\cdots,m$
3) 对 RBM 中所有的样本 $\boldsymbol{v} \in S$
4) $\boldsymbol{v}^{(0)} \leftarrow \boldsymbol{v}$
5) 对每个样本 $t = 0,1,\cdots,k-1$
6) 对 $j = 1,2,\cdots,m$ 采样 $h_j^{(t)} \sim p(h_j \mid v^{(t)})$
7) 对 $i = 1,2,\cdots,n$ 采样 $v_i^{(t+1)} \sim p(v_i \mid h^{(t)})$
8) End for
9) for $i = 1,2,\cdots,n; j = 1,2,\cdots,m$ do
10) 参数更新:
11) $\Delta w_{ij} \leftarrow m^* \cdot \Delta w_{ij} + \eta[v_i^{(0)} h_j^{(0)} - v_i^{(k)} p(h_j = 1 \mid v^{(k)})]$
12) $\Delta a_i \leftarrow m^* \cdot \Delta a_i + \eta[v_i^{(0)} - v_i^{(k)}]$
13) $\Delta b_j \leftarrow m^* \cdot \Delta b_j + \eta[h_j^{(0)} - p(h_j = 1 \mid v^{(k)})]$
14) End for
15) 在每个 RBM 网络中迭代 $iter = 1,2,\cdots,J$
16) 所有的参数批量更新方式:
17) $\boldsymbol{W} \leftarrow \boldsymbol{W} + m^* \cdot \Delta\boldsymbol{W} + \eta[\frac{1}{T}\Delta\boldsymbol{W}]$
18) $\boldsymbol{a} \leftarrow \boldsymbol{a} + m^* \cdot \Delta\boldsymbol{a} + \eta[\frac{1}{T}\Delta\boldsymbol{a}]$
19) $\boldsymbol{b} \leftarrow \boldsymbol{b} + m^* \cdot \Delta\boldsymbol{b} + \eta[\frac{1}{T}\Delta\boldsymbol{b}]$
20) End for
21) 在 BP 网络中迭代 $iter = 1,2,\cdots,J'$
22) 参数 $\boldsymbol{W}$ 更新方式:
23) $\boldsymbol{W} \leftarrow \boldsymbol{W} - m' \cdot \Delta\boldsymbol{W} - \eta'[\frac{1}{T}\Delta\boldsymbol{W}]$
24) End for

4.1　MNIST 数据库手写体识别实验

实验一采用 MNIST 数据集,包含 70000 张 0-9 的 10 个手写体数字图像,每张图片大小是 28×28,随机选取 60000 张用于训练,10000 张用来测试.

若构建一个 RBM 网络:784-400-10,隐单元设置为 400 时效果最好,用本文算法,RBM 迭代 30 次,微调 30 次,错误率可达 1.58%. MNIST 数据库的测试集上不同算法的实验结果如表 1 所示. 以下各表中的耗时指的是训练和测试一起的时间.

表 1　不同 RBM 模型在 MNIST 数据集上的实验情况

不同 RBM / 网络结构		分类错误率	耗时(分)
去噪 RBM[10]	784-1500-10	6.45%	—
MapReduceRBM[12]	784-900-10	2.92%	7.5
本文算法 RBM	**784-900-10**	**1.63%**	**7**
本文算法 RBM	**784-400-10**	**1.58%**	**6.5**
MapReduceRBM[12]	784-900-10	2.89%	12
本文算法 RBM	**784-900-10**	**1.46%**	**11**
动量 RBM[17]	784-400-10	2.04%	—
本文算法 RBM	**784-400-10**	**1.42%**	**10**

从表 1 的结果可以看出,本文提出的算法构成的 RBM 学习时间和分类效果比 Yang[10] 的去噪 RBM 方法、MapReduce RBM 方法[12]、以及 Zareba 等[17] 的动量方法都明显有提高,错误率至少降低了 30%. 且本文算法简单易于实现,不需要面临 Hadoop 平台的 MapReduce 框架设计等问题.

若构建 2 个隐层网络:784-400-1000-10,用本文算法,经过 1 小时 36 分钟,错误率可达 1.02%. 本文算法与其他 DBN 的方法对比结果如表 2 所示.

表 2 不同 DBN 模型在 MNIST 数据集上的实验结果(错误率)

不同模型算法	分类错误率	耗时(小时)
专家乘积方法[8]	1.7%	24
Hinton 的贪婪算法[9]	1.25%	大于 9
异构体并行框架方法[23]	1.39%	—
权衰减动量方法[3]	1.1%	—
训练和微调都无动量方法	**1.3%**	**2**
训练用动量、微调无动量方法	**1.15%**	**1.8**
本文算法	**1.02%**	**1.6**

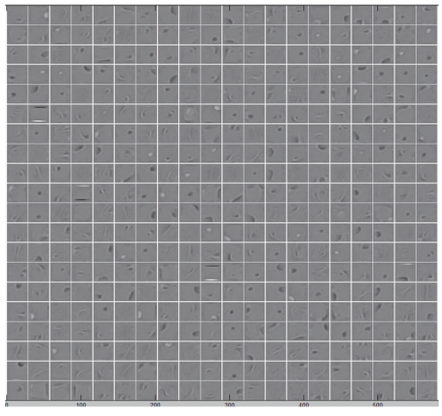
从表 2 的结果显示,本文提出的算法 DBN 分类效果比 Mayraz[8]、Hinton[9]、Swersky 等[3] 和 Wang 等[23] 有关 DBN 方法要好,与 Hinton[9] 的每层学习时间要几个小时相比,本文动量算法能缩短 80% 以上学习时间且分类效果好. 试验结果表明,本文提出的动量算法在保证网络正确率有所提升的同时,还能减少网络训练耗时.

图 2(*a*)为本文算法学习到的权重图像可视化,图 2(*b*)为专家乘积算法[8] 的权重图像可视化. 图像中线条越清晰,说明网络对训练数据学习得越充分,则此网络会因过拟合问题,出现对测试数据分类效果差的现象. 因此,图 2(*a*)中优化网络的特征提取图像线条相对模糊,表明其鲁棒性较好,其网络分类效果更佳.

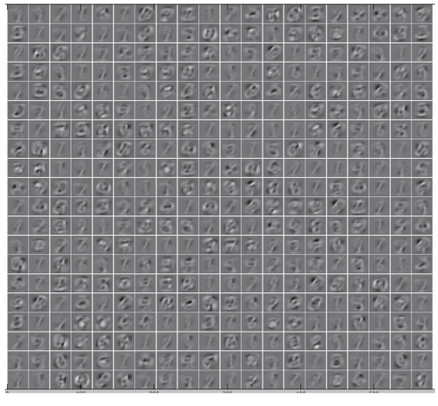
4.2 CMU-PIE 数据库人脸识别实验

实验二采用 CMU-PIE 人脸数据库,包含 68 位志愿者的 41,368 张多姿态、光照和表情的面部图像. 选取库中尺寸为 32×32 的 30 种人脸图像进行分类,每种人脸均有 170 张不同的数据,即共有 5100 个带标数据,随机选取 4500 个作为训练数据,其余 600 个数据作为测试数据;以及选取尺寸为 32×32 的 68 个人的 11554 张图片,其中 10000 张图像作为训练样本,剩下的 1554 张图片用来测试.

实验二在两个不同规模的 DBN 网络上进行优化,在尺寸为 32×32 的人脸数据上,原网络与动量优化网络的分类准确率和耗时上的对比结果,以及 68 人的图像分类结果如表 3 所示.



(*a*) 可视化本文算法学习到的第一层网络的权重



(*b*) 可视化专家乘积方法学习到的第一层网络的权重

图2

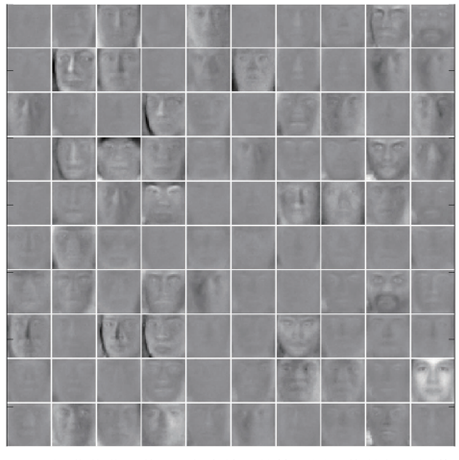
表 3 动量优化网络与原网络在 CMU-PIE 数据集上实验结果对比

网络结构	分类正确率(错误率)	耗时(分)
1024-500-500-30	97.83% (2.17%)	6.5
1024-500-500-30(m)	**99.17%** (**0.83%**)	**6**
文献[24]的方法,30 人	98.83% (1.17%)	17.7
文献[25]的方法,68 人	96.17% (3.83%)	—
1024-100-100-68	**97.8%** (**2.2%**)	**6.5**
1024-100-100-68(m)	**98.47%** (**1.53%**)	**6**

本文算法对 30 人的人脸识别效果要好于 68 人的人脸识别效果,因 68 人的图像数据量大且类别多. 与文献[24]相比,本文算法对 30 人的识别错误率可降低 29% 及时间上节省 66%. 表 3 对比结果显示,本文提出的动量算法无论是对 30 人的还是 68 人的人脸识别,较未用动量时的网络识别正确率都有较大提高. 在没有如此大规模数据时,网络学习时间没有很明显的优势,但更能符合实际问题应用. 本文算法对 68 人的人脸识别效果与文献[25]结果相比错误率降低了 60%,说明 RBM 模型对提出的改进动量算法能够有效地增强图像特征的表达能力,其网络的泛化能力和鲁棒性都较好.

可视化随机的 100 张人脸图片以及网络学习到的权重图像如下图 3 所示.

(*a*) 可视化原数据库中随机的100张人脸图片



(*b*) 可视化优化网络100个隐单元的第一层网络的权重图像

图3

图 3(*a*)为原数据库中人脸图片的可视化,含有丰富的姿态和光照变化的图像. 结合表 3 实验结果分析可知,本文算法对多姿态、光照和表情变化仍然有不错的表现和特征表达能力,鲁棒性较好. 图 3(*b*)为动量优化网络学习到的权重图像可视化,隐约可看见模糊的人脸,它相当于特征提取器,提取到的人脸特征具有较好的分类性能.

5　结论

本文提出了基于动量方法 RBM 的一种有效算法,在 RBM 模型的理论原理上,将不同组合方式的动量项加入到 RBM 训练阶段和微调阶段,起到加速整个网络收敛的作用和提高网络泛化能力,并在 MNIST 手写体数据集和 CMU-PIE 人脸数据库上分别做识别实验,与其他 RBM 同类方法做比较. 试验结果表明,本文提出的动量算法简单有效,在保证网络正确率有所提升的同时,能减少网络学习时间. 说明 RBM 模型对于改进的算法具有较强的特征表达能力和泛化能力及鲁棒性能. 在 MNIST 和 CMU-PIE 公用数据库上的测试结果也说明了本文算法的有效性.

参考文献

[1] 焦李成,杨淑媛,刘芳等. 神经网络七十年:回顾与展望[J]. 计算机学报,2016,39(1):1 – 21.
Jiao Li-cheng, Yang Shu-yuan, Liu Fang. Neural network in seventy: retrospect and prospect [J]. Chinese Journal of Computers, 2016, 39(1): 1 – 21. (in Chinese)

[2] Lee H, Grosse R, Ranganath R. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations [A]. Proceedings of the 26th Annual International Conference on Machine Learning [C]. New York: ACM, 2009: 609 – 616.

[3] Swersky K, Chen B, Marlin B M. A tutorial on stochastic approximation algorithms for training restricted boltzmann machines and deep belief nets [A]. ITA[C]. IEEE, 2010, 80 – 89.

[4] Mei X G, Ma Y, Fan F. Infrared ultraspectral signature classification based on a restricted Boltzmann machine with sparse and prior constraints [J]. International Journal of Remote Sensing, 2015, 36(18): 4724 – 4747.

[5] Hinton G E, Srivastava N, Krizhevsky A, Sutskever I, Salakhutdinov R. Improving neural networks by preventing co-adaptation of feature detectors [DB/OL]. https://arxiv.org/pdf/1207.0580v1.pdf, 2012-7-3.

[6] Wager S, Wang S, Liang P. Dropout training as adaptive regularization [DB/OL]. https://arxiv.org/pdf/1307.1493v2.pdf, 2013-11-1.

[7] Hinton G E. Training products of experts by minimizing contrastive divergence [J]. Neural Computation, 2002, 14(8): 1711 – 1800.

[8] Mayraz G, Hinton G E. Recognizing handwritten digits using hierarchical products of experts [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(2): 189 – 197.

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

[10] 杨杰,孙亚东,张良俊, 刘海波. 基于弱监督学习的去噪受限玻尔兹曼机特征提取算法 [J]. 电子学报,2014,42(12):2365 – 2370.
Yang Jie, Sun Ya-dong, Zhang Liang-jun, Liu Hai-bo. Weakly supervised learning with denoising restricted Boltzmann machines for extracting features [J]. Acta Electronica Sinica, 2014, 42(12): 2365 – 2370. (in Chinese)

[11] Lopes N, Ribeiro B, Goncalves J. Restricted Boltzmann

machines and deep belief networks on multi-core processors [A]. WCCI 2012 IEEE World Congress on Computational Intelligence June [C]. Brisbane, Australia, 2012. 10 – 15.

[12] Zhang Ch Y, Philip-Chen C L, Chen D W. MapReduce based distributed learning algorithm for restricted Boltzmann machine [J]. Neurocomputing, 2016, 198: 4 – 11.

[13] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors [J]. Nature, 1986, 323: 533 – 536.

[14] Hinton G E. A practical guide to training restricted Boltzmann machines [R]. Neural Networks: Tricks of the Trade (2nd ed), 2012. 599 – 619.

[15] Sutskever I, Martens J, Dahl G, Hinton G. On the importance of initialization and momentum in deep learning [A]. Proc International Conference on Machine Learning [C]. Atlanta, USA, 2013: 1139 – 1147.

[16] Nitanda A. Stochastic proximal gradient descent with acceleration techniques [A]. Proc Advances in Neural Information Processing Systems [C]. Montreal, Canada, 2014. 1574 – 1582.

[17] Zareba S, Gonczarek A, Tomczak J M, Swiatek J. Accelerated learning for restricted Boltzmann machine with momentum term [A]. International Conference on Systems Engineering [C]. Coventry, UK, 2015. 187 – 192.

[18] Yuan K, Ying B C, Sayed A H. On the influence of momentum acceleration on online learning [J]. Journal of Machine Learning Research, 2016 (17): 1 – 66.

[19] 李飞, 高晓光, 万开方. 基于权值动量的 RBM 加速学习算法研究[J]. 自动化学报, 2017, 43(7): 1142 – 1159.
Li Fei, Gao Xiao-guang, Wan Kai-fang. Research on RBM accelerating learning algorithm with weight momentum [J]. Acta Automatica Sinica, 2017, 43(7): 1142 – 1159. (in Chinese)

[20] Fischer A, Igel C. Training restricted Boltzmann machines: An introduction [J]. Pattern Recognition, 2014, (47): 25 – 39.

[21] Polyak T. Some methods of speeding up the convergence of iteration methods[J]. USSR Computational Mathematics and Mathematical Physics, 1964, 4(5): 1 – 17.

[22] Goodfellow I, Bengio Y, Courville A 著, 赵申剑等译, 深度学习[M]. 北京: 人民邮电出版社, 2017. 181 – 187.

[23] 王岳青, 窦勇, 吕启, 李宝峰, 李腾. 基于异构体系结构的并行深度学习编程框架 [J]. 计算机研究与发展, 2016, 53 (6): 1202 – 1210.
Wang Yue-qing, Dou Yong, Lv Qi, et al. A parallel deep learning programming framework based on heterogeneous architecture [J]. Journal of Computer Research and Development, 2016, 53(6): 1202 – 1210. (in Chinese)

[24] 付晓, 沈远彤, 付丽华等. 基于特征聚类的稀疏自编码快速算法[J]. 电子学报, 2018, 46(5): 1041 – 1046.
FU Xiao, SHEN Yuan-tong, FU Li-hua, et al. . An optimized sparse auto-encoder network based on feature clustering [J]. Acta Electronica Sinica 2018, 46(5): 1041 – 1046 (in Chinese)

[25] 李倩玉, 蒋建国, 齐美彬. 基于改进深层网络的人脸识别算法 [J]. 电子学报, 2017, 45(3): 619 – 625.
Li Qian-yu, Jiang Jian-guo, Qi Mei-bin. Face recognition algorithm based on improved deep networks [J]. Acta Electronica Sinica, 2017, 45(3): 619 – 625. (in Chinese)

作者简介

沈卉卉 女, 1980 年生于湖北黄冈. 现为湖北经济学院信息管理与统计学院副教授、中国地质大学在读博士生. 研究方向为机器学习与数据处理.
E-mail: sophy0209@126.com

李宏伟 男, 1965 年生于湖南汨罗. 现为中国地质大学(武汉)数理学院教授, 博士生导师. 主要研究方向为信息处理与智能计算.