一种基于修正动量的RBM算法

沈卉卉[1,2,3],刘国武[2],付丽华[1],刘智慧[1],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: 受限玻尔兹曼机(Restricted Boltzmann Machine,RBM)是一种随机网络、概率图模型,它是一种比较有效的的无监督学习模型.针对RBM梯度近似的一种计算方法对动量加速不敏感,以及识别效果不理想等问题,本文提出一种基于修正动量的RBM算法.该算法结合RBM梯度近似方法,通过修改隐单元偏置参数的更新方式,避免RBM模型中隐单元取值采用概率值时导致模型识别效果不理想、动量加速有限等问题.同时,在RBM预训练阶段采用快速上升的动量方式,以加速网络收敛;在微调阶段引入缓慢下降的动量项,以避免陷入局部最优点并提高识别效果.本文算法通过在MNIST手写数字体,Extended Yale B和CMU-PIE人脸数据库上的数值实验结果表明,提出的算法能够有效地提高计算效率和提高网络泛化能力.该算法不仅对RBM的应用领域扩展具有十分积极的实际意义,且为深度学习的应用方法提供一种新的研究思路和借鉴.

关键词: 深度学习;无监督学习;受限玻尔兹曼机;梯度近似算法;Gibbs采样;动量加速;泛化能力
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2019)09-1957-08
电子学报URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2019.09.020

An Algorithm Based on Modified Momentum Using Restricted Boltzmann Machine

SHEN Hui-hui[1,2,3],LIU Guo-wu[2],FU Li-hua[1],LIU Zhi-hui[1],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: Restricted Boltzmann machine (RBM) is a stochastic neural network and probabilistic graphical model, which is one of the most effective models without supervision in deep learning. Focusing on the gradient approximation algorithm insensitivity to the momentum acceleration and recognition effectiveness in RBM, we propose the algorithm based on modified momentum using RBM. When the rule to update the hidden states adopts the probability value instead of sampling a binary value, this calculation method for the RBM gradient approximation leads to the undesirable recognition performance and limited momentum acceleration. Therefore, we modify the updating rule of the hidden bias to avoid these problems. Simultaneously, we use the rapidly ascending momentum method to improve the learning speed in the RBM pre-training phase. An improved slowly descending momentum method is also used in the fine-tuning stage to accurately find the best point, which is far from becoming trapped in poor local optima and improves the classification effect. Through the recognition experiments on MNIST dataset, Extended Yale B and CMU-PIE face dataset, the achieved results show that the proposed algorithm can enhance the computation efficiency and improve the generalization ability of networks. The algorithm not only extends the application fields of RBM, but also provides a new research idea and reference for the application method of deep learning.

Key words: deep learning; unsupervised learning; restricted Boltzmann machine; gradient approximation algorithm; Gibbs sampling; momentum acceleration; generalization ability

1 引言

受限玻尔兹曼机是深度学习模型之一[1],因其表示能力强、是个很好的生成模型等优点被广泛用于深度神经网络[2].RBM训练的优劣将直接影响整个深度网络的性能.因此,如何优化RBM算法以提高网络泛化

收稿日期:2019-01-23;修回日期:2019-03-04;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. D20182203)

能力和鲁棒性,是网络应用中的重要问题[3].

RBM 算法中关于算法收敛和提高计算效率的研究,归纳起来有两类. 一类是在算法中引入动量项[4~17]来加速网络收敛以提高学习效率,常见的有经典动量方法[4~11,15]和 Nesterov 动量方法[13~15],这类方法要么仅与随机梯度下降算法(Stochastic Gradient Descent, SGD)[3]相结合[8~9,12~15],要么仅与随机梯度上升算法(Stochastic Gradient Ascent,SGA)[3]相结合[17]. 这类动量方法的确有加速网络收敛的效果,但分类精度有待进一步提高. 李飞等[17]则分析了这两种动量方法[8,13]加速效果差的原因,提出基于网络权值衰减的动量算法,但仍是以增大网络权值为代价,权值过大会导致网络泛化性能降低[17]. 另一类是通过并行结构来加速,如 Lopes 等[18]和 Zhang 等[19]提出将 RBM 分别在多个处理器和 Hadoop 平台进行,时间有较大提高,但分类效果不理想. Fischer 等[20]发现在他们的算法中用动量项并没有起到加速效果,实验结果并没有改变已有文献的结果.

针对文献[17]和[20]提出的问题,本文提出一种基于修正动量的 RBM 算法. 与文献[20]相比,做了如下 3 方面的改进:首先,在 RBM 梯度上升和梯度下降算法中分别采用不同动量方法. 分别与 SGA 和 SGD 相结合共同作用,从而加速网络收敛和提高图像分类效果. 然后,用 Gibbs 采样[1]来估计 RBM 对数似然函数梯度. 由于文献[20]的平行退火(Parallel Tempering,PT)[21,22]采样方法的多个 Markov 链混合的困难一直存在,而当目标分布不是很复杂的时候,Gibbs 采样[23]通过单个马尔科夫链就能很好的近似目标分布. 最后,本文修改了隐单元偏置的参数更新方式,使其在相同时间内得到较好的分类效果且消除权重密度问题[24].

2 RBM 模型

RBM 是一种随机网络、概率图模型[1],如图 1 所示,它能够无监督地学习到输入数据的概率分布. RBM 有 n 个可见单元和 m 个隐单元,用向量 $\boldsymbol{v}=(v_1,v_2,\cdots,v_n)^{\mathrm{T}}$ 和 $\boldsymbol{h}=(h_1,h_2,\cdots,h_m)^{\mathrm{T}}$ 来表示,其中 v_i,h_j 分别表示第 i 个可见单元的状态和第 j 个隐单元的状态;$\boldsymbol{W}=(w_{ij})_{m\times n}\in\mathbf{R}^{m\times n}$ 表示链接权重的矩阵,其中 w_{ij} 表示第 i 个可见单元和第 j 个隐单元之间的连接权重;$\boldsymbol{a}=(a_1,a_2,\cdots,a_n)^{\mathrm{T}}$ 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,其中 a_i,b_j 分别表示第 i 个可见单元的偏置和第 j 个隐单元的偏置;令 $\theta=\{w_{ij},i=1,2,\cdots,n,j=1,2,\cdots,m;a_i,i=1,2,\cdots,n;b_j,j=1,2,\cdots,m\}$ 表示模型中未知参数的组合,RBM 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.

对于一组给定的状态$(\boldsymbol{v},\boldsymbol{h})$,RBM 模型的能量函数

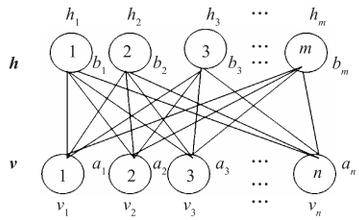


图1 RBM网络结构示意图

定义为[1]:

$$E(\boldsymbol{v},\boldsymbol{h})=-\sum_{i=1}^{n}a_iv_i-\sum_{j=1}^{m}b_jh_j-\sum_{i=1}^{n}\sum_{j=1}^{m}v_iw_{ij}h_j \tag{1}$$

其中,$\forall i,j,v_i,h_j\in\{0,1\}$.

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

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

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

RBM 模型中层内无连接,因此,在给定可见层单元状态时,各隐单元的激活条件独立,可由式(1)和式(2)导出各隐单元的条件激活概率公式:

$$p(h_j=1\mid\boldsymbol{v})=\mathrm{sigmoid}\left(b_j+\sum_{i=1}^{n}v_iw_{ij}\right) \tag{3}$$

由 RBM 对称性可知,当给定隐层单元的状态时,可见层单元的激活也条件独立. 同样可得:

$$p(v_i=1\mid\boldsymbol{h})=\mathrm{sigmoid}\left(a_i+\sum_{j=1}^{m}w_{ij}h_j\right) \tag{4}$$

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:使 RBM 表示的 Gibbs 分布与输入样本表示的分布尽可能的接近. 即使得 p 和 q 之间的 KL 距离:

$$\begin{aligned}\mathrm{KL}(q\parallel p)&=\sum_{v\in\Omega}q(\boldsymbol{v})\ln\frac{q(\boldsymbol{v})}{p(\boldsymbol{v})}\\&=\sum_{v\in\Omega}q(\boldsymbol{v})\ln q(\boldsymbol{v})-\sum_{v\in\Omega}q(\boldsymbol{v})\ln p(\boldsymbol{v})\end{aligned} \tag{5}$$

达到最小[20].

因没整个样本空间 Ω 的完整数据,只有输入样本集 $S=\{\boldsymbol{v}^1,\boldsymbol{v}^2,\cdots,\boldsymbol{v}^{\mathrm{T}}\}$,要使式(5)KL 距离最小,只能利用 Monte Carlo 方法[1]求其近似值. 最小化式(5)转化为最大化似然函数式(6)[20,25]:

$$L(\theta)=\frac{1}{T}\sum_{t=1}^{T}\ln p(\boldsymbol{v}^t) \tag{6}$$

其中 $\boldsymbol{v}^t=(v_1^t,v_2^t,\cdots,v_n^t)^{\mathrm{T}}$ 表示第 t 个训练样本,共有 T 个训练样本. 最大化似然函数常用随机梯度上升算法来求得最优参数 θ^*,于是式(6)变为:

$$\begin{aligned}L(\theta)&=\frac{1}{T}\sum_{t=1}^{T}\ln p(\boldsymbol{v}^t)=\frac{1}{T}\sum_{t=1}^{T}\ln\sum_{h}p(\boldsymbol{v}^t,\boldsymbol{h})\\&=\frac{1}{T}\sum_{t=1}^{T}\left(\ln\sum_{\boldsymbol{h}}\mathrm{e}^{-E(\boldsymbol{v}^t,\boldsymbol{h})}-\ln\sum_{\boldsymbol{v},\boldsymbol{h}}\mathrm{e}^{-E(\boldsymbol{v},\boldsymbol{h})}\right)\end{aligned} \tag{7}$$

为找梯度$\frac{\partial L}{\partial \theta}$,对式(7)求导:

$$\frac{\partial L(\theta)}{\partial \theta} = \frac{\partial\left(\frac{1}{T}\sum_{t=1}^{T}\left(\ln\sum_{h}\mathrm{e}^{-E(\boldsymbol{v}^t,\boldsymbol{h})} - \ln\sum_{v,h}\mathrm{e}^{-E(\boldsymbol{v},\boldsymbol{h})}\right)\right)}{\partial\theta}$$
$$= \frac{1}{T}\sum_{t=1}^{T}\left[E_{p(\boldsymbol{h}\mid\boldsymbol{v}^t)}\left(\frac{\partial(-E(\boldsymbol{v}^t,\boldsymbol{h}))}{\partial\theta}\right) - E_{p(\boldsymbol{v},\boldsymbol{h})}\left(\frac{\partial(-E(\boldsymbol{v},\boldsymbol{h}))}{\partial\theta}\right)\right] \tag{8}$$

式(8)梯度的计算一般采用近似方法来估计. 式(8)的两项期望可采用 Monte Carlo 思想来近似计算[2,25];式(8)中第一项期望也可直接计算,第二项只能近似计算[20],因联合分布 $p(\boldsymbol{v},\boldsymbol{h})$ 涉及可见单元和隐单元,还涉及到归一化因子,该分布很难获取. Hinton 提出对比散度(Contrastive Divergence,CD)算法[1],只需 k 步(通常 $k=1$)Gibbs 采样就可以很好的近似该分布.

本文对式(8)右边的第一项直接计算,第二项采用 Monte Carlo 思想[1,2]近似计算,于是有关 w_{ij}, a_i, b_j 三个参数的导数分别为:

$$\frac{\partial L(\theta)}{\partial w_{ij}} = \frac{1}{T}\sum_{t=1}^{T}\left[\sum_{h}p(\boldsymbol{h}\mid\boldsymbol{v}^t)\times\frac{\partial(-E(\boldsymbol{v}^t,\boldsymbol{h}))}{\partial w_{ij}} - \sum_{v,h}p(\boldsymbol{v},\boldsymbol{h})\times\frac{\partial(-E(\boldsymbol{v},\boldsymbol{h}))}{\partial w_{ij}}\right]$$
$$= \frac{1}{T}\sum_{t=1}^{T}\left[p(h_j=1\mid\boldsymbol{v}^t)\times v_i^t - \frac{1}{T}\sum_{t=1}^{T}p(h_j=1\mid\tilde{\boldsymbol{v}}^t)\times\tilde{v}_i^t\right]$$
$$= \frac{1}{T}\sum_{t=1}^{T}\left[v_i^t\times p(h_j=1\mid\boldsymbol{v}^t) - \tilde{v}_i^t\times p(h_j=1\mid\tilde{\boldsymbol{v}}^t)\right] \tag{9}$$

$$\frac{\partial L(\theta)}{\partial a_i} = \frac{1}{T}\sum_{t=1}^{T}\left[\sum_{h}p(\boldsymbol{h}\mid\boldsymbol{v}^t)\times\frac{\partial(-E(\boldsymbol{v}^t,\boldsymbol{h}))}{\partial a_i} - \sum_{v,h}p(\boldsymbol{v},\boldsymbol{h})\times\frac{\partial(-E(\boldsymbol{v},\boldsymbol{h}))}{\partial a_i}\right]$$
$$= \frac{1}{T}\sum_{t=1}^{T}v_i^t - \frac{1}{T}\sum_{t=1}^{T}\tilde{v}_i^t$$
$$= \frac{1}{T}\sum_{t=1}^{T}(v_i^t - \tilde{v}_i^t) \tag{10}$$

$$\frac{\partial L(\theta)}{\partial b_j} = \frac{1}{T}\sum_{t=1}^{T}\left[\sum_{h}p(\boldsymbol{h}\mid\boldsymbol{v}^t)\times\frac{\partial(-E(\boldsymbol{v}^t,\boldsymbol{h}))}{\partial b_j} - \sum_{v,h}p(\boldsymbol{v},\boldsymbol{h})\times\frac{\partial(-E(\boldsymbol{v},\boldsymbol{h}))}{\partial b_j}\right]$$
$$= \frac{1}{T}\sum_{t=1}^{T}\left[p(h_j=1\mid\boldsymbol{v}^t) - \frac{1}{T}\sum_{t=1}^{T}p(h_j=1\mid\tilde{\boldsymbol{v}}^t)\right]$$
$$= \frac{1}{T}\sum_{t=1}^{T}\left[p(h_j=1\mid\boldsymbol{v}^t) - p(h_j=1\mid\tilde{\boldsymbol{v}}^t)\right] \tag{11}$$

其中 $\boldsymbol{v}^t$ 表示第 t 个训练样本,v_i^t 为第 t 个训练样本第 i 个可见单元的状态;$\tilde{\boldsymbol{v}}^t$ 表示 RBM 网络中采样到的第 t 个样本,$\tilde{\boldsymbol{v}}_i^t$ 表示采样到的第 t 个样本时第 i 个可见单元的状态. 梯度近似的采样方法通常有 Gibbs 采样[1,23]、退火重要性采样[21]、PT 采样[20,22].

本文用 Gibbs 采样来近似梯度,当给定一个训练样本时,引入学习率 η,各参数更新为:

$$\Delta w_{ij} = \Delta w_{ij} + \eta\left[v_i^{(0)}p(h_j=1\mid\boldsymbol{v}^{(0)}) - v_i^{(k)}p(h_j=1\mid\boldsymbol{v}^{(k)})\right] \tag{12}$$

$$\Delta a_i = \Delta a_i + \eta\left[v_i^{(0)} - v_i^{(k)}\right] \tag{13}$$

$$\Delta b_j = \Delta b_j + \eta\left[p(h_j=1\mid\boldsymbol{v}^{(0)}) - p(h_j=1\mid\boldsymbol{v}^{(k)})\right] \tag{14}$$

3　修正动量的 RBM 算法

本文算法与文献[20]有三处不同和改进. 第一个不同之处是,在 RBM 训练和微调阶段分别用不同形式的动量项来加速网络收敛和精确找到最优点;第二个不同之处是,采用 Gibbs 采样来近似似然函数梯度;第三个不同之处是,修改隐单元偏置更新方式来取得更好的识别效果和加速效果.

首先,我们讨论隐单元偏置的更新方式. PT 采样的多个 Markov 链混合困难是一直存在的. 但当目标分布不是很复杂的时候,Gibbs 采样[23]通过单个 Markov 链就能很好的近似目标分布. 因此,本文采用 Gibbs 采样来训练 RBM 模型. 然而,Tieleman[24]提到,当 RBM 隐单元取值采用概率值时会导致模型带来密度问题如图 2(a)所示,但并不影响模型识别[24].

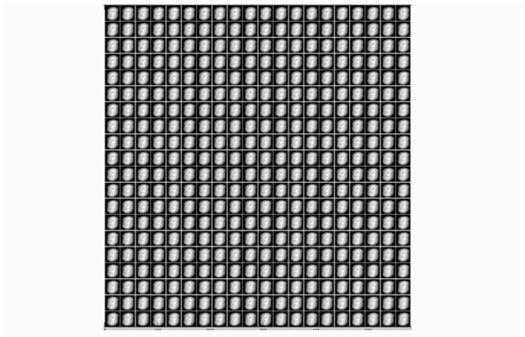
我们在实验中发现,虽不影响识别,但同样时间内,其识别效果要差些. 在 MNIST 数据库上的识别实验情况如下图 2(b)所示.

从图 2(b)中发现,的确如文献[20]所述一样,当不改进隐单元偏置 b_j 更新方式时,即使加入动量项也没起到加速作用(如图 2(b)中的绿线图走势). 当只改进偏置 b_j 的更新步长时,其效果如图 2(b)中的黑色线图走势. 受此启发,本文先增大偏置 b_j 更新步长,再加入经典动量项(如图 2(b)中红色线图所示),说明此思路可行. 从而隐单元偏置采用状态值与概率值的差进行更新,当训练一个样本时,b_j 的参数更新方式改为:

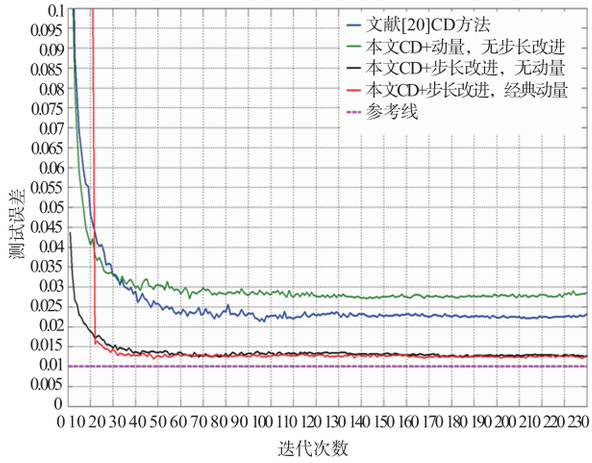
$$\Delta b_j = \Delta b_j + \eta\left[h_j^{(0)} - p(h_j=1\mid\boldsymbol{v}^{(k)})\right] \tag{15}$$

当偏置更新步长加大时,试验中得到的权重就没有密度问题,且同样时间内,其分类效果要好.

下面,我们讨论动量的选取问题. 在神经网络中,动量通常是用来加速和提高 BP 算法的[5,6]. 动量项 m^* 使得当前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,而是采用上一次参数值修改方向与本次梯度方向的组合[5]. 这种组合常见有两种,一种是采取上一次参数值修改方向与本次梯度方向相减的



(*a*)密度问题的权重图



(*b*) 隐单元的偏置采用概率值更新时的MNIST实验情况

图2

经典动量方式[8,9]：

$$\theta_t=\theta_{t-1}+m^*\Delta\theta_{t-1}-\eta\left[\frac{\partial L(\theta)}{\partial\theta_i}\right] \tag{16}$$

经典动量方法的基本思想是，用之前的参数值修改方向来修正当前梯度更新方向.

受 Nesterov 梯度方法[11]启示，有些学者发现另一动量形式即 Nesterov 动量[13]有较好加速效果. Nesterov 动量可看作在经典动量方法中添加了一个校正因子[3,13]：

$$\theta_t=\theta_{t-1}+m^*\Delta\theta_{t-1}-\eta\left[\frac{\partial L(\theta+m^*\Delta\theta_{t-1})}{\partial\theta_i}\right] \tag{17}$$

这两类动量方法能避免算法的不稳定性和对选择不当的梯度方向进行纠正. Sutskever 等[13]提出 Nesterov 动量比经典动量要快. Nitanda[14]将 Nesterov 动量和随机梯度方差相结合，也认为 Nesterov 动量效果好. Zareba 等[15]对两种动量方法做了比较，发现 Nesterov 动量和经典动量具有一样的加速效果，在 MNIST 手写体识别实验中 RBM 取得了 2.04% 的错误率. Yuan 等[16]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，动量的确有加速 RBM 网络收敛和提高学习性能的效果. 以上动量方法要么只与 SGD 结合，要么仅与 SGA 相结合，没有在 RBM 训练和微调阶段同时与 SGA 和 SGD 相结合共同作用，因此，加速效果有限.

针对以上问题，本文提出在 RBM 预训练阶段和微调阶段同时引入不同动量来加速网络收敛和提高识别效果. 此动量方法不同于以上经典动量和 Nesterov 动量方法. 本文动量采取的是上一次参数值修改方向与本次梯度方向相加的方式，因此，在 RBM 训练中采取参数更新方式是：

$$\theta_t=\theta_{t-1}+m^*\Delta\theta_{t-1}+\eta\left[\frac{\partial L(\theta)}{\partial\theta_i}\right] \tag{18}$$

结合梯度上升算法特点，在 RBM 预训练阶段，采用快速上升的动量加速形式. 当许多连续的梯度指向相同的方向时，步长最大，在梯度方向上不停的加速，可快速的达到最优点.

在 RBM 微调阶段，依据梯度下降算法特点，再引入缓慢式下降的动量形式，引入不同学习率 η'，BP 网络中参数 W'的更新取为：

$$nn.W'(n)=nn.W'(n-1)-m'\times nn.\Delta W'(n-1)-\eta'\times nn.\Delta W'(n) \tag{19}$$

在 DBN 网络微调阶段，为使梯度下降过程中能精确找到最优点，微调时采取的更新步长较小，便于不错过每个最低点. 这种采用不同动量项的方法，既可提高识别效果，又可减少网络学习时间.

实际算法中，采用批量进行训练，如每次 T 个样本进行训练，则利用 matlab 中矩阵间相乘运算优势，参数更新过程中使用参数的平均梯度，即：

$$\boldsymbol{W}=\boldsymbol{W}+m^*\cdot\Delta\boldsymbol{W}+\eta\left[\frac{1}{T}\Delta\boldsymbol{W}\right]$$

$$\boldsymbol{a}=\boldsymbol{a}+m^*\cdot\Delta\boldsymbol{a}+\eta\left[\frac{1}{T}\Delta\boldsymbol{a}\right]$$

$$\boldsymbol{b}=\boldsymbol{b}+m^*\cdot\Delta\boldsymbol{b}+\eta\left[\frac{1}{T}\Delta\boldsymbol{b}\right]$$

$$\boldsymbol{W}'=\boldsymbol{W}'-m'\cdot\Delta\boldsymbol{W}'-\eta'\left[\frac{1}{T}\Delta\boldsymbol{W}'\right]$$

采用修正动量的 k 步 Gibbs 采样 RBM 算法如算法 1 所示，得到三个参数 $\boldsymbol{W},\boldsymbol{a},\boldsymbol{b}$ 的更新方式，是为整个 BP 网络初始化做准备，RBM 对数据进行编码后，交给监督学习去分类.

算法 1 基于修正动量的 RBM 算法

输入：RBM($\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_n,\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_m$)，分批训练，每批有 T 个样本，RBM 的学习率为 η，动量为 m^*

输出：参数 $\boldsymbol{W},\boldsymbol{a},\boldsymbol{b}$

(1) for $i=1,2,\cdots,n;j=1,2,\cdots,m$，初始化 $\Delta w_{ij}=\Delta a_i=\Delta b_j=0$

(2) 对所有的样本 $\boldsymbol{v}^t,t=1,2,\cdots,T$

(3) $\boldsymbol{v}^{(0)}\leftarrow\boldsymbol{v}=\boldsymbol{x}$

(4) 对每个样本 $t=0,1,\cdots,k-1$
(5) 对 $j=1,2,\cdots,m$ 采样 $h_j^{(t)} \sim p(h_j|\boldsymbol{v}^{(t)})$
(6) End for
(7) 对 $j=1,2,\cdots,m$ 采样 $v_i^{(t+1)} \sim p(v_i|\boldsymbol{h}^{(t)})$
(8) End for
(9) for $i=1,2,\cdots,n;j=1,2,\cdots,m$
(10) $\Delta w_{ij} \leftarrow m^* \cdot \Delta w_{ij} + \eta[v_i^{(0)} p(h_j=1|\boldsymbol{v}^{(0)}) - v_i^{(k)} p(h_j=1|\boldsymbol{v}^{(k)})]$
(11) $\Delta a_i \leftarrow m^* \cdot \Delta a_i + \eta[v_i^{(0)} - v_i^{(k)}]$
(12) $\Delta b_j \leftarrow m^* \cdot \Delta b_j + \eta[h_j^{(0)} - p(h_j=1|\boldsymbol{v}^{(k)})]$
(13) End for

4　数值实验

数值实验运用 Gibbs 采样及 CD 算法,结合快速上升的动量算法来优化网络加速收敛,并同时采用缓慢下降的动量算法来微调整个网络. MNIST 手写体识别实验,分别构建 RBM 模型和两个隐层的 DBN 模型,来说明本文提出的修正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. Extended Yale B 和 CMU-PIE 人脸识别实验,来体现 RBM 模型对提出的修正动量算法能有效地增强图像特征的表达能力,提高网络的泛化能力和鲁棒性,分类效果更佳.

整个实验都是在 MATLAB R2014a 和 Microsoft Windows 8 的操作系统环境下,CPU 是 Intel(R) core(TM) i7-4770 HQ CPU @2.2GHz,内存是 16GB 来实现的. 首先初始化网络,设置参数的初始值 $w_{ij}=0, a_i=0, b_j=0$.

4.1　MNIST 数据库手写体识别实验

实验一采用 MNIST 手写体数据集,包含 70000 张 0~9 的 10 个手写数字图像,每张图片大小是 28×28,随机选取 60000 张图像用于训练,10000 张用来测试. 分批训练,每批有 100 张图片.

先构建一个 RBM 网络,隐单元个数设置为 400,即网络结构为:784-400-10,用本文动量算法,经过 10 分钟,错误率即可达到 1.33% 的效果. 当建立 784-900-10 的网络结构,其他参数的设置与 400 个隐单元时一样,经过 11 分钟,可达 1.38% 的错误率. 与其他方法的 RBM 模型结果如表 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(min)
MapReduce RBM[19]	784-900-10	2.92%	7.45
本文算法 RBM	**784-900-10**	**1.76%**	**7**
本文算法 RBM	**784-400-10**	**1.47%**	**6.5**
MapReduce RBM[19]	784-900-10	2.89%	12
本文算法 RBM	**784-900-10**	**1.38%**	**11**
动量 RBM[15]	784-400-10	2.04%	—
动量 RBM[25]	784-400-10	1.42%	10
本文算法 RBM	**784-400-10**	**1.33%**	**10**

从表 1 的结果可以看出,本文提出的动量算法的 RBM 模型学习时间和分类效果优于其他同类 RBM 算法. 比 Zhang 等[19] MapReduce RBM 方法、沈卉卉等[25] RBM 方法及 Zareba 等[15] 的两种动量方法的分类效果都有提高,错误率降低了 6% ~35%. 本文算法的优点在于简单易操作且效果好,不需要面临 Hadoop 平台的 MapReduce 框架设计等问题.

当构建 2 个隐层的 DBN 网络:784-400-400-10,经过 42 分钟,错误率就可降到 1.03%,本文算法在 MNIST 数据集上的训练误差和测试误差随着迭代次数的变化趋势如图 3 所示.

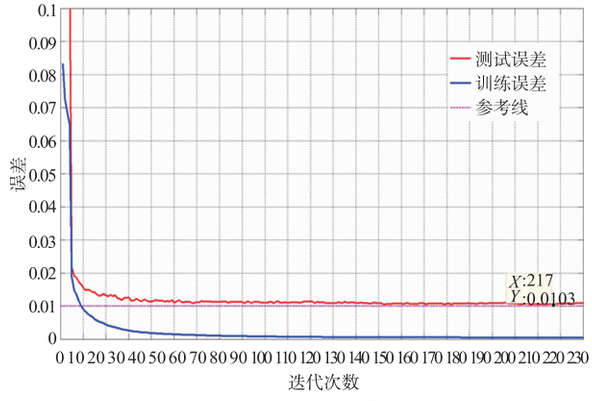


图3　本文算法在MNIST数据集上的训练误差和测试误差变化趋势

本文算法的 DBN 模型与其他有关 DBN 方法在 MNIST 数据集上实验情况的对比结果如表 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(h)
专家乘积 + 动量方法[8]	1.7%	24
Hinton 的贪婪算法[2]	1.25%	大于 9
权衰减动量方法[12]	1.1%	—
多隐层 Gibbs 采样方法[26]	1.09%	—
LogSum 方法[27]	1.02%	—
动量 DBN 方法[25]	1.02%	1.6
本文动量方法,无步长修改	**2.77%**	**0.75**
本文 CD + 步长修改,无动量	**1.28%**	**0.8**
本文 CD + 步长修改 + 经典动量	**1.20%**	**0.76**
本文算法	**1.03%**	**0.7**

表 2 的结果显示,本文提出的 DBN 算法取得了 1.03% 的错误率,其结果略次于 Logsum 方法[27] 和文献[25]动量方法的 1.02%,但时间减少 50% 以上,从网络结构隐单元设置层数和个数上来看,本文修正动量算法比文献[27]和[25]的计算效率要高. 其构成 2 个隐层 DBN 分类效果也优于其他算法,无论识别效果还是效率都优于史科等[26]4 个隐层的 Gibbs 采样方法. 不同 DBN 方法在 MNIST 测试数据集上,其测试误差表现如下图 4 所示. 为便于比较,图 4 中不同 DBN 的参数设置都是在各网络最优情况下设置的,与文献[25]方法相

比,其效果很相似,但是在相同的时间内,本文算法只需50次训练,而文献[25]需100次训练,且其最后一层隐单元个数设置得要足够多才能很好的表达其学习到的特征用于分类,所以文献[25]算法要达到最优需要的时间会长些,本文算法会较早收敛到最优点,略优于文献[25]方法. 当修改隐单元偏置的参数更新步长,而训练和微调都不用动量项,其他参数设置不变,识别错误率为1.28%;本文提出算法与用本文动量而不改隐单元偏置更新步长的方法相比,错误率可降低约63%,由此可见修改隐单元偏置的更新方式很重要. 试验结果表明,本文修正动量算法在DBN网络中也优于其他同类算法.

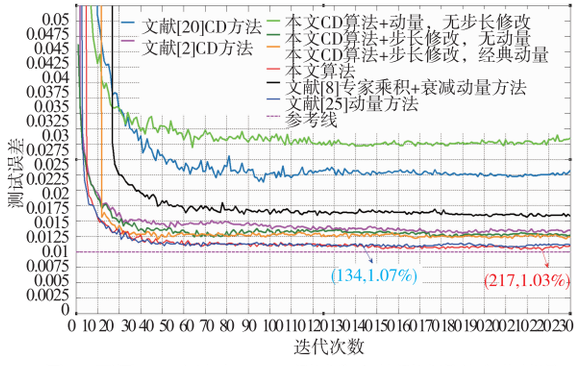


图4 不同DBN方法在MNIST测试集上的测试误差变化情况

4.2 Extended Yale B 数据库人脸识别实验

实验二采用Extended Yale B人脸数据库,该数据库包含38个人的2,414张不同光照和表情的人脸图像. 我们将所有图像剪辑成大小为32×32人脸图像进行分类,随机选取2100张图片来训练,剩下的314张图片用来测试,分批训练,每批样本为100张图片. 构建1024-600-600-38的两个隐层DBN网络,采用本文算法来识别38人的人脸图像,经过6分钟,50次训练和600多次的微调,识别错误率达到2.67%. 与郭继昌等[28]的Fisher约束方法和Tu等[29]的方法同在Extended Yale B人脸数据库上的实验结果相比如下表3所示.

表3 不同方法在Extended Yale B人脸数据集上的实验结果

不同模型算法	识别率	耗时(min)
Fisher约束方法[28]	97.27%	53.6
光照补偿方法[29]	89.4%	—
本文CD算法,无动量	**96%**	**6.2**
本文算法	**97.33%**	**6**

由此可看出本文算法识别结果97.33%略优于文献[28]的Fisher约束方法取得的97.27%,且其计算效率上有明显的优势;比文献[29]的识别结果提高了8.87%.

4.3 CMU-PIE 数据库人脸识别实验

实验三采用CMU-PIE数据库,该数据库包含68位志愿者的41,368张多姿态、光照和表情的面部图像. 选取其中尺寸为32×32的子集,包含68个人,每个人约有170张不同的图像,共11554张图片. 实验分别进行对30人和68人的人脸图像进行识别,30人时,选取5100个带标数据包含30个人的图像,其中随机选取4500张图像来训练,其余600张图片用来测试;68个人时,选10000张图像作为训练样本,剩下的1554张图片用来测试.

分别建立1024-600-600-30和1024-100-100-68的网络识别30人和68人的人脸图像,用本文算法,正确识别率分别可达99.17%和98.67%,不同方法在CMU-PIE数据库上的人脸识别结果如表4所示.

表4 不同方法在CMU-PIE人脸数据集上的实验结果对比

不同模型算法	分类识别率	耗时(min)
特征聚类SAE方法,30人识别[30]	98.83%	17.7
本文CD算法,无动量,30识别	97.33%	10
本文算法,30人识别	**99.17%**	**9.5**
CNN+SAE方法,68人识别[31]	96.17%	—
动量DBN方法,68人识别[25]	98.47%	6
本文CD算法,无动量	**97.93%**	**6.5**
本文算法	**98.67%**	**6**

因30人的数据量少且类别也少,其识别率99.17%比68人的人脸识别率98.67%要稍好. 与文献[30]相比,本文算法对30人的识别率提高了0.34%,且时间上缩短了46%. 与文献[31]、[25]相比,本文算法对68人的识别率分别提高了2.5%和0.2%. 说明本文提出的动量算法能够有效地增强RBM模型对图像本质特征的提取能力,其网络泛化能力和鲁棒性都较好.

综合以上4组实验分析,本文提出的修正动量算法在图像识别中有效. 构建的RBM网络不同阶段使用不同的修正动量项,与其他方法相比,可使得优化后网络在同样时间内,其分类效果更好.

5 结论

本文提出了一种基于修正动量的RBM算法. 针对RBM梯度的计算,采用第一项利用样本直接计算,第二项利用Gibbs采样近似计算,但当隐单元的取值采用概率值时会导致模型识别效果不理想,动量项也无法起到很好的加速作用,还会带来密度问题,于是,本文修改了隐单元偏置的更新方式来避免以上问题. 再结合不同方式不同效用的动量项加入到RBM训练和微调阶段,共同作用以加速网络收敛和提高分类效果. 此算法在MNIST数据集,Extended Yale B和CMU-PIE人脸数据库上分别进行了实验与分析,并与同类RBM算法,DBN算法做比较. 试验结果表明,本文提出的修正动量算法在图像识别上快速且有效,尤其在人脸识别上效果较为明显,说明RBM梯度计算与提出的修正动量项

相结合,能使 RBM 具有较强的特征表达能力和分类能力,不同动量优化后的网络具有很好的泛化能力和鲁棒性.

参考文献

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

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

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

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

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

[6] Qian N. On the momentum term in gradient descent learning algorithms[J]. Neural Networks, 1999, 12(1): 145 – 151.

[7] Attoh-Okine N O. Analysis of learning rate and momentum term in back-propagation neural network algorithm trained to predict pavement performance[J]. Advances in Engineering Software, 1999, 30(4):291 – 302.

[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. A Practical Guide to Training Restricted Boltzmann Machines[R]. Berlin, Springer: Neural Networks, Tricks of the Trade (2nd ed), 2012. 599 – 619.

[10] Hamid N A, Nawi N M, Ghazali R, Mohd Salleh M N. Accelerating learning performance of back propagation algorithm by using adaptive gain together with adaptive momentum and adaptive learning rate on classification problems[J]. International Journal of Software Engineering and Its Applications, 2011, 5(4):31 – 44.

[11] Nesterov Y. A method of solving a convex programming problem with convergence rate O(1/sqr(k))[J]. Soviet Mathematics Doklady, 1983, 27(2):372 – 376.

[12] Swersky K, Chen B, Marlin B. A tutorial on stochastic approximation algorithms for training restricted Boltzmann machines and deep belief nets[A]. Information Theory and Applications Workshop (ITA)[C]. USA: IEEE Press, 2010. 80 – 89.

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

[14] Nitanda A. Stochastic proximal gradient descent with acceleration techniques[A]. Proceedings Advances in Neural Information Processing Systems[C]. Canada: MIT Press, 2014. 1574 – 1582.

[15] 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: Springer, 2015. 187 – 192.

[16] 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.

[17] 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)

[18] 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: IEEE, 2012. 10 – 15.

[19] 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.

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

[21] Salakhutdinov R, Murray I. On the quantitative analysis of deep belief networks[A]. Proceedings of the International Conference on Machine Learning[C]. Helsinki, Finland: ACM, 2008. 872 – 879.

[22] Desjardins G, Courville A, Bengio Y, Vincent P, Dellaleau O. Parallel tempering for training of restricted Boltzmann machines[A]. Proceedings of the International Conference on Artificial Intelligence and Statistics[C]. Italy: JMLR, 2010. 145 – 152.

[23] Hastings W K. Monte Carlo sampling methods using Markov chains and their applications[J]. Biometrika, 1970, 57(1):97 – 109.

[24] Tieleman T. Training restricted Boltzmann machines using approximations to the likelihood gradient[A]. Proceedings of the Twenty-first International Conference on Machine Learning (ICML)[C]. New York, USA: ACM, 2008. 1064 – 1071.

[25] 沈卉卉,李宏伟. 基于动量方法的受限玻尔兹曼机的一

种有效算法[J]. 电子学报,2019,47(1):176 – 182.
Shen Hui-hui, Li Hong-wei. An effective algorithm of restricted Boltzmann machine based on momentum method [J]. Acta Electronica Sinica, 2019, 47(1): 176 – 182. (in Chinese)

[26] 史科,陆阳,等. 基于多隐层 Gibbs 采样的深度信念网络训练方法[DB/OL]. 自动化学, 2018-10-11. doi: 10.16383/j.aas.c170669.
Shi Ke, Lu Yang, et al. A deep belief networks training strategy based on multi hidden layer Gibbs sampling[DB/OL]. Acta Automatica Sinica, 2018-10-11. doi: 10.16383/j.aas.c170669. (in Chinese)

[27] Ji N N, Zhang J S, Zhang C X, et al. Enhancing performance of restricted Boltzmann machines via log-sum regularization[J]. Knowledge-Based Systems, 2014, 63(1): 82 – 96.

[28] 郭继昌,张帆,王楠. 基于 Fisher 约束和字典对的图像分类[J]. 电子与信息学报,2017,39(2):270 – 277.
Guo Ji-chang, Zhang Fan, Wang Nan. Image classification based on Fisher constraint and dictionary pair[J]. Journal of Electronics & Information Technology, 2017, 39(2): 270 – 277. (in Chinese)

[29] Tu X, Gao J, Xie M, Qi J, Ma Z. Illumination normalization based on correction of large-scale components for face recognition [J]. Neurocomputing, 2017, 266(C): 465 – 476.

[30] 付晓,沈远彤,付丽华,杨迪威. 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, Yang Di-wei. An optimized sparse auto-encoder network based on feature clustering[J]. Acta Electronica Sinica, 2018, 46(5): 1041 – 1046. (in Chinese)

[31] 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

刘国武 男,1965 年生于湖北随州. 现为湖北经济学院教授,硕士生导师.
E-mail: luckyly2000@sohu.com

付丽华 女,1979 年生于湖北枝江. 现为中国地质大学(武汉)数理学院教授,硕士生导师. 研究方向为信息处理与智能计算.
E-mail: lihuafu@cug.edu.cn

刘智慧 女,1979 年生于湖南沅江. 现为中国地质大学(武汉)数理学院副教授,硕士生导师. 研究方向为信号与信息处理.
E-mail: zhhliu@cug.edu.cn

李宏伟(通讯作者) 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.
E-mail: hwli@cug.edu.cn