基于受限玻尔兹曼机的专家乘积系统的一种改进算法

沈卉卉[①②③] 李宏伟*[①③]

[①](中国地质大学数理学院 武汉 430074)

[②](湖北经济学院信息管理与统计学院 武汉 430205)

[③](中国地质大学地球内部多尺度成像湖北省重点实验室 武汉 430074)

摘 要：深度学习在高维特征向量的信息提取和分类中具有很强的能力，但深度学习训练时间也比较长，超参数搜索空间大，从而导致超参数寻优较困难。针对此问题，该文提出一种基于受限玻尔兹曼机(RBM)专家乘积系统的改进方法。先将专家乘积系统原理与RBM算法相结合，采用全是真实概率值的参数更新方式会引起模型识别效果不理想和带来密度问题，为此将其更新方式进行改进；为加快网络收敛和提高模型识别能力，采取在RBM预训练阶段和微调阶段引入不同组合方式动量项的一种改进算法。通过对MNIST数据库中的0~9的手写数字体的识别和CMU-PIE数据库的人脸识别实验，提出的算法减少了学习时间，提高了超参数寻优的效率，进而构建的深层网络能获得较好的分类效果。试验结果表明，提出的改进算法在处理高维大量的数据时，计算效率有较大提高，其算法有效。

关键词：深度学习；专家乘积；神经网络；受限玻尔兹曼机；动量

中图分类号：TP182; TP183 **文献标识码**：A **文章编号**：1009-5896(2018)09-2173-09

DOI: 10.11999/JEIT170880

An Improved Algorithm of Product of Experts System Based on Restricted Boltzmann Machine

SHEN Huihui[①②③] LI Hongwei[①③]

[①](*Institute of Geophysics & Geomatics*, *China University of Geosciences*, *Wuhan* 430074, *China*)

[②](*School of Statistics & Information Management*, *Hubei University of Economics*, *Wuhan* 430205, *China*)

[③](*Hubei Subsurface Multi-scale Imaging Key Laboratory*, *China University of Geosciences*, *Wuhan* 430074, *China*)

Abstract: Deep learning has a strong ability in the high-dimensional feature vector information extraction and classification. But the training time of deep learning is so long that the optimal hyper-parameters combination can not be found in a short time. To solve these problems, a method of product of experts system based on Restricted Boltzmann Machine (RBM) is proposed. The product of experts theory is combined with the RBM algorithm and the parameter updating way is all adopted the probability value, which leads to the undesirable recognition effect and slightly worse density models, so the parameter updating way is improved. An improved algorithm with momentum terms in different combinations is used not only in the RBM pre-training phase but also in the fine-tuning stage for both classification accuracy enhancement and training time decreasing. Through the recognition experiments on the MNIST database and CMU-PIE face database, the proposed algorithm reduces the training time, and improves the efficiency of hyper-parameters optimization, and then the deep belief network can achieve better classification performance. The result shows that the improved algorithm can improve both accuracy and computation efficiency in dealing with high-dimensional and large amounts of data, the new method is effective.

Key words: Deep learning; Product Of Experts (POE); Neural network; Restricted Boltzmann Machine (RBM); Momentum

收稿日期：2017-09-18；改回日期：2018-05-24；网络出版：2018-07-12

*通信作者： 李宏伟 hwli@cug.edu.cn

基金项目：湖北省教育厅科学技术研究计划重点项目(D20182203)

Foundation Item: The Science and Technology Research Program Key Project of Hubei Provincial Education Department (D20182203)

1 引言

专家系统是一类具有专门知识和经验的计算机智能程序系统，它是人工智能的一个重要分支[1]。专家系统应用人工智能技术和计算机技术，根据某领域1个或多个专家提供的知识和经验，进行推理和判断，模拟人类专家的决策过程[1]。其核心结构是知识库和推理机。神经网络通过自动学习获取知识的特点能解决专家系统实现过程中的一些难点问题[2]。比如，分类识别[1,3]、办公自动化[2]等问题。神经网络专家系统知识的获取无需人工整理和总结专家知识，而是通过网络学习自动获取知识[2]；神经网络专家系统的推理机制是一种并行计算过程。神经网络专家系统与传统的专家系统相比在知识获取、并行推理、适应性学习、容错能力等方面有明显的优越性[2]。因此，基于神经网络模型的专家系统是一种有效的综合智能系统，其核心在于网络学习算法的改进，从而使神经网络及其专家系统具有更高的智能水平。

受限玻尔兹曼机(Restricted Boltzmann Machine, RBM)是深度学习中的经典模型之一[4]。由于RBM表示力强、易于推理等优点被成功用于深度神经网络中。RBM学习算法有对比散度算法(Contrastive Divergence, CD)[5]和专家乘积的方法(Product Of Experts, POE)[3]等。RBM在无监督特征学习方面具有很强的能力[4]，但多个RBM的堆叠而构成的深度信念网络(Deep Belief Nets, DBN)的学习时间会相应地加长。

Mayraz等人[3]提出专家乘积方法，结合权衰减算法，在MNIST数据库上做识别实验，耗时太长。罗剑江等人[6]和王岳青等人[7]都提出新的RBM算法，识别精度有所提高，但模型中的超参数增多，人工设置较麻烦以及面临深度学习并行编程框架的设计等问题。Zhang等人[8]提出分布式学习算法MapReduce RBM，在时间上有提高，但精度有待进一步提高，且要面临Hadoop平台的MapReduce框架设计等问题。

针对深度神经网络学习时间长的问题，Polyak[9]提出动量方法，在随机梯度方法中引入动量项[9–13]，的确有加速网络收敛的效果。动量方法有经典动量[10,13]和Nesterov动量[11]这两种动量方式，虽然分别都取得了一定的效果，但分类精度有待进一步提高。

本文提出基于RBM的专家乘积系统的一种改进算法。主要做了3个工作：首先，将专家乘积系统原理进行详细推导和梳理，并与RBM学习算法相结合，采用全是真实概率值的参数更新方式，但会导致模型识别效果不理想以及带来密度问题；因此，接着，修改了其中一个偏置的参数更新方式，使其在相同的时间内得到较好的分类效果且消除权重的密度问题；最后，为了加速网络收敛，提出一种改进的动量算法，对RBM的参数更新方式做了不同于以上两种动量方式[10,11]的算法改进，同时微调整个网络也用不同的动量算法。在MNIST和CMU-PIE数据库上，都取得了不错的识别效果。与RBM同类传统算法相比，实验结果表明，该算法在运行时间和识别精度上具有一定的优势。

2 专家乘积系统原理

专家系统与神经网络结合起来能解决专家系统中知识获取的瓶颈问题[2]。且RBM神经网络有极强的表示能力和学习能力，对于新的模式和样本可通过权值的改变进行学习、记忆、联想和存储，进而在以后的运行中能判断这些新的模式。

2002年Hinton[5]提出专家乘积系统，它将每个独立的专家系统模型的概率相乘并规范化。专家乘积系统模型是由多个专家系统组成的概率模型，所有n个专家模型组合如式(1)：

$$p(\boldsymbol{d}\,|\theta_1,\theta_2,\cdots,\theta_n)=\frac{\prod_m p_m(\boldsymbol{d}\,|\theta_m)}{\sum_{\boldsymbol{c}}\prod_m p_m(\boldsymbol{c}\,|\theta_m)} \tag{1}$$

式中，$\boldsymbol{d}$表示离散数据向量，θ_m表示模型m的参数，$p_m(\boldsymbol{d}\,|\theta_m)$表示数据$\boldsymbol{d}$在模型$m$下的条件概率，$\boldsymbol{c}$表示数据空间所有数据向量的索引，对于一个连续的空间，其和可用近似的常数来代替[5]。

为了找到符合独立同分布的一组观察向量的专家乘积系统，对式(1)取对数求导得

$$\begin{aligned}&\frac{\partial[\ln p(\boldsymbol{d}\,|\theta_1,\theta_2,\cdots,\theta_n)]}{\partial\theta_m}\\&=\frac{\partial[\ln p_m(\boldsymbol{d}\,|\theta_m)]}{\partial\theta_m}-\sum_{\boldsymbol{c}}p(\boldsymbol{c}\,|\theta_1,\ \theta_2,\cdots,\theta_n)\\&\quad\cdot\frac{\partial[\ln p_m(\boldsymbol{c}\,|\theta_m)]}{\partial\theta_m}\end{aligned} \tag{2}$$

式(2)右边的第2项是一个期望值，$\boldsymbol{c}$是从专家乘积系统中生成的理想数据。则最大的困难是在此专家乘积系统下如何产生理想数据的概率分布。对于离散的数据，Hinton用吉布斯采样(Gibbs)的马尔可夫链蒙特卡洛方法来产生理想数据的概率分布[3]。

专家乘积系统目的是得到理想的数据分布，使得专家乘积系统达到平衡时在可见层产生的理想的数据分布P_θ^∞与观察的数据向量的概率分布P^0更接近。从信息熵的角度，也就是使KL($P^0||P_\theta^\infty$)距离

最小，两种分布越接近[14]。P^0和P_θ^∞之间的KL距离定义为[14]

$$\begin{aligned}\mathrm{KL}(P^0\,\|P_\theta^\infty)&=\sum_{\boldsymbol{d}}P^0(\boldsymbol{d})\ln\frac{P^0(\boldsymbol{d})}{P_\theta^\infty(\boldsymbol{d})}\\&=\sum_{\boldsymbol{d}}P^0(\boldsymbol{d})\ln P^0(\boldsymbol{d})\\&\quad-\sum_{\boldsymbol{d}}P^0(\boldsymbol{d})\ln P_\theta^\infty(\boldsymbol{d})\\&=-H(P^0(\boldsymbol{d}))\\&\quad-E_{P^0(\boldsymbol{d})}\left[\ln P_\theta^\infty(\boldsymbol{d})\right]\end{aligned}\tag{3}$$

其中，$P^0(\boldsymbol{d})$为观察的数据向量的概率分布；$P_\theta^\infty(\boldsymbol{d})$为样本空间的分布，即当专家乘积系统达到平衡时可见层产生的理想的数据分布。

对式(3)两边对θ_m求导：

$$\begin{aligned}\frac{\partial\left[\mathrm{KL}\left(P^0\,\|P_\theta^\infty\right)\right]}{\partial\theta_m}&=\frac{\partial\left[-H\left(P^0\right)\right]}{\partial\theta_m}\\&\quad-\frac{\partial\left\{E_{P^0}\left[\ln P_\theta^\infty\left(\boldsymbol{d}\right)\right]\right\}}{\partial\theta_m}\\&=-E_{P^0}\left[\frac{\partial[\ln P_\theta^\infty(\boldsymbol{d})]}{\partial\theta_m}\right]\end{aligned}$$

$H(P^0)$表示观察数据分布的熵，由于P^0不依赖于模型参数θ_m，所以在优化过程中$H(P^0)$可以被忽略，$P_\theta^\infty(\boldsymbol{d})$是式(1)中$p(\boldsymbol{d}\,|\theta_1,\theta_2,\cdots,\theta_n)$的另一种写法，则式(2)，在观察数据向量的数据分布上的平均值为，也即对式(2)两边求期望：

$$\begin{aligned}E_{P^0}\left[\frac{\partial[\ln P_\theta^\infty(\boldsymbol{d})]}{\partial\theta_m}\right]&=E_{P^0}\left\{\frac{\partial[\ln p_m(\boldsymbol{d}\,|\theta_m)]}{\partial\theta_m}\right.\\&\quad-\sum_{\boldsymbol{c}}p(\boldsymbol{c}\,|\theta_1,\theta_2,\cdots,\theta_n)\\&\quad\left.\cdot\frac{\partial[\ln p_m(\boldsymbol{c}\,|\theta_m)]}{\partial\theta_m}\right\}\\&=E_{P^0}\left\{\frac{\partial[\ln p_m(\boldsymbol{d}\,|\theta_m)]}{\partial\theta_m}\right.\\&\quad\left.-E_{P_\theta^\infty}\left[\frac{\partial[\ln p_m(\boldsymbol{c}\,|\theta_m)]}{\partial\theta_m}\right]\right\},\\-\frac{\partial[\mathrm{KL}(P^0||P_\theta^\infty)]}{\partial\theta_m}&=E_{P^0}\left[\frac{\partial[\ln P_\theta^\infty(\boldsymbol{d})]}{\partial\theta_m}\right]\\&=E_{P^0}\left\{\frac{\partial[\ln p_m(\boldsymbol{d}\,|\theta_m)]}{\partial\theta_m}\right\}\\&\quad-E_{P_\theta^\infty}\left\{\frac{\partial[\ln p_m(\boldsymbol{c}\,|\theta_m)]}{\partial\theta_m}\right\}\end{aligned}\tag{4}$$

用相对偏差来代替$P^0\,\|P_\theta^\infty$之间的偏差，经过专家乘积系统进行一次完整的Gibbs采样后对观察数据向量$\boldsymbol{d}$进行一步重构出的数据向量$\hat{\boldsymbol{d}}$的分布P_θ^1作为一个桥梁，即用最小化$P^0\,\|P_\theta^\infty$与$P_\theta^1\,\|P_\theta^\infty$之间的偏差来代替最小化$P^0\,\|P_\theta^\infty$之间的偏差，因Gibbs采样达平衡状态需很长时间，只要保证系统能不断向正确的方向更新，就可用相对偏差来代替初始状态和最终状态之间的偏差[5]。因为相对偏差是一个非负值，在马尔科夫链中所有的转换都可得到一非零的概率值，这样就保证数据不断趋向完美。

根据式(4)，可以得到相对偏差的式(5)：

$$\begin{aligned}&-\frac{\partial\left\{\mathrm{KL}\left[\left(P^0\,\|P_\theta^\infty\right)-\left(P_\theta^1\,\|P_\theta^\infty\right)\right]\right\}}{\partial\theta_m}\\&\quad=E_{P^0}\left\{\frac{\partial[\ln p_m(\boldsymbol{d}\,|\theta_m)]}{\partial\theta_m}\right\}\\&\qquad-E_{P_\theta^1}\left\{\frac{\partial[\ln p_m(\hat{\boldsymbol{d}}\,|\theta_m)]}{\partial\theta_m}\right\}-\frac{\partial\left[H\left(P_\theta^1\right)\right]}{\partial\theta_m}\end{aligned}\tag{5}$$

Hinton(2002)证明式(5)右边的第3项的值相对前面两项的值很小，可以忽略[5]。因此，式(5)的对比散度(相对偏差)简写为

$$\begin{aligned}\Delta\theta_m&=E_{P^0}\left\{\frac{\partial[\ln p_m(\boldsymbol{d}\,|\theta_m)]}{\partial\theta_m}\right\}\\&\quad-E_{P_\theta^1}\left\{\frac{\partial[\ln p_m(\hat{\boldsymbol{d}}\,|\theta_m)]}{\partial\theta_m}\right\}\end{aligned}\tag{6}$$

实践表明这种算法的效果非常好，我们可以随机初始化专家乘积系统，尽管这时初始的概率分布并不理想，但是使用式(6)的学习算法后，可以获得理想数据的概率分布[3]。

3　受限玻尔兹曼机

RBM是Smolensky[15]提出的一种概率模型。RBM模型如图1所示，有一个可见层，一个隐层，层内无连接，所以，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。因此，尽管RBM所表示的分布无法有效计算，但通过Gibbs采样可以得到服从RBM所表示分布的随机样本。文献[16]证明，只要隐单元的数目足够多，RBM能够拟合任意离散分布。

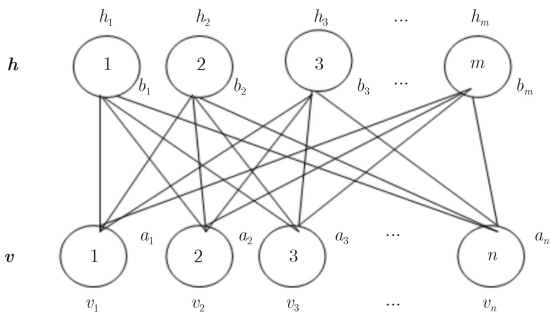


图 1 RBM网络结构示意图

在图1中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。用向量$\boldsymbol{v}=(v_1,v_2,\cdots,v_n)^{\mathrm{T}}$和$\boldsymbol{h}=(h_1,h_2,\cdots,h_m)^{\mathrm{T}}$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；$\boldsymbol{W}=(w_{ij})_{m\times n}\in\mathbf{R}^{m\times n}$表示连接权重的矩阵。令$\theta=\{w_{ij},a_i,b_j\}$表示模型中未知参数的组合，RBM的任务就是求出这些参数θ以拟合给定训练数据的分布。

那么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[17]

$$E(\boldsymbol{v},\boldsymbol{h})=-\sum_{i=1}^{n}a_iv_i-\sum_{j=1}^{m}b_jh_j-\sum_{i=1}^{n}\sum_{j=1}^{m}v_iw_{ij}h_j \tag{7}$$

$\forall i,j$，有$v_i,h_j\in\{0,1\}$。

基于该能量函数，定义给定状态$(\boldsymbol{v},\boldsymbol{h})$的联合概率分布为[17]

$$p(\boldsymbol{v},\boldsymbol{h})=\frac{\mathrm{e}^{-E(\boldsymbol{v},\boldsymbol{h})}}{Z_\theta} \tag{8}$$

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

RBM模型只定义了式(7)和式(8)，根据RBM的性质和学习RBM的目标，调整参数θ，使得参数θ下RBM网络表示的边缘概率分布尽可能地与训练数据的概率分布相符合。

文献[18]已证明在RBM中，生成的可见层向量的概率与每个隐单元独立作用而产生的可见层向量(重构的可见层数据向量)的概率乘积成正比。因此，可以把一个RBM看成是一个特殊的专家乘积系统[3,18]，它的每个隐单元都被作为一个独立的专家系统。每个专家对可见层节点的状态产生的影响不够强，但所有专家的观察结果连乘起来就足够强大了。我们最关心的是观测数据$\boldsymbol{v}$的概率分布$p(\boldsymbol{v})$，它是RBM模型$p(\boldsymbol{v},\boldsymbol{h})$的边缘分布。

$$\begin{aligned}p(\boldsymbol{v})&=\sum_{\boldsymbol{h}}p(\boldsymbol{v},\boldsymbol{h})=\sum_{\boldsymbol{h}}\frac{\mathrm{e}^{-E(\boldsymbol{v},\boldsymbol{h})}}{Z_\theta}=\frac{1}{Z_\theta}\sum_{\boldsymbol{h}}\mathrm{e}^{-E(\boldsymbol{v},\boldsymbol{h})}\\&=\frac{1}{Z_\theta}\sum_{\boldsymbol{h}}\mathrm{e}^{\sum_{i=1}^{n}a_iv_i+\sum_{j=1}^{m}b_jh_j+\sum_{i=1}^{n}\sum_{j=1}^{m}v_iw_{ij}h_j}\\&=\frac{1}{Z_\theta}\sum_{h_1}\sum_{h_2}\cdots\sum_{h_m}\mathrm{e}^{\sum_{i=1}^{n}a_iv_i+\sum_{j=1}^{m}h_j\left(b_j+\sum_{i=1}^{n}v_iw_{ij}\right)}\\&=\frac{1}{Z_\theta}\cdot\prod_{i=1}^{n}\mathrm{e}^{a_iv_i}\cdot\prod_{j=1}^{m}\left(1+\mathrm{e}^{b_j+\sum_{i=1}^{n}v_iw_{ij}}\right)\end{aligned}$$

可见$p(\boldsymbol{v})$是由独立专家乘积组成，称为专家乘积系统(Product Of Experts, POE)[5]，在学习的过程中利用Gibbs采样和KL偏差的方法，使高维模型获得更理想的概率分布[5]。

当训练一个样本，第t个样本$\boldsymbol{v}^t$时，RBM模型中参数w_{ij}更新方式是[17]

$$\Delta w_{ij}=E_{p(\boldsymbol{h}|\boldsymbol{v}^t)}(\boldsymbol{v}_i^th_j)-E_{p(\boldsymbol{v},\boldsymbol{h})}(v_ih_j) \tag{9}$$

式(2)专家乘积系统的算法等价于RBM的学习算法式(9)[3]。考虑数据$\boldsymbol{d}$的对数似然函数对第i个可见层单元到第j个隐单元的连接权值w_{ij}求导，则式(2)右侧的第1项变为

$$\frac{\partial[\ln p_j(\boldsymbol{d}\,|\,\boldsymbol{W}_j)]}{\partial w_{ij}}=E_{\boldsymbol{d}}[v_ih_j]-E_{P_\theta^\infty(j)}[v_ih_j] \tag{10}$$

其中，$\boldsymbol{W}_j$表示第j个隐单元到所有可见层的权值向量，$E_{\boldsymbol{d}}[v_ih_j]$代表期望值，其中$\boldsymbol{d}$表示固定在可见层观察向量，$v_i$表示数据$\boldsymbol{d}$在第$i$个可见层中的状态，$h_j$表示根据$\boldsymbol{d}$决定的后验概率采样得到的第$j$个隐层状态。$E_{P_\theta^\infty(j)}[v_ih_j]$表示使用Gibbs采样后第$j$个隐单元与可见层之间反复交替采样达到平衡状态时的概率分布的平均值。

同理，式(2)右侧的第2项也可以写成

$$\begin{aligned}&\sum_{\boldsymbol{c}}p(\boldsymbol{c}\,|\,\boldsymbol{W})\frac{\partial[\ln p_j(\boldsymbol{c}\,|\,\boldsymbol{W}_j)]}{\partial w_{ij}}\\&\quad=E_{P_\theta^\infty}\left[\frac{\partial[\ln p_j(\boldsymbol{c}\,|\,\boldsymbol{W}_j)]}{\partial w_{ij}}\right]\\&\quad=E_{P_\theta^\infty}[v_ih_j]-E_{P_\theta^\infty(j)}[v_ih_j]\end{aligned} \tag{11}$$

其中，$E_{P_\theta^\infty}[v_ih_j]$表示对所有的可见层单元和隐层单元进行Gibbs采样达到平衡后的期望值。

用式(10)减去式(11)，得到式(2)对具体参数w_{ij}的导数为

$$\begin{aligned}\frac{\partial[\ln P_\theta^\infty(\boldsymbol{d})]}{\partial w_{ij}}&=\frac{\partial[\ln p_j(\boldsymbol{d}\,|\,\boldsymbol{W}_j)]}{\partial w_{ij}}\\&\quad-\sum_{\boldsymbol{c}}p(\boldsymbol{c}\,|\,\boldsymbol{W})\frac{\partial[\ln p_j(\boldsymbol{c}\,|\,\boldsymbol{W}_j)]}{\partial w_{ij}}\\&=E_{\boldsymbol{d}}[v_ih_j]-E_{P_\theta^\infty}[v_ih_j]\end{aligned} \tag{12}$$

从而，由式(5)和式(10)、式(11)、式(12)进一步可得：

$$\begin{aligned}&-\frac{\partial\{\mathrm{KL}[(P^0\,\|\,P_\theta^\infty)-(P_\theta^1\,\|\,P_\theta^\infty)]\}}{\partial w_{ij}}\\&\quad=-\frac{\partial\{\mathrm{KL}[(P^0\,\|\,P_\theta^\infty)]\}}{\partial w_{ij}}+\frac{\partial\{\mathrm{KL}[(P_\theta^1\,\|\,P_\theta^\infty)]\}}{\partial w_{ij}}\\&\quad=E_{P^0}[v_ih_j]-E_{P_\theta^1}[v_ih_j]\end{aligned} \tag{13}$$

则式(13)与RBM学习算法式(9)等价[5]。其中

$E_{P_\theta^1}[v_i h_j]$表示第1次重构数据的期望值。

依据概率分布特性和专家乘积系统原理，及文献[19]用简单的平均场方法，即全部用真实的概率值来代替随机的二值状态来近似达到系统平稳时的分布[5,19]。又因专家乘积系统将每个独立模型的概率相乘并重新规范化，像素强度被规范化为在0与1之间，可被视为概率，则对式(13)中的所有二值状态值不再采用，全部用概率值来代替状态值，并用蒙特卡洛思想[17]求式(13)右边的两个期望值，从而式(13)改写为

$$\begin{aligned}&-\frac{\partial\{\mathrm{KL}[(P^0\,\|P_\theta^\infty)-(P_\theta^1\,\|P_\theta^\infty)]\}}{\partial w_{ij}}\\&\approx E_{P^0}[p_ip_j]-E_{P_\theta^1}[p_ip_j]\\&\approx p_ip_j-\widehat{p}_i\widehat{p}_j\end{aligned}\tag{14}$$

4　改进的动量算法

针对式(14)，假设只有一个观察数据向量规范化后$\boldsymbol{v}^{(0)}=\boldsymbol{d}\in[0,1]$，则式(14)中的4个概率值分别为：$p_i=\boldsymbol{v}^{(0)}, p_j=p\left(h_j=1|\boldsymbol{v}^{(0)}\right)$; $\widehat{p}_i=p\left(v_i=1|\boldsymbol{h}^{(0)}\right)$, $\widehat{p}_j=p\left(h_j=1|\boldsymbol{v}^{(1)}\right)$。其中$\boldsymbol{h}^{(0)}$是初始观察数据$\boldsymbol{v}^{(0)}$通过Gibbs采样计算得到隐层的第1次数据分布；$\boldsymbol{v}^{(1)}$是对$\boldsymbol{v}^{(0)}$的第1次重构的数据分布；$\boldsymbol{h}^{(1)}$是对隐层$\boldsymbol{h}^{(0)}$的一次重构数据分布。

为了解决分类效果不理想和权重的密度问题，于是，结合CD算法，改进偏置b_j的更新步长。RBM训练过程是在高维曲面上寻找全局最优解的过程，若两个偏置a_i, b_j的更新步长要稍微长点，每经过一次训练，搜寻点应该更加靠近最优点所在的区域范围，偏置的更新步长的加长有利于将搜寻范围限制在该范围内，而不至于跳出这个搜索圈，逐渐缩小搜索范围，最终找到全局最优解所对应的点，则网络收敛。3个参数w_{ij}, a_i, b_j中，主要是为了得到较好的权重w_{ij}来学习和提取数据样本的特征，因此根据模型需要，对隐单元的偏置b_j的更新方式做如下改进，使其更新步长加大，这样得到的权重就没有密度问题，并且在同样的时间内，其分类效果要好些。当训练一个观察数据向量$\boldsymbol{v}^{(0)}=\boldsymbol{d}$时，各参数更新方式改为

$$\begin{aligned}\Delta w_{ij}\approx{}&v_i^{(0)}\times p\left(h_j=1|\boldsymbol{v}^{(0)}\right)\\&-p\left(v_i=1|\boldsymbol{h}^{(0)}\right)\times p\left(h_j=1|\boldsymbol{v}^{(1)}\right)\end{aligned}\tag{15}$$

$$\Delta a_i\approx v_i^{(0)}-p\left(v_i=1|\boldsymbol{h}^{(0)}\right)\tag{16}$$

$$\begin{aligned}\Delta b_j\approx{}&p\left(h_j=1|\boldsymbol{v}^{(0)}\right)\\&-p\left(h_j=1|\boldsymbol{v}^{(1)}\right)=h_j^{(0)}-p\left(h_j=1|\boldsymbol{v}^{(1)}\right)\end{aligned}\tag{17}$$

参数的设置，最开始时设w_{ij}, a_i, b_j都为0。为了使算法稳定，引入学习率η, η设置较大时，收敛比较快，但可能引起不稳定；η设置比较小时，可避免不稳定，但是收敛速度很慢。为克服这一情况，Rumelhart等人[20]提出了一种既加快学习速度又保持算法稳定的动量方法。

文献[3]用专家乘积系统的方法，采用权值衰减的动量方法，MNIST手写体识别DBN网络取得1.7%的错误率[3]。Sutskever等人[10]提出Nesterov动量方法。2015年Zareba等人[11]用经典动量和Nesterov动量两种方法做了比较，MNIST识别实验，得出Nesterov动量效果较好，RBM识别错误率是2.04%。Yuan等人[12]分析了动量随机梯度法的收敛性和效果性能。这些研究表明动量方法在随机梯度方法中，的确有加速网络收敛和提高学习性能的效果。

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]，在某些情况下，这样可以避免算法过早地收敛到局部最优点。这种组合方式常见的有两种，一种是采取上一次参数值修改方向与本次梯度的负方向组合的经典动量方式[10,13]，另外一种是Nesterov动量方式[11]。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，梯度计算在施加当前速度之后。因此，Nesterov动量可看作在经典动量方法中添加了一校正因子[11]。

基于以上研究，本文采取不同于经典动量[10,13]和Nesterov动量[11]的两种方式。而是采用上一次参数值修改方向与本次梯度的正方向相加的动量组合方式，加大了参数更新步长，这样更加快速收敛到最优，本文在RBM训练中采取参数更新方式是

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

而在微调阶段，结合梯度下降的特点，为了能精确找到最优点，采用的动量组合方式也不同于以上，本文微调时的参数更新方式是

$$\Delta\boldsymbol{W}_t=-m'\cdot\Delta\boldsymbol{W}_{t-1}-\eta'\Delta\boldsymbol{W}_{t-1}\tag{19}$$

得到3个参数w_{ij}, a_i, b_j的更新方式后，整个DBN训练的过程可看作是对一个深层BP网络参数的初始化，这样克服了BP网络因随机初始化权值参数而易陷入局部最优的缺点。DBN最后接的BP神经网络的输入是最顶层RBM学习到的特征输出，并用BP算法微调整个网络参数，DBN网络由信息的正向传播和误差的反向传播两个过程组成，不断调整网络的权值和偏置参数，使整个网络的误差平方和最小。

5 实验

5.1 实验1

本文采用MNIST数据库中的70000张手写数字样本图片，其中60000张图片用于训练，10000张用来测试。每幅图片大小是28×28，将一张图片拉成784维的向量，则可见层的神经元$\boldsymbol{v}$就设置为784个神经元。其RBM学习算法主要步骤见表1。

表 1 RBM学习算法的主要步骤

(1) 输入训练样本集合$S=\{\boldsymbol{v}^1,\boldsymbol{v}^2,\cdots,\boldsymbol{v}^T\}$或者$S=\{\boldsymbol{d}^1,\boldsymbol{d}^2,\cdots,\boldsymbol{d}^T\}$，每一批有$S=B=T$个训练样本，设置可见层的单元个数$n$，隐单元个数$m$，学习率$\eta$，动量项$m^*$

(2) 初始化：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) 在每个RBM中，对所有的训练样本$\boldsymbol{d}\in S$

(4) $\boldsymbol{v}^{(0)}\leftarrow\boldsymbol{v}=\boldsymbol{d}$

(5) For $t=0,1,\cdots,k-1$

(6) Gibbs 采样： For $j=1,2,\cdots,m$，采样$h_j^{(t)}\sim p\left(h_j\left|\boldsymbol{v}^{(t)}\right.\right)$

(7) For $i=1,2,\cdots,n$，采样$v_i^{(t+1)}\sim p\left(v_i\left|\boldsymbol{h}^{(t)}\right.\right)$

(8) End for

(9) For $i=1,2,\cdots,n;j=1,2,\cdots,m$

(10) 一个训练样本时，参数更新：

(11) $\Delta w_{ij}\leftarrow m^*\cdot\Delta w_{ij}+\eta\left[v_i^{(0)}p\left(h_j=1|v^{(0)}\right)-p\left(v_i=1\left|\boldsymbol{h}^{(1)}\right.\right)\cdot p\left(h_j=1\left|\boldsymbol{v}^{(1)}\right.\right)\right]$

(12) $\Delta a_i\leftarrow m^*\cdot\Delta a_i+\eta\left[v_i^{(0)}-p\left(v_i=1\left|\boldsymbol{h}^{(1)}\right.\right)\right]$

(13) $\Delta b_j\leftarrow m^*\cdot\Delta b_j+\eta\left[h_j^{(0)}-p\left(h_j=1\left|\boldsymbol{v}^{(1)}\right.\right)\right]$

(14) End for

整个实验都是在MATLAB R2014a和Microsoft Windows 7的操作系统环境下，CPU是Intel(R) core (TM) i7-4770 HQ CPU @2.2 GHz，内存是16 GB来实现的。

若构建一个隐层的RBM：784-400-10(可见单元个数-隐单元个数-类别数)，用本文算法，迭代不同次数得到结果如表2所示。以下各表中耗时指的是训练和测试一起的时间。

表 2 两种RBM模型在MNIST数据集上的实验情况(隐单元个数、时间、错误率)

不同RBM	网络结构	分类错误率(%)	耗时(min)
MapReduce RBM[8]	784-900-10	2.92	7.5
本文算法 RBM	**784-400-10**	**2.16**	**6.5**
MapReduce RBM[8]	784-900-10	2.89	12.0
本文算法 RBM	**784-400-10**	**1.70**	**10.0**

从表2的结果可看出，本文提出的算法构成1个隐层的RBM，无论在学习时间和分类效果上都比2016年的MapReduce RBM方法[8]要好，错误率降低了42%。且本文算法简单，无需面临MapReduce程序框架设计等问题。

当构建DBN网络：784-500-500-10，用本文算法，错误率可达1.32%。其结果如表3所示。

表 3 不同模型在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)
2002 POE DBN[3]	1.70	24.00
2016 AtanDBN[6]	1.39	–
SVM[21]	1.40	–
784-500-500-10 无动量	**1.63**	**1.60**
本文算法 784-500-500-10(m)	**1.32**	**1.25**

从表3的结果显示，本文提出的算法构成2个隐层的DBN的分类效果比AtanDBN[6]、专家乘积系统(POE)[3]和SVM[21]要稍好些，尤其是网络学习时间能缩短90%以上。若训练和微调时都不用动量，则无动量时的DBN识别错误率1.63%与其他方法比，分类效果没优势。而网络784-500-500-10(m)表示用本文动量算法，其错误率1.32%比没用动量时降低约20%，与未用动量时的网络时间上也能减少22%。试验结果表明，深度学习的结果要比浅层学习效果好，本文提出的改进算法在保证网络正确率有所提升的同时，也能缩短网络训练耗时。

为了图像可视化的效果，设网络隐单元为400个，可视化学习到的特征都是抽象的，如图2所示；可视化训练样本中最后一批随机的100个数字和网络识别结果如图3所示。

5.2 实验2

实验2采用CMU-PIE人脸数据库，该数据库包含68位志愿者的41368张多姿态、光照和表情的面

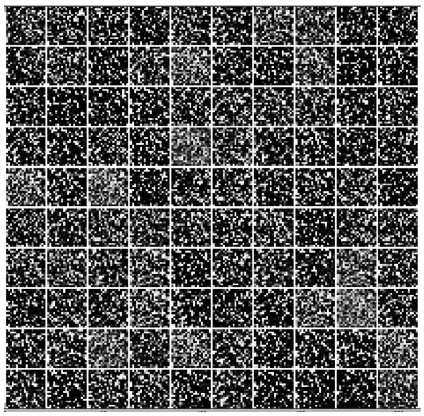
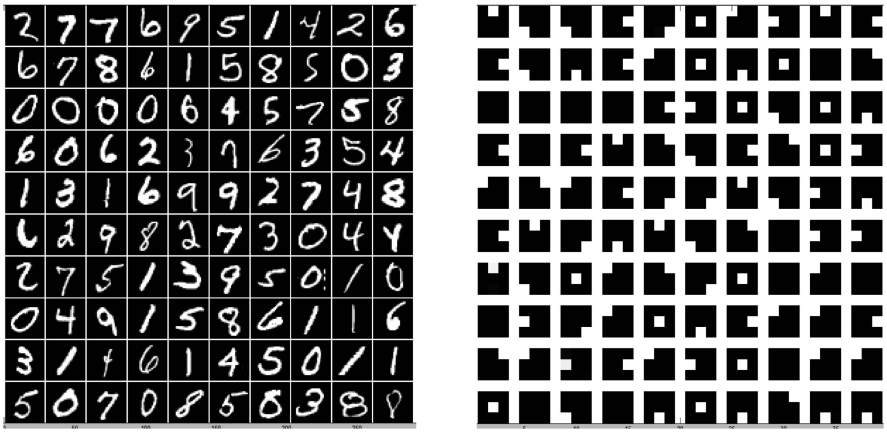


图 2 可视化第1层隐单元学习到的特征



(a) 最后一批 100 个手写体数字　　(b) 其识别结果

图 3 可视化随机的最后一批100个手写体数字和相应的识别结果

部图像。选取数据库中尺寸为32 × 32的30种人脸图像进行分类，其中每种人脸均有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个人脸的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。

人脸识别的实验有很多，如文献[22]提出基于Fisher约束的字典对学习方法在Extended YaleB图像库中进行人脸识别[22]；文献[23]提出用特征聚类的稀疏自编码网络进行人脸识别，在CMU-PIE人脸数据库中，30人的识别取得了98.17%的正确率[23]；文献[24]提出改进的深层网络，也即将卷积神经网络和稀疏自编码网络相结合的方法对CMU-PIE人脸数据库进行人脸识别，68人的图像识别中取得了96.17%的正确识别率[24]。

实验2对3种不同规模的DBN网络进行优化，比如网络规模：1024-600-600-30表示2个隐层单元都设为600个且未用动量的30人的识别原网络，1024-600-600-30(m)表示用本文动量算法的优化网络。在尺寸为32 × 32的人脸数据上，对30人和68人的人脸图像分类，原网络与动量优化网络的分类情况和耗时的对比结果如表4所示。

表4对比结果显示，本文提出的优化算法在处理人脸数据时，有较好的识别能力。对30人的识别，随着隐单元的增多，识别率也随着提升，但升到一定程度后，识别率又开始下降，隐单元设置600个最佳，30人识别率可达98.83%；且动量优化后的网络较原网络错误率降低了53%，与最新文献[23]相比，识别率是一样的，但时间上节约了50%。在没有如此大规模数据的时候，网络学习时间没有很明显的优势，但更能符合实际问题的应用。在68人的识别中，隐单元设为100个，其效果

表 4　优化网络与原网络在CMU-PIE人脸数据集上的实验结果对比

网络结构	分类正确率(错误率)(%)	耗时(min)
1024-600-600-30	97.50 (2.50)	9.50
1024-600-600-30(m)	**98.83 (1.17)**	**9.00**
文献[23]的方法，30人	98.83 (1.17)	17.70
文献[24]方法，68人	96.17 (3.83)	–
1024-100-100-68	**97.83 (2.17)**	**7.00**
1024-100-100-68(m)	**98.13 (1.87)**	**6.25**

最佳，因68人图像数据量大且类别多，则微调次数要700多才达到最优，且动量优化后的网络较未用动量的原网络识别错误率能降低14%。本文动量算法对68人脸图像的识别错误率1.87%比最新文献[24]中68人的识别错误率3.83%降低了约51%。实验结果表明，本文算法在人脸数据库中的应用，也有较好的分类效果。

可视化网络权重图像如图4所示，图4(a)为原网络学习到的权重图像，图4(b)为优化网络学习到的权重图像。图像中线条越清晰，说明网络对训练数据的学习越充分，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4(b)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，所以其优化网络的分类预测效果更佳。

综合以上实验数据分析，表明相对于传统深度学习算法，本文将专家乘积方法与改进的RBM动量方法相结合的新算法，不仅在1个隐层的RBM网络，而且在2个隐层的DBN网络中提取的特征，都具有较好的分类效果；无论是对手写体的特征提取，还是对人脸图像的光照、表情、姿态等改变都有较强的适应性和特征提取能力，系统具有较强的泛化能力和鲁棒性。本文基于RBM的专家乘积系统的改进算法，有加速网络收敛和提高图像分类效果的作用。

(a) 原网络 100 个隐单元的第 1 层网络的权重 (b) 优化网络 100 个隐单元的第 1 层网络的权重

图 4 可视化网络权重

6 结束语

本文针对深度学习的网络学习时间较长，以及RBM的算法在采取全用真实概率值的参数更新方式会导致模型识别效果不理想和带来的密度问题，提出基于RBM的专家乘积系统的一种改进算法。在专家乘积系统原理上，结合RBM改进的动量算法，采用全是真实概率值的参数更新方式，用概率值可以减少采样噪声且使模型学得更快，并有较好的识别能力。在MNIST手写体数据集和CMU-PIE人脸数据库分别做识别实验，试验结果表明，本文提出的改进算法简单有效，采取算法改进和动量优化后的网络提取的图像特征具有更好的分类效果，且能较大程度地缩短网络学习时间。在处理高维复杂大数据量的数据、多次重复实验和训练深层网络的情况下，本文的优化算法效果较为明显。但本文模型的超参数仍然需要凭经验去人工设置，如何快速有效地设置或模型自适应地调整参数是下一步研究的重点方向。

参 考 文 献

[1] LIAO S H. Expert system methodologies and applications-a decade review from 1995 to 2004[J]. *Expert Systems with Applications*, 2005, 28: 93–103. doi: 10.1016/j.eswa.2004.08.003.

[2] VUNDAVILLI PANDU R, PHANI KUMAR J, SAI PRIYATHAM CH, *et al*. Neural network-based expert system for modeling of tube spinning process[J]. *Neural Computing and Application*, 2015, 26(6): 1481–1493. doi: 10.1007/s00521-015-1820-4.

[3] MAYRAZ G and 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. doi: 10.1109/34.982899.

[4] 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. doi: 10.11897/SP.J.1016.2016.01697.

JIAO Licheng, YANG Shuyuan, LIU Fang, *et al*. Neural network in seventy: Retrospect and prospect[J]. *Chinese Journal of Computers*, 2016, 39(1): 1–21. doi: 10.11897/SP.J.1016.2016.01697.

[5] HINTON G E. Training products of experts by minimizing contrastive divergence[J]. *Neural Computation*, 2002, 14(8): 1711–1800. doi: 10.1162/089976602760128018.

[6] 罗剑江, 王振友. 一种提高受限玻尔兹曼机性能的反正切函数逼近L0范数方法[J]. 小型微型计算机系统, 2016(11): 2562–2566.

LUO Jianjiang and WANG Zhenyou. Enhancing performance of restricted Boltzmann machine using Arctan approximation of L0 norm[J]. *Journal of Chinese Computer Systems*, 2016(11): 2562–2566.

[7] 王岳青, 窦勇, 吕启, 等. 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. doi: 10.7544/issn1000-1239.2016.20150147.

WANG Yueqing, Dou Yong, Lü 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. doi: 10.7544/issn1000-1239.2016.20150147.

[8] ZHANG Chunyang, CHEN Philip, CHEN Dewang, *et al*. MapReduce based distributed learning algorithm for Restricted Boltzmann Machine[J]. *Neurocomputing*, 2016(198): 4–11. doi: 10.1016/j.neucom.2015.09.129.

[9] POLYAK T. Some methods of speeding up the convergence of iteration methods[J]. *USSR Computational Mathematics and Mathematical Physics*, 1964, 4(5): 1–17. doi: 10.1016/0041-5553(64)90137-5.

[10] SUTSKEVER I, MARTENS J, DAHL G, *et al*. On the importance of initialization and momentum in deep learning[C]. Proceedings of International Conference on Machine Learning, Atlanta, USA, 2013: 1139–1147.

[11] ZAREBA S, GONCZAREK A, TOMCZAK J M, *et al*. Accelerated learning for restricted Boltzmann machine with momentum term[C]. Proceedings of International Conference on Systems Engineering, Coventry, UK, 2015: 187–192. doi: 10.1007/978-3-319-08422-0_28.

[12] YUAN Kun, YING Bicheng, and SAYED A H. On the influence of momentum acceleration on online learning[J]. *Journal of Machine Learning Research*, 2016(17): 1–66.

[13] HINTON G E. A practical guide to training restricted Boltzmann machines[R]. Toronto: Machine Learning Group, University of Toronto, 2012: 599–619. doi: 10.1007/978-3-642-35289-8_32.

[14] FISCHER A and CHRISTIAN I. Training restricted Boltzmann machines: An introduction[J]. *Pattern Recognition*, 2014, 47: 25–39. doi: 10.1007/s13218-015-0371-2.

[15] SMOLENSKY P. Information Processing in Dynamical Systems: Foundations of Harmony Theory[M]. Cambridge, MA: MIT Press, 1986: 195–280.

[16] ROUX N L and BENGIO Y. Representational power of restricted Boltzmann machines and deep belief networks[J]. *Neural Computation*, 2008, 20(6): 1631–1649. doi: 10.1162/neco.2008.04-07-510.

[17] HINTON G E, OSINDERO S, and TEH Y W. A fast learning algorithm for deep belief nets[J]. *Neural Computation*, 2006, 18(7): 1527–1554. doi: 10.1162/neco.2006.18.7.1527.

[18] FREUND Y and HAUSSLER D. Unsupervised learning of distributions on binary vectors using two layer networks[J]. *Advances in Neural Information Processing Systems*, 1992, 4: 912–919.

[19] PETERSON C and ANDERSON J R. A mean field theory learning algorithm for neural networks[J]. *Complex Systems*, 1987, 1: 995–1019.

[20] RUMELHART D E, HINTON G E, and WILLIAMS R J. Learning representations by back-propagating errors[J]. *Nature*, 1986, 323: 533–536. doi: 10.1038/323533a0.

[21] DECOSTE D and SCHOELKOPF B. Training invariant support vector machines[J]. *Machine Learning*, 2002, 46: 161–190.

[22] 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. doi: 10.11999/JEIT160296.

GUO Jichang, ZHANG Fan, and WANG Nan. Image classification based on Fisher constraint and dictionary pair[J]. *Journal of Electronics & Information Technology*, 2017, 39(2): 270–277. doi: 10.11999/JEIT160296.

[23] 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. doi: 10.3969/j.issn.0372-2112.2018.05.003.

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. doi: 10.3969/j.issn.0372-2112.2018.05.003.

[24] 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. doi: 10.3969/j.issn.0372-2112.2017.03.017.

LI Qianyu, JIANG Jianguo, and QI Meibin. Face recognition algorithm based on improved deep networks[J]. *Acta Electronica Sinica*, 2017, 45(3): 619–625. doi: 10.3969/j.issn.0372-2112.2017.03.017.

沈卉卉：女，1980 年生，博士生，副教授，研究方向为机器学习与数据处理.

李宏伟：男，1965 年生，教授，博士生导师，主要研究方向为信息处理与智能计算.