文章编号: 1001-0920 (2015) 02-0256-05 DOI: 10.13195/j.kzyjc.2013.1390

DBN 网络的深度确定方法

潘广源, 柴 伟, 乔俊飞

(北京工业大学 电子信息与控制工程学院，北京 100124)

摘 要: 针对DBN网络隐含层层数难以选择的问题, 首先从数学生物学角度分析了随机初始化的梯度下降法导致网络训练失败的原因, 并进行验证, 证明了 RBM 重构误差与网络能量的正相关定理; 然后根据隐含层和误差的关系, 提出一种基于重构误差的网络深度判断方法, 在训练过程中自组织地训练网络, 使其能够以一种接近人类处理问题的方式解决 AI 问题. 手写数字识别的实验表明, 该方法能够有效提高运算效率, 降低运算成本.

关键词: 深度信念网络；网络深度；无监督学习；数字识别

中图分类号: TP273 **文献标志码:** A

Calculation for depth of deep belief network

PAN Guang-yuan, *CHAI Wei*, *QIAO Jun-fei*

(College of Electronic Information and Control Engineering，Beijing University of Technology，Beijing 100124, China．Correspondent：QIAO Jun-fei，E-mail：pgy_yuki@outlook.com)

Abstract: In order to calculate the depth of deep belief network(DBN) in its applications, the reason of failure in training by using random initialization in gradient-based is analyzed in both math and biology, and then verified by the test. The theorem that the reconstruction error of restricted boltzmann machine(RBM) is related to network’s energy function is proved. After that, a method to calculate the depth by using restructure error in RBM is proposed based on the relationship between hidden layers and errors. DBN approaches human-level performance in AI tasks after the self-training. The experiment of hand writing digital recognition shows that the proposed method can improve the efficiency and lower the cost.

Keywords: deep belief network；depth；unsupervised learning；digital recognition

0 引 言

人工神经网络是对大脑工作方式的模拟, 在机器学习 (ML) 和人工智能 (AI) 等领域发挥着重要作用. 生物学家发现哺乳动物的大脑具有一种层次的结构[1], 这种结构使动物在认知世界时, 不断提取从外界接收的信号, 每次提取信号中一个或多个方面的特征, 最终在最后一层将一个抽象的概念传递给大脑, 形成动物对世界的认知[2]. 人的认知过程是逐层进行、逐步抽象的, 并且深层结构的神经网络能够有效提高工作效率, 避免所谓维数灾难或降低其危害. 根据这种思想, 多伦多大学的 Hinton 教授提出了深度信念网络 (DBN)[3], 实现了人工神经网络在多隐含层建立工作中的突破.

DBN 已成功应用于多个领域[4-8], 但仍处于发展初期, 许多问题值得深入研究. 目前 DBN 缺乏有效的并行训练算法, 因此其在应用中仍使用经验法选择隐含层层数和神经元个数, 这样不仅误差较大, 而且不利于网络的扩展应用, 造成计算成本较高, 效率较低. 蒙特利尔大学的 Bengin 教授在文献 [9] 中指出: 能否确定一个合适的网络深度, 使 DBN 能够向人类处理问题那样, 解决几乎所有的 AI 问题? 该问题是开放性的, 研究该问题, 探索 DBN 在 AI 领域的应用研究, 具有重要意义.

因 DBN 涉及的范围较广, 难以给出一个标准答案, 所以可将问题进行转化, 即不以人为方式对网络深度进行规定, 而是通过设置一个机制, 让网络自身来计算最合适的深度. 因此, 根据问题和要求的不同, 得出的结果也不同. 首先, 本文通过分析 DBN 中有监督学习和无监督学习的训练过程, 得出网络深度和训练误差之间的关系, 并以此为基础, 提出一种基于

收稿日期: 2013-10-09；**修回日期:** 2014-01-12.

基金项目: 国家杰出青年科学基金项目(61225016)；国家自然科学基金重点项目(61034008)；北京市自然科学基金青年基金项目(4144067).

作者简介: 潘广源(1987−), 男, 博士生, 从事深度神经网络的结构优化、智能控制的研究；乔俊飞(1968−), 男, 教授, 博士生导师, 从事神经网络结构与算法优化、污水处理控制与优化等研究.

RBM 重构误差的深度确定方法, 让网络在计算中自组织地训练, 计算出符合要求的深度, 既能满足精度的要求, 又能有效降低成本.

1　深度信念网及特征分析

DBN 由一系列叠加的受限玻尔兹曼机 (RBM) 和一层 BP 网络构成, 其结构如图 1 所示. DBN 的训练过程可以分为两步: 首先, 使用无监督学习方法训练每一层 RBM, 且每个 RBM 的输入为上一个 RBM 的输出, 即每一层 RBM 都要单独训练, 确保特征向量映射到不同的特征空间时, 尽可能多地保留特征信息; 然后, 使用最后一层的 BP 网络接收最后一个 RBM 的输出, 用有监督的方式训练整个网络, 对其进行微调.

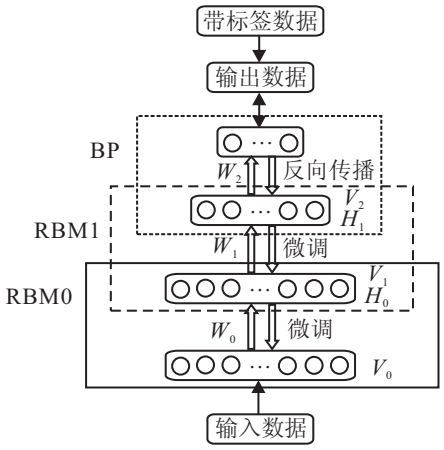


图 1　DBN 网络结构

由已知的可视层节点可得到隐含层节点的值, 即

$$p(h_j=1)=\frac{1}{1+\mathrm{e}^{-b_j-\Sigma_i v_i w_{ij}}}. \tag{1}$$

由于 RBM 是对称网络, 可以得到由隐含层节点得到可视层节点的值, 即

$$p(v_i=1)=\frac{1}{1+\mathrm{e}^{-c_i-\Sigma_j h_j w_{ji}}}. \tag{2}$$

其中: v_i 和 h_j 分别表示可视层和隐含层第 i、j 个节点值, b 和 c 为该两层偏置值, w_{ij} 为可视单元 i 和隐藏单元 j 的连接权值. 可视层的特征向量 v 和隐含层的特征向量 h 的联合概率分布

$$p(v,h)\propto\exp(-\mathrm{E}(v,h))=\mathrm{e}^{h^{\mathrm{T}}Wv+b^{\mathrm{T}}v+c^{\mathrm{T}}h}. \tag{3}$$

其中: W 为可视层和隐含层之间的权值, $\mathrm{E}(v,h)$ 为特征向量 v 和 h 的数学期望值. 网络训练的目的是求解 $\theta=(W,b,c)$, 使式 (3) 的联合概率分布 $P(v,h)$ 最大[10]. 常规方法是马氏链蒙特卡洛法 (MCMC), 实际上, 由于难以确定步长, 利用马尔科夫链的方法求得的 $P(v,h)$ 和末端的联合概率分布 $P(v_i^{\infty},h_j^{\infty})$ 难以保证收敛性, 实验中可以使用 Contrastive Divergence 准则 (CD 准则)[13]来提高计算速度和保证计算精度.

无监督学习和有监督学习能够影响到网络误差, 因此其在识别过程中发挥了重要的作用. 在 DBN 中, 反向微调的梯度下降法的初始化权值使用的是无监督训练得到的权值, 因为实验已经证明[9], 如果使用随机初始化的梯度下降法, 则容易导致训练失败, 而失败的原因一直是研究人员研究 DBN 的重要课题之一. 要讨论随机初始化的梯度下降算法失败的原因, 需要从 DBN 网络机制和生物学机制两个角度来分析, 这有助于深入了解 DBN 的工作原理, 进而展开下一步工作.

通过机器学习和生物学的前期研究[2,11], 已经知道以下事实:

1) 无监督学习在生物认知过程中发挥重要作用;

2) 带有标注信息的样本数量较少并且包含的信息较少, 因此, 无监督学习有助于增加先验知识, 使网络权值 w 处于较好的初始位置, 从而提高网络的性能;

3) 在生物系统中, 稳定的神经网络一旦形成, 则难以改变.

如果没有无监督学习, 只使用随机初始化的有监督学习, 则会导致网络训练失败. 本文提出以下两点假设:

1) 随机初始化的梯度下降法在训练时由于缺乏先验知识而陷入局部极小;

2) 难以选择合适的批处理方式和步长来帮助算法跳出局部极小.

假设 1) 为理论原因, 要证明此假设, 可以使用数学方法来分析梯度下降法在 DBN 中不同隐含层中的训练误差.

令 δ_l 为反向传播过程中第 l 层的输出, 则在自顶向下的传播中, 第1层(即最顶层)的误差为

$$\begin{aligned}&e_l=D-y_l=D-\delta_{l+1}W_{l+1}=\\&D-y_{l+1}[1-y_{l+1}]\delta_{l+2}W_{l+2}W_{l+1}=\\&D-e_L\Big(\prod_{l=1}^{L}y_i[1-y_i]W_i\Big).\end{aligned}$$

因 $y_i\in[0,1]$, 所以 $y_{i-1}=y_iW_i\in[0,1]$, 故有

$$\begin{aligned}&e_l=D-e_L\Big(\prod_{l=1}^{L}y_i[1-y_i]W_i\Big)=\\&D-e_L\Big(\prod_{l=1}^{L}y_iW_i\Big)(1-y_i)>\\&D-e_L\Big(\prod_{l=2}^{L}y_i[1-y_i]W_i\Big)=e_{l+1}.\end{aligned}$$

可以看出, 在 DBN 中, 反向微调的梯度下降法会导致训练误差逐层扩大. 若使用随机初始化的方法, 则网络权值会均匀分布于状态空间, 初试误差已经较大, 随着训练的逐层进行, 误差会逐层扩大, 最终导致

训练失败.

假设2) 是技术原因. 由于梯度下降法具有“梯度的弥散”特性[12], 即当使用反向传播方法计算导数时, 随着网络深度的增加, 反向传播的梯度幅值会快速而急剧地减小, 结果造成整体的损失函数相对于最初几层的权重的导数非常小. 这样, 当使用梯度下降法时, 最初几层的权重变化非常缓慢, 以至于它们不能从样本中进行有效地学习.

2 网络深度的确定方法

前文证明, 无监督学习是网络训练的核心方法之一, 其对具有不同深度的网络训练的误差是不同的, 而不同深度的网络成本是不一样的, 成本随着深度的增加而增加, 因此在解决实际问题时, 选择合适深度的网络, 既能满足精度的要求, 又能最大程度地节约成本. 在目前的DBN研究中, 基本是凭借经验知识对网络的深度和每个隐含层的单元数进行选择[13,15], 这样不利于充分利用网络优势, 也容易造成计算成本过高, 不利于效率的提高.

本文提出一种基于重构误差的方法, 用于计算并确定DBN的深度, 在此引入两条引理作为理论依据.

引理 1 RBM的训练精度随着深度的增加而提高.

引理 2 在DBN网络训练中, 通过无监督学习, 网络权值已处于较好的位置, 而基于梯度下降的反向运算只是在某些小的方面调节权值.

重构误差是以训练数据作为初始状态, 经过RBM的分布进行一次Gibbs转移后与原数据的差异量(一般以一阶范式或二阶范式来评估), 即

$$\mathrm{RError}=\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(p_{i,j}-d_{i,j})}{nmp_x}. \tag{4}$$

其中: n 为样本个数, m 为像素个数, p 为网络计算的值, d 为真实值, px 为取值个数或范围.

规则如下所示:

$$\begin{cases} L=N_{\mathrm{RBM}}+1,\ \mathrm{RError}>\epsilon;\\ L=N_{\mathrm{RBM}},\ \mathrm{RError}<\epsilon. \end{cases} \tag{5}$$

其中: ϵ 为目标重构误差预设值, L 为隐含层个数. 如果此时网络通过训练, 重构误差达到目标, 即低于预设值, 则开始进行梯度反向微调; 否则, 令网络深度自动加一, 继续进行训练. 由于测试样本的取值范围和真实值是可以提前预知的, 在应用中可通过计算得到重构误差的值(一般取值令正确率为95%以上).

RBM的训练是基于模拟退火原理, 因此DBN具有模拟退火的某些特性. 寻求重构误差和网络能量之间的耦合关系, 能够从理论上证明该方法的正确性. 可使用RBM特征向量的期望值来描述网络能量, 即

$$E(v,h)=h^{\mathrm{T}}Wv+b^{\mathrm{T}}v+c^{\mathrm{T}}h. \tag{6}$$

定理 1 重构误差和网络能量正相关.

证明 令 P 为计算值, D 为真实值, 有

$$P=P(v),D=P(v_0). \tag{7}$$

由条件概率公式可得

$$P=P(v)=P(v_0)P(h|v_0)P(v|h), \tag{8}$$

由全概率公式 $P(v|h)=P(v,h)/P(h)$ 可知

$$P=P(v_0)\frac{P(v_0,h)}{P(v_0)}\frac{P(v,h)}{P(h)}, \tag{9}$$

消除 $P(v_0)$ 得

$$P=P(v_0,h)\frac{P(v,h)}{P(h)}. \tag{10}$$

再一次使用条件概率公式, 可得

$$P=P(v_0|h)P(h)\frac{P(v,h)}{P(h)}=P(v_0|h)P(v,h). \tag{11}$$

因此, 将重构误差公式代入, 可得

$$\mathrm{RError}=\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(p_{i,j}-d_{i,j})}{nmp_x}=P-D. \tag{12}$$

将式(7)和(12)代入, 得

$$RE=P(v_0|h)P(v,h)-P(v_0), \tag{13}$$

于是有

$$RE=P(v_0)(P(v,h)-1). \tag{14}$$

根据目标函数和式(7), 可得

$$RE\propto D(\mathrm{E}(v,h)-1)\propto \mathrm{E}(v,h). \tag{15}$$

重构误差正相关于网络能量, 定理得证. □

由定理1可知, 重构误差判断法既能保证运算简单, 便于实现, 又因为与网络能量具有耦合特性而从网络机理的角度提出了合理的判断机制, 使运算结果更具说服力. 网络训练流程如图2所示.

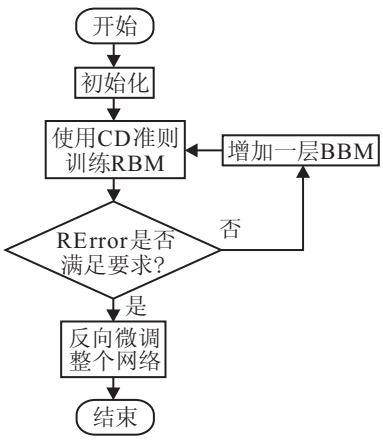


图 2 使用RBM重构误差计算DBN深度

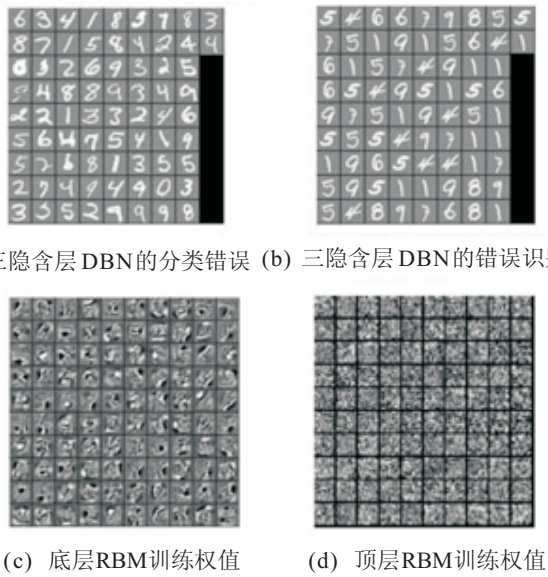
3 实验与分析

为了测试本文提出的方法是否有效, 设计了一个数字识别实验. 该实验使用MNIST手写数据库, 该

数据库拥有60 000个训练图像和10 000个测试图像, 库中数字均为手写体, 每个数字都用相当多数量的手写方式来显示, 已经有许多不同模式识别技术应用到这个数据库中, 因此这个数据库是一个公认的、理想的评估新方法的方式. 对于MNIST学习任务的基本版本, 由于不提供几何知识, 也没有特殊的预处理或训练组增强, 像素点的随机但固定的排列并不会影响学习算法. 取5 000个样本用于无监督学习, 从中取出1 000个样本用于有监督学习, 再取1 000个样本进行测试. 数据库所含样本为0～9的阿拉伯数字, 均为手写体, 每个图像为28×28的像素, 5 000个样本分为50批次, 每批100个样本, 因此每层的神经元默认100个, 重构误差条件设定正确率为99%以上, 通过计算得出RError = 1.59e-5.

网络在隐含层层数达到3 (即深度4) 时停止增加, 此时通过测试1 000个样本, 产生了74处错误, 原图像和产生的错误分别如图3(a)和图3(b)所示. 通过分析图中数字, 可进一步统计出网络容易在判断什么样的图像、提取什么样的特征时产生的错误, 这有助于对进一步提高网络性能提供参考.

图3(c)为网络训练产生的最底层(即第1层)RBM的权值, 图3(d)为最顶层(即最后层)RBM的权值. 这两张图显示的是DBN训练到的权值具象化, 可以看出, 随着深度的增加, 网络权值愈加抽象, 表明网络识别的信息是对这些抽象数据的组合(实际上这种组合具有稀疏特性[17]).

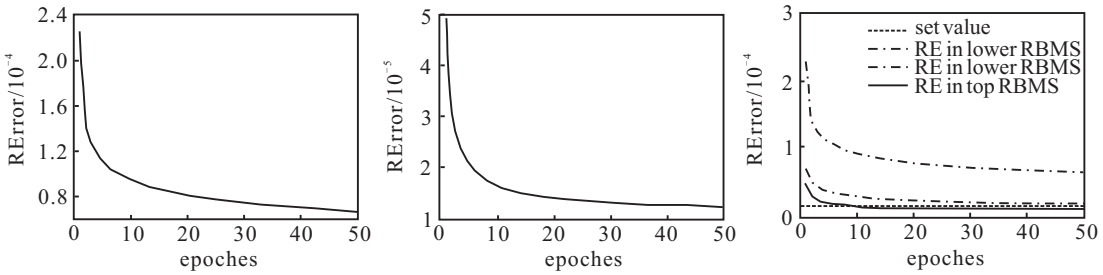


(a) 三隐含层DBN的分类错误　(b) 三隐含层DBN的错误识别

(c) 底层RBM训练权值　(d) 顶层RBM训练权值

图 3　训练结果

图4(a)为第1层RBM得到的重构误差, 图4(b)为最后1层RBM得到的重构误差, 从中可以看出重构误差在每一层RBM中呈下降趋势. 将3个RBM的重构误差放到同一张图里, 如图4(c)所示.



(a) 底层RBM的重构误差　(b) 顶层RBM的重构误差　(c) 各个RBM的重构误差

图 4　重构误差曲线

图4(c)中实线为最后1层RBM重构误差, 此时已经达到预设的目标. 关于为什么每次RBM重构误差的初始值比上一个RBM的终值高, 这是因为每增加一层RBM, 其初始权值是随机给定的, 故开始的重构误差处于较高的状态. 而如何更有效地利用先验知识对其后一层RBM初始化, 也值得进一步研究, 这将有助于提高网络性能.

为了便于比较和数据分析, 进一步将网络深度增加至6, 并计算了DBN训练数据, 如表1所示.

由表1可以看出, 随着深度的增加, 网络重构误差逐渐降低, 运算时间逐渐增大, 这符合网络的特性. 发生的错误个数(即正确率)在深度为4时达到最高, 为92.6%. 而如果将网络深度进一步增加, 此时, 虽然重构误差仍然继续降低, 但正确率却出现了下降, 最终在深度为6时降低为88.8%.

表 1　不同深度的DBN训练数据

深度	重构误差	错误个数	正确率/%	运算时间/s
2	6.618 1e-5	88	91.2	22.9
3	2.074 2e-5	77	92.3	26.9
4	1.232 6e-5	74	92.6	32.9
5	0.778 5e-5	89	91.1	38.6
6	0.534 4e-5	112	88.8	44.6

为什么随着深度的增加, 正确率会下降呢? 针对这种现象, 可能造成的原因如下:

1) 只有最后1层RBM的重构误差能够达到要求, 而前 $L-1$ 层RBM的重构误差会通过误差累加使正确率降低;

2) 隐含层 L 增加导致反向梯度下降算法的误差累加过大;

3) 隐含层 L 增加导致运算时间复杂度增加, 效率降低.

因此, 研究如何合理设置重构误差, 或者寻求网络计算成本与深度之间的关系, 将有助于网络在自学习过程中更具智能化.

4 结 论

本文介绍了深度信念网 (DBN) 的结构和算法, 针对 DBN 网络隐含层层数难以选择问题, 分析了无监督学习初始化和随机初始化梯度下降算法的性能, 得出了网络隐含层与训练误差之间的关系, 进一步提出让网络根据任务的不同, 自组织自训练并确定网络深度. 主要工作如下: 1) 基于随机初始化的梯度下降算法容易导致网络训练失败, 因此提出两点假设, 利用反向微调的公式, 从数学角度进行了证明, 并从生物学角度进行论述; 2) 提出一种基于重构误差的判断方法, 用于确定网络深度, 在训练过程中自组织地训练网络, 以解决深度学习网络在模式识别问题中的隐含层层数选择问题, 有效提高运算效率, 降低运算成本.

此外, 由于本文仅将网络深度作为唯一自组织变量, 将每个 RBM 的隐层单元数固定, 忽略了不同隐含层之间的关系. 因此, 下一步的工作可以引入生物学和数学相关领域的知识, 寻求智能化设置阈值, 并将其扩展到其他领域例 (如过程控制等) 的应用研究方法.

参考文献(References)

[1] Lee T S, Mumford D. Hierarchical Bayesian inference in the visual cortex[J]. Optical Society of America, 2003, 20(7): 1434-1448.

[2] Rossi A F, Desimone R, Ungerleider L G. Contextual modulation in primary visual cortex of macaques[J]. J of Neuroscience, 2001, 21(5): 1689-1709.

[3] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786): 504-507.

[4] Dahl G E, Yu D, Deng L, et al. Large vocabulary continuous speech recognition with context-dependent DBN-HMMS[C]. Proc of IEEE Int Conf on Acoustics, Speech and Signal Processing. Prague, 2011: 4688-4691.

[5] Deselaers T, Hasan S, Bender O, et al. A deep learning approach to machine transliteration[C]. Proc of the 4th Workshop on Statistical Machine Translation. Athens, 2009: 233-241.

[6] Fasel I, Berry J. Deep belief networks for real-time extraction of tongue contours from ultrasound during speech[C]. Proc of the 20th Int Conf on Pattern Recognition. Stroudsburg: Association for Computational Linguistics, 2010: 1493-1496.

[7] Deng L, Seltzer M L, Yu D, et al. Binary coding of speech spectrograms using a deep auto-encoder[C]. Proc of the 11th Annual Conf on Int Speech Communication Association. Makuhair, 2010: 1692-1695.

[8] 陈宇, 郑德权, 赵铁军. 基于Deep Belief Nets 的中文名实体关系抽取[J]. 软件学报, 2012, 23(10): 2572-2585. (Chen Y, Zheng D Q, Zhao T J. Chinese relation extraction based on deep belief nets[J]. J of Software, 2012, 23(10): 2572-2585.)

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

[10] Yoshua Bengio, Pascal Lamblin, Dan Popovici, et al. Greedy layer-wise training of deep networks[C]. Advances in Neural Information Processing Systems 19 (NIPS 2006). Vancouver, 2007: 153-160.

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

[12] Yang Hu, Yong Yu. Learning Restricted Boltzmann Machines using Mode-Hopping MCMC[C]. The 4th Int Conf on Machine Learning and Computing. Xi'an, 2012, 20: 105-110.

[13] Le Roux, Nicolas, Yoshua Bengio. Representational power of restricted boltzmann machines and deep belief networks[J]. Neural Computation, 2008, 20(6): 1631-1649.

[14] Thomas P, Karnowski. Deep spatiotemporal feature learning with application to image classification[C]. The 9th Int Conf on Machine Learning and Applications. Washington, 2010: 883-889.

[15] Li Deng, Dong Yu, John Platt. Scalable stacking and learning for building deep architectures[C]. ICASSP. Kyoto, 2012: 2133-2137.

[16] Bengio Yoshua, Olivier Delalleau. On the expressive power of deep architectures[C]. Algorithmic Learning Theory. Berlin: Springer /Heidelberg, 2011: 18-36.

[17] MarcAurelio Ranzato, Christopher Poultney, Sumit Chopra, et al. Efficient learning of sparse representations with an energy-based model[C]. Advances in Neural Information Processing Systems(NIPS 2006). Vancouver, 2007: 1137-1144.

(责任编辑: 孙艺红)