知识堆叠降噪自编码器

刘国梁 [1] 余建波 [1]

摘 要 深度神经网络是具有复杂结构和多个非线性处理单元的模型, 广泛应用于计算机视觉、自然语言处理等领域. 但是, 深度神经网络存在不可解释这一致命缺陷, 即 "黑箱问题", 这使得深度学习在各个领域的应用仍然存在巨大的障碍. 本文提出了一种新的深度神经网络模型 —— 知识堆叠降噪自编码器 (Knowledge-based stacked denoising autoencoder, KBSDAE). 尝试以一种逻辑语言的方式有效解释网络结构及内在运作机理, 同时确保逻辑规则可以进行深度推导. 进一步通过插入提取的规则到深度网络, 使 KBSDAE 不仅能自适应地构建深度网络模型并具有可解释和可视化特性, 而且有效地提高了模式识别性能. 大量的实验结果表明, 提取的规则不仅能够有效地表示深度网络, 还能够初始化网络结构以提高 KBSDAE 的特征学习性能、模型可解释性与可视化, 可应用性更强.

关键词 深度学习, 堆叠降噪自编码器, 知识发现, 符号规则, 分类规则

引用格式 刘国梁, 余建波. 知识堆叠降噪自编码器. 自动化学报, 2022, **48**(3): 774−786

DOI 10.16383/j.aas.c190375

Knowledge-based Stacked Denoising Autoencoder

LIU Guo-Liang[1] YU Jian-Bo[1]

Abstract Deep neural network is a complex structure and multiple nonlinear processing units models. It has achieved great success in computer vision, natural language processing and speech recognition. However, deep neural networks have unexplained fatal flaws, namely the "black box" problem, which makes the application of deep neural networks (DNNs) in various fields have huge obstacles. This paper proposes a new deep network system, knowledge-based stacked denoising autoencoder (KBSDAE). Try to extract and insert knowledge from the stacked denoising autoencoder (SDAE) with a simple logic language, and ensure that logic rules can perform deep reasoning. The system not only accurately represents the SDAE, but also adaptively builds the network model and improves network performance. Experiments show that this modular learning system can effectively explain SDAE and improve network performance.

Key words Deep learning, stacked denoising autoencoder (SDAE), confidence rules, knowledge discovery, pattern recognition

Citation Liu Guo-Liang, Yu Jian-Bo. Knowledge-based stacked denoising autoencoder. *Acta Automatica Sinica*, 2022, **48**(3): 774−786

知识的表述和推理一直是人工智能的热点话题, 其中知识所代表的是数据特征与标签间存在的一般规律. 在人工智能发展早期, 符号规则用来表述知识并进行推理, 研究者企图通过这种方式来以人类的思维模式解释机器的结论, 而在这一过程中的规则定义为知识[1]. 符号形式的优点在于可以通过推导对知识进行验证, 并且其规则推导过程都是可以理解的.

在大数据时代, 以连接主义为核心的神经网络相较于符号系统具有更好的适应性. 其中, 深度学习[2] 凭借其良好的特征学习性能近些年已经在各个领域得到了广泛应用. 通过深度学习得到的深度网络即为具有深度结构的神经网络. 在深度神经网络领域, Hinton 等[3−4] 基于深度置信网络 (Deep belief network, DBN) 提出非监督贪心逐层训练算法, 为解决深层结构优化难题提供了解决办法, 并进一步提出了堆叠自动编码器 (Stacked auto encoder, SAE). Lecun 等[5] 提出了卷积神经网络 (Convolutional neural network, CNN), 利用空间相对关系以减少参数数目来提高反向传播算法的训练效果, 在图像识别方面应用前景广阔. 此外, 深度学习还出现了一些变形结构, 如堆叠降噪自编码器 (Stacked denoising auto encoder, SDAE)[6]. 深度网络凭借其强大的学习能力广泛地应用于各个领域, 但同时也

收稿日期 2019-05-16 录用日期 2019-08-22
Manuscript received May 16, 2019; accepted August 22, 2019
国家自然科学基金 (71771173) 资助
Supported by National Natural Science Fundation of China (71771173)
本文责任编委 张民
Recommended by Associate Editor ZHANG Min
1. 同济大学机械与能源工程学院 上海 201804
1. School of Mechanical and Energy Engineering, Tongji University, Shanghai 201804

据有不可忽视的 “黑箱问题”[7], 即人类不能通过了解网络内部的结构和数值特性来得到数据特征和数据标签之间的关系, 这一问题从根本上限制了深度神经网络的发展.

近年来, 一些研究者开始探究如何将符号系统和神经网络相结合, 其中一部分人希望通过符号规则所表示的逻辑关系来解释网络内部的结构和数值特性, 另一部分希望将人类已知的知识通过符号系统传入神经网络以提高网络性能. Gallant[8] 最先提出了一种使用 IF-THEN 规则解释推理结论的神经网络专家系统. 其后 Towell 等[7] 提出基于知识的人工神经网络 (Knowledge-based artificial neural network, KBANN), 利用 MofN 规则实现对神经网络的知识抽取和插入, 通过这种方式解释网络并增强网络性能. Garcez 等[9] 在 KBANN 的基础上提出了 CILP (Connectionist inductive learning and logic programming) 系统, 该系统将逻辑规则应用到初始化网络过程中, 使网络可以更好地学习数据和知识. Fernando 等[10] 在 KBANN 的基础上提出了 INSS (Incremental neuro-symbolic system) 系统, 成功利用包含实数的分类规则初始化人工神经网络. Setiono[11] 在前人的基础上从标准的三层前馈神经网络中抽取了 IF-THEN 规则, 该抽取算法最大的亮点在于将隐藏节点的激活值离散化. 袁静等[12] 尝试利用符号逻辑语言描述神经网络, 并通过激活强度从理论上帮助规则进行推导. 钱大群等[13] 根据神经网络节点的输出值建立约束并生成规则. 这种规则被用来解释神经网络的行为. 黎明等[14] 将模糊规则与神经网络相结合以提高模型的模式识别能力. 在深度学习上更进一步的研究中, Penning 等[15] 提出神经符号认知代理模型 (Neural-symbolic cognitive agent, NSCA), 试图将时间符号知识规则与 RTRBM (Recurrent temporal restricted Boltzmann machine) 结合实现在线学习. Odense 等[16] 将受限玻尔兹曼机 (Restricted Boltzmann machine, RBM)与 MofN 规则相结合. 深度置信网络 (DBN) 是由 RBM 堆叠形成的, 而这一研究的意义则在于这是对 DBN 网络进行模块化解释的主要基础, 也是对神经网络与符号结合的一种新思考. Tran 等[17] 在 NSCA 的基础上将置信度规则与 DBN 相结合, 实现知识的抽取和插入. Li 等[18] 通过将符号系统与神经网络相结合形成神经符号系统, 将符号主义与连接主义的优点集成, 形成新的推理学习模型. Garcez 等[19] 提出了神经符号计算的概念, 其中知识以符号的形式表示, 而学习和推理由神经元计算. 通过这种方式将神经网络的鲁棒学习和有效推理与符号的可解释性相结合. 论文从知识的表示、提取、推理和学习方面进行讨论, 并分别对基于规则、基于公式和基于嵌入的神经符号计算方法进行了论述. 但是, 上文所提到的知识提取与插入方法通常在浅层神经网络模型实施, 对于深度神经网络模型的知识提取与插入有待深入展开.

本文提出了一种新的深度神经网络模型 —— 知识堆叠降噪自编码器 (Knowledge-based stacked denoising autoencoder, KBSDAE), 实现了符号系统与深度 SDAE之间的有效集成, 解决了深度神经网络的知识发现, 特征提取与网络可视化问题. 本文的主要贡献如下: 1) 提出了一种新的深度知识神经网络 KBSDAE 模型, 显著地提高了特征学习及模式识别性能; 2) 提出了一种从深度网络发现知识的方法, 实现了深度网络可解释的目的; 3) 有效地将符号规则与分类规则相结合, 获得了一种具有高推导性能的规则系统. 最后采用各类标杆数据验证了本文所提出方法的有效性与可应用性.

1 堆叠降噪自编码器

自编码器 (Autoencoders, AE) 是基于神经网络的特征表达网络, 由输入层 (x)、隐藏层 (h) 和输出层 (y) 构成, 是深度学习的典型模型之一[6]. 它通过编码和解码运算重构输入数据, 使得重构误差最小. 通过这种方式得到输入数据的隐藏层表达, 以达到特征提取的目的. 因为学习过程中不存在数据标签, 但是又以输入数据作为重构目标, 所以认定该模型为自监督学习过程.

自编码器的编码阶段是输入层 x 到隐藏层 h 的过程, 具体表示为

$$h = f_\theta(x) = \sigma(wx + b) \tag{1}$$

其中, σ 是 Sigmoid 非线性激活函数: $\sigma(x) = 1/(1+\mathrm{e}^{-x})$, 参数集合 $\theta = \{w, b\}$. 解码阶段是隐藏层 h 重构输出层 y 的过程, 具体表示为

$$y = g_{\theta'}(h) = \sigma'(w'h + b') \tag{2}$$

其中, σ' 是 Sigmoid 非线性激活函数, 参数集合 $\theta' = \{w', b'\}$.

通过最小化重构误差函数 $L(x, y) = \|x - y\|^2$ 来逐步地调整网络内部的参数 θ, θ', 优化方式选择随机梯度下降法. 最优参数表示为

$$\theta, \theta' = \arg\min_{\theta, \theta'} L(x, g_{\theta'}(f_\theta(x))) \tag{3}$$

降噪自编码器 (Denoising auto-encoder, DAE) 是基于 AE 的一种变形, 通过噪声污染训练输入数据以增加网络的鲁棒性, 防止过拟合[7]. 从图 1 可以

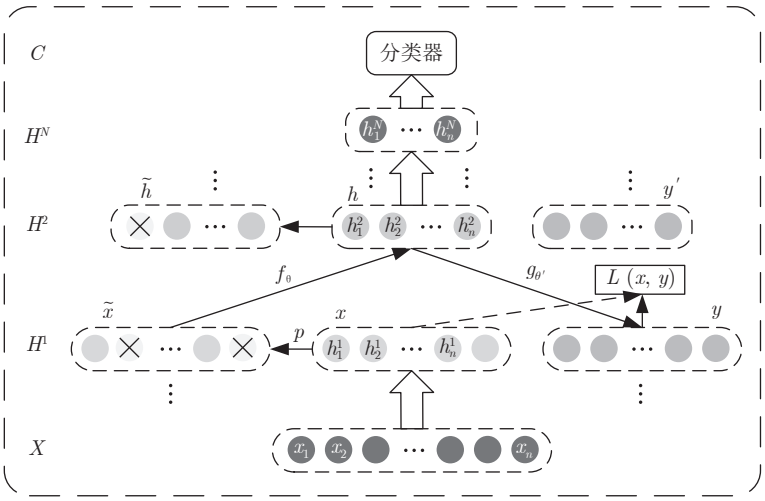


图 1 堆叠降噪自编码器工作原理示意图

Fig. 1 Stacked denoising auroencoder working principle diagram

看到 DAE 的训练过程, 首先利用随机函数以一定的概率 p 将原训练数据 x 中的一些单元置零得到被污染的数据 $\tilde{x}$; 通过编码和解码对 $\tilde{x}$ 进行重构; 最后调整网络参数 θ, θ'. 在利用受污染的数据学习后, 网络可以具有更好的鲁棒性. 可以说, DAE 相较于传统的 AE 具有更强的泛化能力.

将若干个 DAE 堆叠起来, 就可以形成 SDAE, 如图 1 所示. 每一个 DAE 都以前一个 DAE 的隐藏层输出作为原始输入数据, 添加噪声后利用被污染的数据进行训练. 其训练过程首先是逐个 DAE 贪婪训练, 最后通过 BP 算法[20] 微调整个网络以获得最佳网络模型.

本文提出的 KBSDAE 实现了深度网络知识抽取和插入目的, 形成的 KBSDAE 系统如图 2 所示. 下面将阐述知识抽取和插入的详细过程.

2 规则抽取与推理

本文提出的规则集是由置信度符号规则和分类规则合并而成的, 两种形式规则的合成有助于规则集具有更高的可理解性和推理精度. 对两种规则集分别建立了相应的规则抽取算法, 并且面向规则推理过程中两者形式不同的问题, 建立了一套基于惩罚逻辑[21] 的完整推理算法.

2.1 混合规则

符号规则方面, 传统逻辑符号规则有很多种表达形式, 但是它们在复杂问题上的逻辑推理能力较弱. 为了能描述深度神经网络, 本文选择了一种数值和符号相结合的规则 —— 置信度规则. 这种规则存在以下特性: 规则节点与网络神经元一一对应; 规则节点间的逻辑关系是从网络中拓扑而出; 规则置信值是对网络权值进行拟合得出的; 即便面对复杂的大型规则结构也可以进行有效的数学推理. 这些特性赋予符号规则两种能力: 1) 符号规则的结构与网络基本相同且元素一一对应, 网络内部的逻辑关系可以被迁移到规则上作为一种网络内部关系的表现; 2) 规则可以作为深度神经网络的一种简化表示, 具备一定的网络能力. 所以符号规则的运行其实是对神经网络行为的一种简化模仿, 而这种模仿过程是人类所能理解的.

置信度规则是一个符合充要条件的等式 $c:\ h \leftrightarrow x_1 \wedge \cdots \wedge x_n$, 其中 c 属于实数类型, 定义为置信值; h 和 $x_i\,(i \in [1, n])$ 为假设命题, 这种符号规则形式与文献 [16] 中的规则相似, 但由于面向的网络不同, 规则符号的意义也不同. 本文定义具体的置信度符号规则为

$$\lambda^l = \begin{cases} c_j^l : h_j^l \leftrightarrow (\wedge_{\forall p \in P} h_p^{l-1}) \wedge (\wedge_{\forall n \in N} \neg h_n^{l-1}), & \text{若 } 1 < l < N \\ c_j^l : h_j^l \leftrightarrow (\wedge_{\forall p \in P} x_p) \wedge (\wedge_{\forall n \in N} \neg x_n), & \text{若 } l = 1 \end{cases} \tag{4}$$

该规则可解释为: 当 $x_1, \cdots, x_n$ 命题成立时, h 命题也成立的置信值为 c, 反之亦成立. 其中, λ_j^l 是符号规则标签, 解释为第 l 层第 j 个符号规则; h_j^l 代表 DAE 中第 l 个隐藏层中第 j 个神经元; $x_i\,(i \in [1, n])$ 代表 DAE 输入层中第 i 个神经元, P 和 N 分别代表对 h_j^l 产生积极和消极影响的输入层神经元集合. 根据表达式可以看出, 符号规则的整体结构和堆叠的自编码器具有相似的堆叠嵌套结构, 可以最大化复现网络结构.

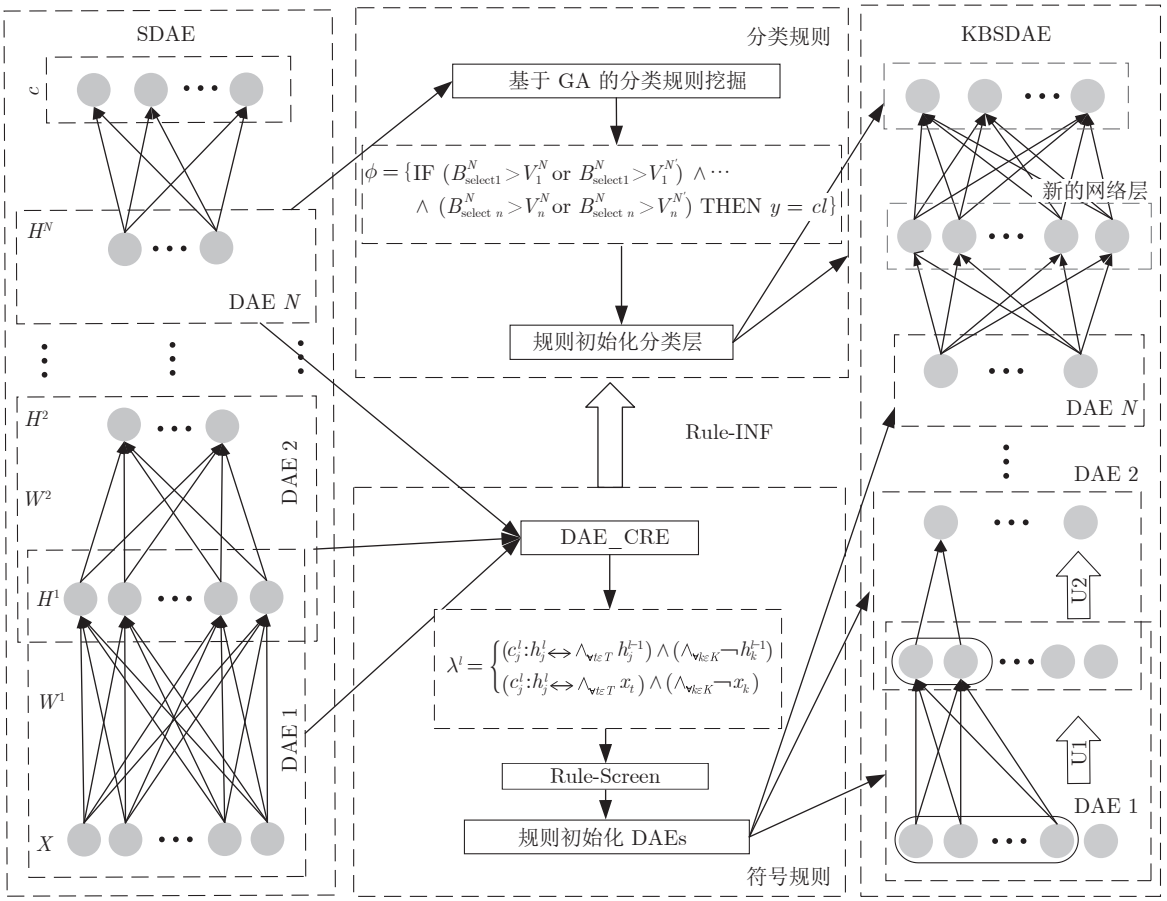


图 2　KBSDAE 模型结构图

Fig. 2　KBSDAE model structure diagram

分类规则的解释逻辑是对网络分类过程的一种模仿, 由数值和逻辑符号组成. 这种规则是从功能上对神经网络分类层的模仿, 通过数值和符号定义不同类别的不同区间可以最大限度的对网络分类过程进行模仿和解释. 同时, 由于分类规则组成元素与置信度符号规则相同, 故两者可以进行有效结合. 具体的规则形式表示为

$$\phi=\Big\{\text{IF}\,(B^N_{\text{select1}}>V^N_1 \text{ or } B^N_{\text{select1}}>V^{N'}_1)\wedge\cdots\wedge (B^N_{\text{selectn}}>V^N_n \text{ or } B^N_{\text{selectn}}>V^{N'}_n)\ \text{THEN}\, y=cl\Big\} \tag{5}$$

其中, ϕ 是规则的标签, B^N 表示从置信度符号规则集中推导输出的可信值集, V^N 代表对应的实数集, y 为类值标签, 可以被赋值为某一类 cl. 该规则的解释为: 当符号规则集输出的 B^N 符合相应的条件时, 可以判定这组数据属于某一类 cl.

两种规则的合并即为本文提出的混合逻辑规则 R_{mix}

$$R_{\text{mix}}=\{\wedge^N_{l=1}\lambda^l\wedge\phi\} \tag{6}$$

可以看出, 它是由堆叠的多层置信度符号规则和分类规则混合而成.

2.2　符号规则抽取

符号规则抽取目标是每一个自编码器的编码阶段[17], 抽取过程的核心原理是将规则置信度 $c_j s_j$ 最大化拟合权重值 w_j. 根据 DAE 基本原理, 其输入数据 x 到隐含表示 h 的映射表示为

$$h_j=\sigma\left(w^{\mathrm{T}}_j x+b_j\right) \tag{7}$$

根据式 (7) 提出新的函数, 同样可以将数据 x 映射到隐藏层空间中

$$h_{j1}=\sigma\left(c_j s^{\mathrm{T}}_j x+b_j\right) \tag{8}$$

其中, c_j 是连续实数, $s_j\in\{1, 0, -1\}$ 是对 w_j 的符号项表示, 可以理解为 $s_{ij}=\text{sign}(w_{ij})$. 当 x_i 或 $\neg x_i$ 在规则 j 中时, s_{ij} 分别等于 1 或 −1, 其余情况下 s_{ij} 等于 0. 对比式 (7) 和式 (8) 可以看出, 二式的形式和元素基本相同. 为了使 h_{j1} 可以有效地代表隐藏层空间, 需要找到合适的 c_j 和 s_j, 使 h_{j1} 近似于 h_j. 本文通过最小化 w_j 与 $c_j\times s_j$ 之间的欧氏距离实现

拟合过程

$$d(w, cs) = \sum_{ij} \|w_{ij} - c_j s_{ij}\|^2 \tag{9}$$

在提取 c_j 的过程中, 对式 (8) 进行求导并令其等于零, 即

$$\sum_i 2(w_{ij} - c_j s_{ij}) s_{ij} = 0, \quad \forall j \tag{10}$$

经过数学推导, 最终可以得到 c_j 的表达式为

$$c_j = \frac{\sum\limits_i w_{ij} s_{ij}}{\sum\limits_i s_{ij}^2} \tag{11}$$

进一步对式 (9) 分析, 可得

$$\|w_{ij} - c_j s_{ij}\|^2 = \begin{cases} (|w_{ij}| + c_j)^2, & 若\ s_{ij} = -1 \\ (|w_{ij}| - c_j)^2, & 若\ s_{ij} = 1 \\ |w_{ij}|^2, & 若\ s_{ij} = 0 \end{cases} \tag{12}$$

由于 c_j 是正实数, 分析式 (12) 可得

$$(|w_{ij}| + c_j)^2 > (|w_{ij}| - c_j)^2 \cap (|w_{ij}| + c_j)^2 > |w_{ij}|^2$$

可知, 如果想要欧氏距离最小, 需要在 $s_{ij} = 1$ 或 $s_{ij} = 0$ 的情况下. 进一步, 如果一个元素对应的欧氏距离只有在 $s_{ij} = 0$ 的情况下最小, 那么可以判定该元素 x_i 不应该出现在规则 h_j 中. 即

$$\begin{aligned} |w_{ij}|^2 &\le (|w_{ij}| - c_j)^2 \\ \frac{c_j}{2} &\ge |w_{ij}| \end{aligned} \tag{13}$$

可知, 当 $2|w_{ij}| \le c_j$ 时, 对应的元素不应出现在规则中. 通过式 (11) 和式 (13) 可以得到一种具有强连接关系和判定系数的符号规则集, 能够更加紧凑地描述 DAE 网络.

根据上述分析, 从 DAE 网络中抽取置信度符号规则的具体算法 DAE_CRE 可参考文献 [17] 的规则抽取算法. SDAE 是由 DAE 编码网络堆叠而成的, 只需将 DAE_CRE算法迭代运行, 就能抽取出 SDAE 编码器部分的置信度符号规则集. 同时, 通过上述方法得到的规则集具有堆叠嵌套结构, 这有助于规则集的推理性.

2.3 分类规则抽取

置信度符号规则可以表示 SDAE 的堆叠编码器部分, 但是在分类层的表示上存在较大的信息损失[17]. 为了更加准确地解释 SDAE, 本文运用数据挖掘领域分类规则的相关知识, 从统计意义上解释分类层. 为了与之前的符号规则保持一致性, 分类规则的基本形式也是 IF-THEN.

利用遗传算法 (Genetic algorithm, GA)[22] 挖掘符号规则推导出的置信值与分类标签之间的关系是分类规则抽取算法的核心思想. 在实验中发现, 从符号规则中推导出的置信度集合存在数据间差异过小的问题. 这使得一般的 GA 编码方式得到的规则性能较低. 针对上述问题, 本文改进了 Yu 等[23] 的编码方式, 使分类规则性能提高的同时让规则更加紧凑简单. 算法基因的构成如表 1 所示, 每一个基因由状态判别量 (Active, Act)、符号判别量 (Distinguished symbol, DS)、具体数值 (Value, V) 三个元素组成. N 个这样的基因按照特征的顺序组成了染色体, N 为特征数量.

表 1 遗传算法基因编码示意表

Table 1 Genetic algorithm gene coding schematic

Gene 1			Gene 2			···	Gene N		
Act_1	DS_1	V_1	Act_2	DS_2	V_2	···	Act_N	DS_N	V_N

本文选用的 GA 算法适应度函数是较为普遍的规则质量评估函数:

$$F = \frac{TP}{TP + FN} \times \frac{TN}{FP + TN} \tag{14}$$

其中, TP 表示被判断为正样本, 实际上也是正样本; FN 表示被判断为负样本, 实际上也是负样本; TN 表示被判断为负样本, 但实际上是正样本; FP 表示被判断为正样本, 但实际上是负样本. $(TP/(TP + FN))$ 和 $(TN/(FP + TN))$ 分别表示每条规则对该类别的敏感度和特异性. 通过这个适应度函数, 可以得到对每一个分类都具有高敏感度和特异性的规则.

2.4 规则推理

由于符号规则与分类规则的表现形式不同, 需要特定的推导方法才能使混合规则集具有可推导性, 抽取的规则集的意义才能显现. 本文提出的推导方法 (Rule-INF) 分成两个部分: 符号规则部分和分类规则部分. 符号规则的推导借鉴了文献 [21] 中惩罚逻辑的思想并加以改善. 通过置信度值的数值特性, 符号规则的推导突破了二值的限制, 可以用于推导大量的连续性数据. 分类规则的推导将符号规则输出的置信值与规则一一对照, 寻找出符合条件的规则并进行分类. 具体的推导算法如算法 1 所示.

算法 1. Rule-INF 算法

输入. 规则集 R_{mix}, 数据集 X

1) 对 X 进行归一化 $X \leftarrow \mathrm{Norm}(X)$

2) **For** $l = 1$ to N **do** // N 是隐层的数量

3) 初始化置信向量 $B^l = \{\}$;

4) **For** 每个符号规则 λ_j^l **do**

5) 将 X_t, X_k 分别赋值给 α_t, $\alpha_k(t \in T, k \in K)$;
6) $\alpha = c_j^l \times (\sum_t \alpha_t - \sum_k \alpha_k)$;
7) 置信值 $B_j^l = \text{Norm}(\alpha)$;
8) **end For**
9) 将 B^l 赋值给 X;
10) **end For**
11) **For** 每个分类规则 ϕ_j **do**
12) **IF** $B^N \leftrightarrow \phi_j$ **THEN** $y = c$; // 比较置信值和规则以确定类别
13) **end For**

输出. 符号规则集 c_j 和 r_j

3 KBSDAE

本节将讨论如何利用混合规则集 R_{mix} 初始化并训练 KBSDAE, 这种初始化过程可以看作是一种迁移学习, 即在网络创建的初始阶段就赋予深度网络有效的知识. 通过这种方式得到的 KBSDAE 网络具有更快的收敛速度和更高的识别精度. 混合规则包含符号规则和分类规则, 它们分别对应初始化 KBSDAE 的编码部分和分类器部分, 本节将分别讨论初始化 KBSDAE 不同训练阶段的方法.

3.1 DAE 学习阶段

基于 Tran 等[17] 的符号规则与 DBN 相结合的思想, 本文首先筛选出置信度较高的分类规则束, 其次利用符号规则初始化 DAE 网络, 最后利用特殊的参数更新策略完成训练.

符号规则束筛选算法 Rule-Screen 基于符号规则的置信度运行. 由于规则结构具有迭代嵌套特性, 可以看作一种类似树状结构, 所以将规则集分割为规则束以作为筛选目标. 每个规则束的置信度值通过推导算法得到, 最终选取置信度较高的规则用于初始化网络. 具体的算法流程如算法 2 所示, 将第 1 层符号规则集的置信度值作为输入在符号规则集中进行推导, 并最终筛选出置信度高的 n 个规则束. 通过这种方法将规则集中最可信的规则用于网络初始化, 在不造成信息过度损失的情况下简化初始化过程. 在本文中, 规则筛选率为 $1/n = 0.3$, 该参数的选取依据参照文献 [24]. 实验发现, 利用算法 2 得到的规则束囊括了各层符号规则中的高置信度规则, 这从另一个方面证明了算法的有效性.

算法 2. Rule-Screen 算法

输入. 符号规则集 λ^l

1) 对 c^1 进行归一化并赋值给 X, $X \leftarrow \text{Norm}(c^1)$
2) **For** $l = 2$ to N **do** // N 是隐层的数量
3) $B^l = \{\}$; // Initialize belief vector
4) **For** 初始化置信向量 λ_j^l **do**
5) 将 X_t, X_k 分别赋值给 α_t, $\alpha_k(t \in T, k \in K)$;
6) $\alpha = c_j^l \times (\sum_t \alpha_t - \sum_k \alpha_k)$;
7) 置信值 $B_j^l = \text{Norm}(\alpha)$;
8) **end For**
9) 将 B^l 赋值给 X;
10) **end For**
11) 对规则 B^N 进行排序, $RANK \leftarrow \text{sort}(B^N)$;
12) 基于 $RANK$ 排序从 λ^N 中选择 $1/n$ 个规则放入 SR_N
13) **For** $l = N - 1$ to 2 **do**
14) 在 λ^l 中选择与 SR_{l+1} 相关的规则放入 SR_l;
15) **end For**

输出. SR

由于符号规则的结构与 DAE 的编码阶段基本相同, 本文将规则中的元素和置信度值分别初始化为对应网络中的神经元和连接权重, 并且在网络训练过程中抑制被初始化参数的更新, 使知识得以保留. 具体训练步骤如下.

步骤 1. 建立一个 SDAE, 其结构和抽取规则的原网络相同. 对每一个规则 $c_j : h_j \leftrightarrow x_1 \wedge \cdots \wedge x_n$, h_j & $x_1 \wedge \cdots \wedge x_n$ 分别对应目标网络的隐藏层神经元以及输入层神经元集. 注意, 这些神经元集的排列位置应与原网络位置相同.

步骤 2. 确定在 h_j 与 $x_1 \cdots x_n$ 之间的连接权重. 如果输入神经元对应规则中的 x_p, 那么 $s = 1$; 反之, 则 $s = -1$. 其余的与 h_j 没有关联以及隐藏层与输出层之间的权重设为较小的随机值. 神经元偏差设为随机值.

步骤 3. 利用 BP 算法训练 DAE, 其中被规则初始化的连接权重不被更新. 为了保证代入的规则在训练过程中与网络较好嵌合, 利用随机数对隐藏层神经元输出进行二值化处理: 随机生成一个数值为 0-1 的随机数 R, 如果 $h_j > R$, 那么 $h_j = 1$; 反之, 则 $h_j = 0$.

步骤 4. 自上而下地对每一个 DAE 重复执行步骤 1 ~ 3, 直到编码部分训练完成.

3.2 Fine-tuning 阶段

分类规则的结构形式与 SDAE 的分类层结构不同, 不能直接进行初始化过程. 本文将 KBANN[7] 的相关知识融入到 KBSDAE 的分类层中, 训练过程如图 3 所示. 但是本文所定义的分类规则包含实数集和符号集, 而只针对符号规则的 KBANN 方法在这里并不能完全适用. 为了将分类规则中的实数知识初始化入网络中, 本文借鉴了 Fernando 等[10] 提出的 INSS 系统思想, 利用实数型规则初始化网络.

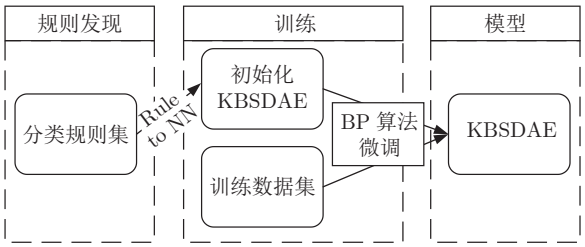


图 3 KBSDAE 的 Fine-tuning 阶段示意图

Fig. 3 Fine-tuning diagram of KBSDAE

初始化分类层的核心步骤是, 利用分类规则在分类层和 DAEs 中间新增加一个隐藏层, 定义为 l^{N+1}. 对每一个规则基本元素 IF $(B^N_{\text{selectn}} > V^N_n$ or $B^N_{\text{select}n} > V^{N'}_n)$ THEN $y = cl_j$, B^N_{selectn} 对应网络中 DAEs 部分最后一个隐藏层的神经元, V^N_n 对应新建隐藏层 l^{N+1} 中的神经元, 类标签值 cl_j 对应分类层神经元. 具体的权值和偏置值初始化方法如图 4 所示. 图 4 以一条分类规则为例, 展示了这条规则初始化网络的全过程, 其中 P 代表连向该单元的神经元数目, w 和 C 分别代表分类规则的灵敏度和可信度. 为了更清晰地表达, 图 4 省略了大部分连接线. 本文中定义 $Weight_2 = 1$. 需要注意的是, 为了提高被初始化网络的泛化性, 微调过程中在新隐藏层 l^{N+1} 和编码阶段顶层 l^N 分别新加 25 个神经元, 这种操作在 KBANN[7] 中被证明是有效的.

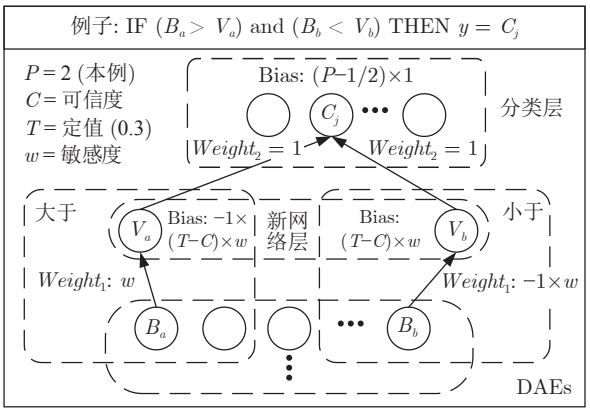


图 4 分类规则初始化网络算法示意图

Fig. 4 Classification rule initialize network algorithm diagram

4 实验与结果分析

4.1 规则有效性验证

为验证抽取知识的有效性. 利用数据关系已知的简单数据训练自编码器, 并从编码过程中抽取知识. 如果抽取出规则所代表的知识与已知数据关系相符, 那么可以证明从网络中抽取特征间知识的有效性. 本文利用具有异或关系的数据对自编码器 AE 进行训练, 并从编码过程中抽取符号规则集. 该实验的目的在于验证从不同规模的 AE 提取出的符号规则是否也具有异或特性. 输入数据具有三个维度 x, y, z, 它们具有 $z \leftrightarrow x \oplus y$ 关系的同时还具有 $x \leftrightarrow y \oplus z$ 和 $y \leftrightarrow x \oplus z$ 关系. 通过训练让 DAE 学习异或关系.

从不同 DAE 中提取的部分符号规则如表 2 所示. 其中与输入数据关系相符的规则具有较大的置信度, 反之只具有很小的置信度, 并且从不同规模网络中抽取出的规则都基本符合数据关系. 这证实了本文所提出的符号规则体系可以有效拓扑并抽取出网络中的知识. 尽管抽取过程存在一定的信息损失, 但是依旧可以很好地表示 DAE 的编码过程.

表 2 部分 DAE 符号规则抽取结果

Table 2 DAE symbol rule extraction result

隐藏单元	DAE 的置信度符号规则束
3	$2.2874 : h_2 \leftrightarrow \neg x \wedge y \wedge z$ $2.9129 : h_3 \leftrightarrow \neg x \wedge \neg y \wedge \neg z$
10	$1.4163 : h_1 \leftrightarrow \neg x \wedge \neg y \wedge \neg z$ $2.4803 : h_2 \leftrightarrow \neg x \wedge \neg y \wedge \neg z$ $1.9159 : h_3 \leftrightarrow x \wedge \neg y \wedge z$ $1.0435 : h_4 \leftrightarrow \neg x \wedge \neg y \wedge \neg z$ $0.6770 : h_5 \leftrightarrow \neg x \wedge y \wedge \neg z$ $1.9298 : h_6 \leftrightarrow x \wedge \neg y \wedge z$ $1.9785 : h_7 \leftrightarrow x \wedge \neg y \wedge z$ $1.9448 : h_8 \leftrightarrow \neg x \wedge y \wedge z$ $2.4405 : h_9 \leftrightarrow x \wedge y \wedge \neg z$ $2.0322 : h_{10} \leftrightarrow \neg x \wedge y \wedge z$

为了验证符号规则是否也能够像 DAE 一样提取特征, 本文利用三种特征维度较高且数据特征与标签间的关系不明显的数据集分别对不同规模的 DAEs 进行训练并抽取规则. 一般情况下, 相较于原始数据, 经过 DAE 编码后输出的数据具有更低的维度, 而利用这种低纬度数据训练的支持向量机 (Suppprt vector machine, SVM) 一般具有更好的分类结果. 如果相应置信度符号规则对输入的原始数据经过推导后所输出的低纬度数据也可以提升 SVM 的分类结果, 那么可以证明从 DAE 中提取出来的规则具备同样的特征提取能力.

利用 DAE 无监督训练后输出的降维数据和对应规则推导得到的置信值集合分别对 SVM 进行训练并测试分类性能, 其中 SVM 选用线性核函数. 数据集包含 MNIST 手写数字数据集、HAR 人体活动识别数据集、数据集的图像中每个像素有 256 (0 ~ 255) 灰度级, 将这些图像归一化到区间 [0, 1]. MNIST 数据集包含 60 000 个训练数据和 10 000 个

测试数据集. 为了提高训练效率, 本文首先选取 10000 个训练数据和 2 000 个验证数据训练初始网络. 之后选取分类效果最好的初始网络并利用完整的 60000 个数据进行训练. 最后再利用 10 000 个测试数据进行测试. HAR 数据包含由智能手机收集的 561 维特征以及对应的 6 种人物动作, 有 8 239 个训练数据集和 2 060 个测试数据集. 训练模型的参数: DAE 的学习率为 0.01 ~ 1 之间, 噪声率在 0.2 ~ 0.5 之间. SVM 的惩罚系数在 0.01 ~ 1 之间.

利用 SVM 对不同数据进行 10 折交叉训练的结果如表 3 所示. 相较于原始数据, 经过 DAEs 或对应符号规则提取特征的低维数据具有更好的可分性. 虽然规则的特征提取能力和相应 DAEs 之间存在差距, 但是从某种意义上讲, DAEs 中抽取的规则可以在一定程度上代表 DAEs, 故规则也具有稳定降维的能力.

表 3 复杂数据集降维后 SVM 10 折交叉分类结果 (%)
Table 3 Ten-fold cross-classification results of dimensionally reduced complex data on SVM (%)

	MNIST	HAR
One DAE ($J = 500$)	98.00	97.27
Symbolic rule	94.43	96.73
Two DAEs (top $J = 100$)	98.74	98.07
Symbolic rule	96.03	96.84
Two DAEs (top $J = 200$)	98.90	97.74
Symbolic rule	95.42	97.33
SVM (linear)	92.35	96.55

4.2 混合规则可推导解释性验证

为了验证混合规则的可推导性和可解释性, 利用 DNA promoter[25] 数据集训练网络并抽取规则. 该数据集普遍应用于验证网络−符号系统, 数据集具有 106 例数据, 分成 53 个激活项数据和 53 个不激活项数据, 每一例数据由 DNA 的−50 ~ 7 位置上的染色体状态组成 (A, T, G, C). 将染色体状态进行 one-hot 编码以方便网络学习, 如: A = (1, 0, 0, 0), T = (0, 1, 0, 0) 等. 经过数据处理, 原数据变为 228 维数据. 随机选取 86 例训练数据, 20 例预测数据. 首先利用数据分别训练 5 个单层 SDAE, 网络规模为 228-3-2 (即: 单层 SDAE 的输入神经元个数为 228 (输入数据为 228 维数据), SDAE 的隐层神经元个数为 3, 输出层神经元个数为 2 (根据分类数确定, 此处为激活项与不激活项两类)), 两个训练阶段的学习率分别为 0.01 和 1, DAE 训练阶段噪声率为 0.1. 这些网络在测试集上的平均识别精度为 90%, 而对应抽取出的规则在测试集上的平均推导识别精度为 91.52%. 具体的分类规则如表 4 所示, 可以看出本文提出的混合规则具有较高的推导判别性能.

表 4 基于 DNA promoter 的分类规则明细表 (%)
Table 4 Classification rule schedule based on DNA promoter (%)

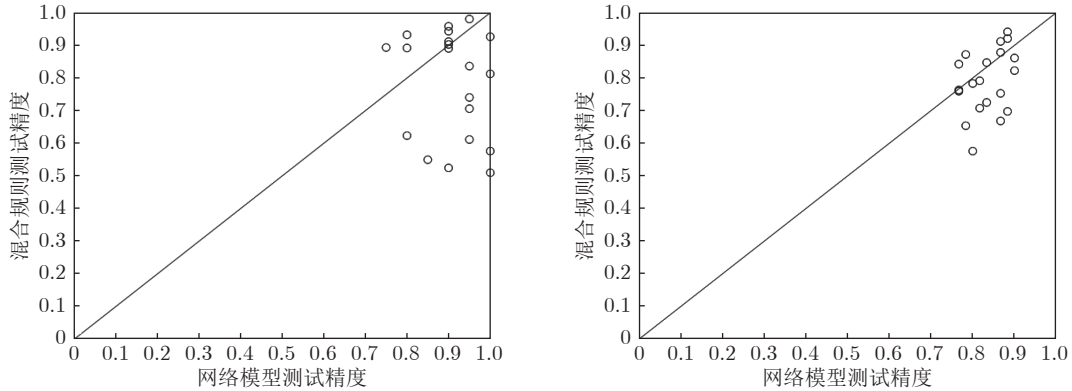
分类规则	可信度	覆盖率
IF $(h_2^1 > 0.771 \wedge h_3^1 > 0.867)$ THEN $promoter$	98.62	50.00
IF $(h_1^1 < 0.92 \wedge h_2^1 < 0.634 \wedge h_3^1 < 0.643)$ THEN $\neg promoter$	84.42	50.00

在模型的可解释方面, 本文所提出模型的核心思想为符号与网络相结合, 而规则就是对网络的一种模拟和解释. 接下来我们将尝试利用规则对网络进行解释并验证. 基于 DNA promoter 数据集的模型规则如表 3 和表 4 所示. 表 5 包含了 3 个隐藏神经元所对应的部分规则以及 DNA promoter 数据集固有的基本规则, 其中 "[ ]"表示为任意碱基都成立; "T (A)"表示 T 或 A 在该位置出现都算假设成立. 以 h_2^1 为例解释网络行为: 如果输入数据在规则片段 1 对应位置的碱基排列为 "T T G (T or A) C"; 在规则片段 2 对应位置的碱基排列为 "(T or A) A A A G C"; 在规则片段 3 对应位置的碱基排列为 "A A T A A", 那么 DAE 的隐藏层神经元 h_2 的输出值则尽可能大. 根据表 4 的分类规则, 如果 h_2 和 h_3 足够大, 就将这一数据定义为 "promoter". 上述的过程为规则推导的过程, 也在一定意义上解释了神经网络内部的运行机理. 为了验证这种解释是否正确, 将抽取出的规则与数据的固有基本规则进行对比, 可以发现基本规则与生成规则所对应的地方基本相符, 这也验证了利用规则抽取网络内部知识的有效性.

对比规则和标准网络在相同测试数据下的推导精度, 以验证规则的高精度推导能力和稳定性. 根据不同训练数据量, 分别连续训练 20 个标准单层 DAE 网络并从中抽取规则. 利用 20 例测试数据分别对每一个网络和其对应的规则进行精度测试, 结果如图 5 所示. 图中横坐标表示标准网络在测试集上的预测精度, 纵坐标表示规则在测试集上的推导精度. 可以看出, 测试结果大部分都落在对角线周围, 这证明规则推导精度和对应的标准网络大体相当, 且具有较高稳定性. 即便训练数据量发生变化, 推导精度也不会发生突变.

表 5 基于 DNA promote 数据集的部分符号规则明细
Table 5 Partial symbol rule details based on DNA promote

节点	置信度	规则片段 1	起始位 @-36	终止位 @-32	规则片段 2	起始位 @-12	终止位 @-7	规则片段 3	起始位 @-45	终止位 @-41
h_1^1	0.76	A C [ ] G T			G G T C (T) G C			G C T A T (A)		
h_2^1	1.29	T T G T (A) C			T (A) A A A G C			A A T A A		
h_3^1	1.47	G (A) T G T (A) C			T (A) T (A) G (A) T C (T) G (A)			A A T C A		
基本规则	L1	minus-35: T T G A C			minus-10: T A [ ] [ ] [ ] T			Conformation: A A [ ] [ ] A		
	L2	contact←minus-35∧minus-10								
	L3	promoter←contact∧conformation								



(a) 86 例数据训练后模型预测精度对比
(a) Comparison of model prediction accuracy after 86 data training

(b) 46 例数据训练后模型预测精度对比
(b) Comparison of model prediction accuracy after 46 data training

图 5 SDAE 和对应混合规则 DNA promoter 的识别率对比 (%)
Fig. 5 Comparison of DNA promoter recognition rate between SDAE and corresponding blending rules (%)

4.3 标准数据集验证

KBSDAE 在构建过程中创造性地同时利用两种不同类型的规则进行网络确定与权重参数初始化, 其网络结构以及训练方法都与传统 SDAE 有所不同. 因此, 本文采用 UCI 数据库[26] 中的经典标准数据 (如表 6 所示) 对 KBSDAE 的分类性能进行测试, 并与其他典型分类器进行比较.

KBSDAE 的网络结构为两个 DAE 堆叠 (第 1 个 DAE 隐藏层神经元数稍大于输入特征数, 第 2 个 DAE 隐藏层神经元数稍小于输入特征数), 外加一层由分类规则初始化出的隐藏层 l^{N+1}, 该层神经元数根据分类规则变化, 最后堆叠一个 Softmax 分类层. 两个训练阶段的学习率在 0 ~ 0.01 之间, 噪声率在 0.01 ~ 0.3 之间. DAE 迭代训练 200 次, Fine-tuning 阶段迭代 50 次, 其余分类器的迭代次数与之相似. 对比的网络有 DBN, 网络结构为两个 RBM 堆叠, 其结构与 KBSDAE 相似, 两个训练阶段学习率在 0.1 ~ 1 之间, RBM 训练阶段的动量为 0.05 ~ 0.3; Sym-DBN 模型来自于文献 [17], 是一种符号规则和 DBN 相结合的模型, 其网络结构与 DBN 相同, 两阶段学习率在 0.01 ~ 0.2 之间, RBM 训练阶段的动量为 0.05 ~ 0.3. SDAE 为两层 DAE 堆叠而成, 网络结构与参数和 KBSDAE 相同. INSS-KBANN[10] 中的规则由 GA 算法直接从训练数据集中挖掘, 包含一个隐藏层, 学习率在 0.1 ~

表 6 UCI 数据集信息
Table 6 UCI dataset information

数据集	特征数量	类别数	数据量
Credit card	14	2	690
Diabetes	8	2	768
Pima	8	2	759
Wine	13	3	178
Cancer	9	2	689
Vehicle	8	4	846
Heart	13	2	270
German	24	2	1 000
Iris	4	3	150

1 之间, 迭代训练 200 次. BPNN 模型结构与 KBSDAE 相同, 学习率在 0.1 ~ 1 之间, 迭代训练 300 次.

如表 7 所示, 其中所有识别器的测试结果都是经过 5 折交叉[27] 后得到. 在实验过程中, 将数据随机等分成 5 份, 其中 4 份作为训练数据, 另外一份作为测试数据. 实验连续进行 5 次, 保证每一份数据都成为过测试集和训练集. 实验结果证明, KBSDAE 的分类性能相较于传统的机器学习分类器有较为明显的提升. 特别地, KBSDAE 明显优于 SDAE, 这说明符号规则与分类规则融合的 SDAE设计方法显著提高了其特征学习与识别性能.

知识代入过程中对网络的影响在于初始化阶段. 网络的初始化对网络的训练过程具有较大的影响[7]. 为了验证利用知识初始化网络是否可以带来积极影响, 本文利用 HAR 数据集建立 KBSDAE 网络, 并记录了 KBSDAE 在无监督训练和微调阶段的均方误差 (Mean square error, MSE) 变化. MSE 可以很好地描述网络在训练过程中的分类性能变化. 从图 6 可以看出, 无论是在无监督还是在微调阶段, KBSDAE 的 MSE相较于规模相同的 SDAE 都具有更低的起点、更快的收敛速度和更低的收敛区间. 这证明了利用知识初始化网络所带来的积极影响, 进一步证明了本文提出方法的有效性.

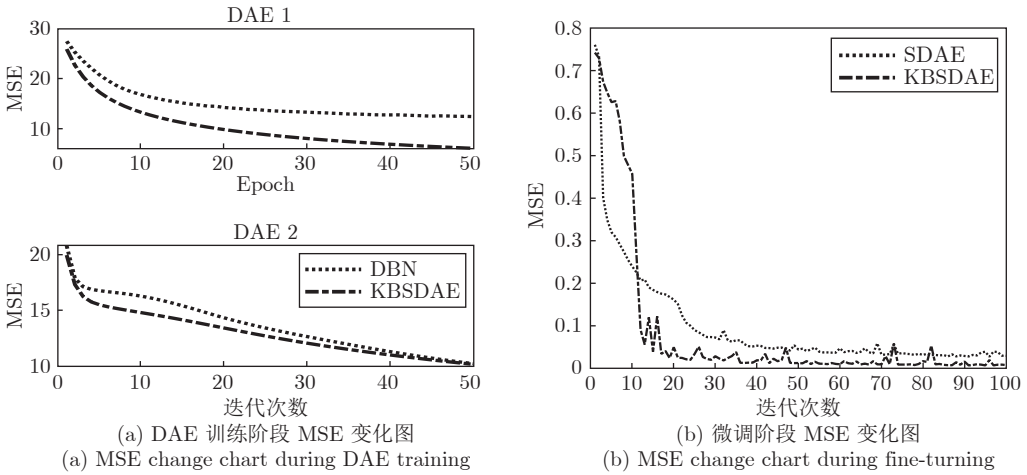
为了进一步验证 KBSDAE 模型在处理复杂分类问题时能否有效进行分类, 选取了两种标准数据集: USPS 手写数据集以及 HAR 数据集[28]. 对比了 SDAE 和 KBSDAE 的类识别性能. 具体的实验数据如表 8 所示. 从表中可以看出, 在结构参数基本相同的情况下, KBSDAE 与 SDAE 的分类情况类似, 故认定 KBSDAE 不存在类识别不平衡问题. 值得注意的是, 文献 [27] 利用改进的支持向量机处理 HAR 数据集的分类问题, 识别精度最高仅达到 89.3%, 而 KBSDAE 可以达到 98.5% 的识别精度.

为了验证 KBSDAE 在复杂数据集上的分类性能是否足够优越. 本文选取了目前较为主流的一维分类器和符号神经模型在 USPS 和 HAR 数据集上

表 7　UCI 数据集信息 (%)

Table 7　UCI dataset information (%)

数据集	DBN	SDAE	INSS-KBANN	BPNN	Sym-DBN	KBSDAE
Credit card	84.29	84.14	81.17	85.00	85.57	87.18
Diabetes	73.20	73.47	74.00	72.40	76.53	78.27
Pima	72.57	70.00	73.73	73.73	74.00	76.40
Wine	96.67	96.00	97.67	96.00	98.00	97.00
Cancer	96.92	97.38	97.21	96.31	97.69	97.12
Vehicle	75.29	73.97	71.82	68.69	74.67	75.85
Heart	81.60	76.80	78.80	78.40	82.40	84.00
German	70.90	71.30	71.60	69.40	71.30	79.10
Iris	84.00	82.00	93.00	92.33	92.33	94.33



(a) DAE 训练阶段 MSE 变化图　(b) 微调阶段 MSE 变化图

(a) MSE change chart during DAE training　(b) MSE change chart during fine-turning

图 6　KBSDAE 和 SDAE 在 HAR 数据集上训练过程的均方误差变化对比

Fig. 6　Comparison of mean square error of KBSDAE and SDAE training process on HAR dataset

表 8 复杂数据集分类结果对比 (%)

Table 8 Classification results of comparison on complex datasets (%)

数据集及网络参数	数据标签	SDAE	KBSDAE
USPS:			
SDAE: 256-100-20-10/learning rate: 0.01/noising rate: 0.1 KBSDAE: 256-100-25-18-10/learning rate: 0.01/noising rate: 0.1	0	98.97	98.38
	1	99.13	99.06
	2	96.46	97.28
	3	96.37	95.20
	4	96.33	96.74
	5	94.19	94.53
	6	97.62	97.97
	7	97.49	97.44
	8	94.46	96.61
	9	98.25	97.88
Mean		97.24	97.33
HAR			
SDAE: 561-300-20-6/learning rate: 0.01/noising rate: 0.1 KBSDAE: 561-300-25-11-6/learning rate: 0.01/noising rate: 0.1	Walking	98.84	100.00
	Walking upstairse	88.17	98.77
	Walking downstairse	92.31	98.49
	Sitting	97.05	98.85
	Standing	89.36	94.99
	Laying	100	100.00
Mean		94.10	98.50

进行分类结果横向对比. 结果如表 9 所示, 其中所有结果都是 5 折交叉后得到的. 可以看出, KBSDAE 相较于其他分类模型具有更好的分类效果.

表 9 复杂数据集 5 折交叉分类结果对比 (%)

Table 9 Comparison of five-fold cross-classification results on complex datasets (%)

数据集	KBSDAE	Sym-DBN	DBN	SDAE	BPNN	SVM
USPS	97.43	97.47	96.72	97.24	97.22	93.37
HAR	98.40	97.09	96.89	95.32	95.84	96.55

4.4 灵敏度分析

为了验证规则中的知识是否能赋予网络一定的分类性能以及 KBSDAE 对数据的敏感度, 对比了不同训练数据量下网络模型的预测精度. 本文利用 DNA promoter 数据分别训练 SDAE 和 KBSDAE. 训练数据量从 10 开始逐渐递增. 训练后的网络利用 20 个测试数据进行识别性能测试, 结果如图 7 所示, 在训练数据量很小的情况下, KBSDAE 依旧具有高识别精度, 这是由于知识代入网络的结果. 并且, 随着训练数据量的增加 KBSDAE 识别精度也稳定高于传统 SDAE.

为进一步验证知识代入网络的过程是否有效, 对比了 KBSDAE 和 SDAE 在不进行 Fine-tuning 和只进行几步 Fine-tuning 后的测试精度. 利用 DNA promoter 数据集分别建立了结构和训练参数相同的 SDAE 和 KBSDAE, 其中两个训练阶段的

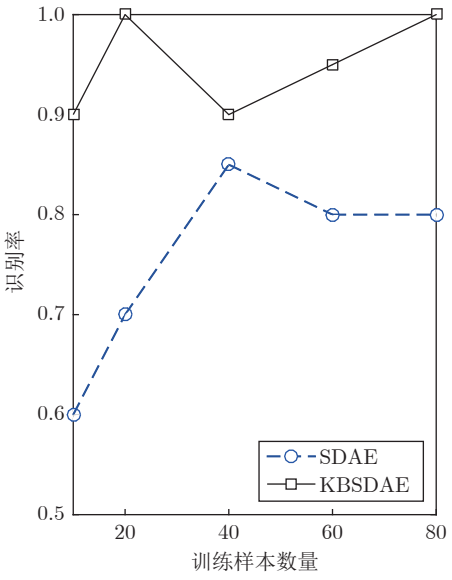


图 7 不同 DNA promoter 数据量训练的 SDAE 与 KBSDAE 分类性能对比

Fig. 7 Comparison of classification performance between SDAE and KBSDAE trained by different DNA promoter data

学习率分别为 0.01 和 1, DAE 训练阶段噪声率为 0.1. 实验结果如图 8 所示, 可以看到 KBSDAE 在不进行 Fine-tuning 的情况下仍具有 80% 的测试精度, 与 SDAE 的 51% 测试精度相比提升明显, 这进一步证明了利用规则将知识代入网络中的方法是有效的. 经过前几步 Fine-tuning 后的 KBSDAE 测试精度普遍高于 SDAE, 证明了将知识代入网络可以显著提高网络的分类性能.

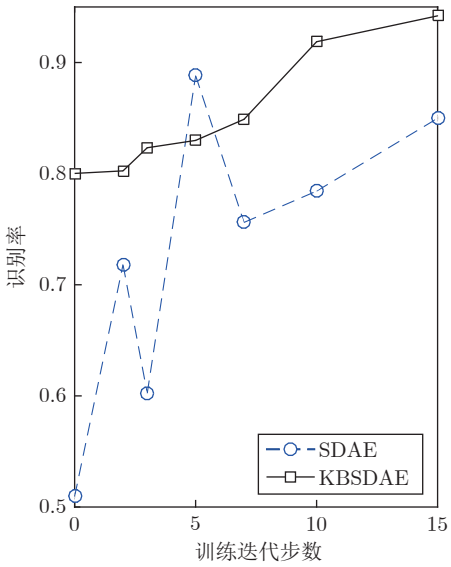


图 8 不同 Fine-tuning 训练步数的 SDAE 与 KBSDAE 分类性能对比

Fig. 8 Comparison of SDAE and KBSDAE classification performance of different fine-tuning training steps

5 结束语

面对深度神经网络的 “黑箱问题”, 本文提出了一套全新的知识表达规则系统, 尝试解释并强化深度神经网络. 该系统可以对 SDAE 网络进行简单表示并从 SDAE 中抽取和插入知识. 通过这套系统, 可以理解到网络内部的知识并建立性能更加强大的 KBSDAE. 规则系统创新性的将符号类型和数值类型的规则进行有机结合, 使得这种混合规则具有较高的推导性能和可理解性, 在网络规模愈加复杂的当下这种规则形式不失为一条具有研究价值的路径. 实验证明, 混合规则系统可以有效表示网络并提取网络知识, 具有高推导精度和稳定性. 利用规则初始化后的 KBSDAE 相较于传统 SDAE 具有更快的收敛速度, 更高的预测精度和数据灵敏度. 下一步可以将规则与深度网络可视化相结合以提升对网络的解释能力. 可以尝试解释更复杂和庞大的网络, 如卷积神经网络.

References

1 Towell G G. Extracting refined rules from knowledge-based neural networks. *Machine Learning*, 1993, **13**(1): 71−101

2 Lecun Y, Bengio Y, Hinton G E. Deep learning. *Nature*, 2015, **521**(7553): 436−444

3 Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets. *Neural Computation*, 2006, **18**(7): 1527−1554

4 Bengio Y, Lamblin P, Popovici D, Larochelle H. Greedy layer-wise training of deep networks. In: Proceedings of the 2006 Advances in Neural Information Processing Systems 19, Proceedings of the 20th Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4−7, 2006. DBLP, 2007.

5 Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998, **86**(11): 2278−2324

6 Vincent P, Larochelle H, Bengio Y, Manzagol P A. Extract-ing and composing robust features with denoising autoen-coders. In: Proceedings of the 25th International Conferenceon Machine Learning. Helsinki, Finland: ACM, 2008. 1096−1103

7 Towell G G, Shavlik J W. Knowledge-based artificial neural networks. *Artificial Intelligence*, 1994, **70**(1−2): 119−165

8 Gallant S I. Connectionist expert systems. *Communications of the ACM*, 1988, **31**(2): 152−169

9 Garcez A D A, Zaverucha G. The connectionist inductive learning and logic programming system. *Applied Intelligence: The International Journal of Artificial, Intelligence, Neural Networks, and Complex Problem-Solving Technologies*, 1999, **11**(1): 59−77

10 Fernando S. Osório, Amy B. INSS: A hybrid system for constructive machine learning. *Neurocomputing*, 1999, **28**(1−3): 191−205

11 Setiono R. Extracting rules from neural networks by pruning and hidden-unit splitting. *Neural Computation*, 2014, **9**(1): 205−225

12 Yang Li, Yuan Jing, Hu Shou-Ren. The problem solving mechanism of neural networks. *Chinese Journal of Compuers*, 1993, (11): 814−822
(杨莉, 袁静, 胡守仁. 神经网络问题求解机制. 计算机学报, 1993, (11): 814−822)

13 Qian Da-Qun, Sun Zhen-Fei. Knowledge acquisition and behavioral explanation on neural network. *Acta Automatica Sinica*, 1994, **20**(3): 348−351
(钱大群, 孙振飞. 神经网络的知识获取与行为解释. 自动化学报, 1994, **20**(3): 348−351)

14 Li Ming, Zhang Hua-Guang. Research on the method of neural network modeling based on rough sets theory. *Acta Automatica Sinica*, 2002, **28**(1): 27−33
(黎明, 张化光. 基于粗糙集的神经网络建模方法研究. 自动化学报, 2002, **28**(1): 27−33)

15 Penning L D, Garcez A S D, Lamb L C, Meyer J J C. A neural-symbolic cognitive agent for online learning and reasoning. In: Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, 2011. 1653−1658

16 Odense S, Garcez A S D. Extracting *M* of *N* rules from restricted Boltzmann machines. In: Proceedings of the 2017 International Conference on Artificial Neural Networks. Springer, Cham, 2017. 120−127

17 Tran S N, Garcez A S D. Deep logic networks: Inserting and extracting knowledge from deep belief networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2018, **29**(2): 246−258

18 Li S, Xu H R, Lu Z D. Generalize symbolic knowledge with

neural rule engine. arXiv preprint arXiv: 1808.10326v1, 2018.

19 Garcez A D A, Gori M, Lamb L C, Serafini L. Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. arXiv preprint arXiv: 1905.06088, 2019.

20 Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors. *Nature*, 1986, **323**: 533−536

21 Pinkas G. Reasoning, nonmonotonicity and learning in connectionist networks that capture propositional knowledge. *Artificial Intelligence*, 1995, **77**(2): 203−247

22 Mitchell T M. *Machine Learning*. McGraw-Hill, 2014.

23 Yu J B, Xi L, Zhou X. Deep logic networks: Intelligent monitoring and diagnosis of manufacturing processes using an integrated approach of KBANN and GA. *Computers in Industry*, 2008, **59**(5): 489−501

24 Tran S N, Garcez A S D. Knowledge extraction from deep belief networks for images. In: Proceedings of the 2013 IJCAI-Workshop Neural-Symbolic Learning and Reasoning, 2013. 1−6

25 Towell G G, Shavlik J W. The extraction of refined rules from knowledge-based neural networks. *Machine Learning*, 2018, **13**(1): 71−101

26 Murphy P M, Aha D W. UCI repository of machine learning databases. *Depth Information Compute Science*, University California, Irvine, CA, 1994.

27 Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence. San Francisco, CA, USA: Morgan Kaufmann, 1995. 1137−1143

28 Anguita D, Ghio A, Oneto L, Parra X, Reyes-Ortiz J L. Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine. In: Proceedings of the 4th International Workshop of Ambient Assited Living, IWAAL 2012, Vitoria-Gasteiz, Spain, 2012. 216−223

刘国梁 同济大学机械与能源工程学院硕士研究生. 2018 年获上海大学机械工程及其自动化学院学士学位. 主要研究方向为机器学习, 深度学习, 智能质量管控.
E-mail: guoliangliutt@163.com
(**LIU Guo-Liang** Master student at the School of Mechanical and Energy Engineering, Tongji University. He received his bachelor degree from Shanghai University in 2018. His research interest covers machine learning and intelligent quality control.)

余建波 同济大学机械与能源工程学院教授. 2009 年获上海交通大学机械工程学院博士学位. 主要研究方向为机器学习, 深度学习, 智能质量管控, 过程控制, 视觉检测与识别. 本文通信作者.
E-mail: jbyu@tongji.edu.cn
(**YU Jian-Bo** Professor at the School of Mechanical and Energy Engineering, Tongji University. He received his Ph.D. degree from Shanghai Jiao Tong University. His research interest covers machine learning, deep learning, intelligent quality control, process control, and visual inspection and identification. Corresponding author of this paper.)