基于深度信念网络和线性单分类 SVM 的高维异常检测

李昊奇，应娜，郭春生，王金华
（杭州电子科技大学，浙江 杭州 310018）

摘　要：针对目前高维数据异常检测存在的困难，提出一种基于深度信念网络和线性单分类支持向量机的高维异常检测算法。该算法首先利用深度信念网络具有良好的特征提取功能，实现高维数据的降维，然后基于线性核函数的单分类支持向量机实现异常检测。选取 UCI 机器学习库中的高维数据集进行实验，结果表明，该算法在检测正确率和计算复杂度上均有明显优势。与 PCA-SVDD 算法相比，检测正确率有 4.65%的提升。与自动编码器算法相比，其训练和测试时间均有显著下降。

关键词：异常检测；高维数据；深度信念网络；单分类支持向量机
中图分类号：TP183
文献标识码：A
doi: 10.11959/j.issn.1000−0801.2018006

High-dimensional outlier detection based on deep belief network and linear one-class SVM

LI Haoqi, YING Na, GUO Chunsheng, WANG Jinhua
Hangzhou Dianzi University, Hangzhou 310018, China

Abstract: Aiming at the difficulties in high-dimensional outlier detection at present, an algorithm of high-dimensional outlier detection based on deep belief network and linear one-class SVM was proposed. The algorithm firstly used the deep belief network which had a good performance in the feature extraction to realize the dimensionality reduction of high-dimensional data, and then the outlier detection was achieved based on a one-class SVM with the linear kernel function. High-dimensional data sets in UCI machine learning repository were selected to experiment, result shows that the algorithm has obvious advantages in detection accuracy and computational complexity. Compared with the PCA-SVDD algorithm, the detection accuracy is improved by 4.65%. Compared with the automatic encoder algorithm, its training time and testing time decrease significantly.

Key words: outlier detection, high-dimensional data, deep belief network, one-class SVM

1　引言

异常检测是数据挖掘中的重要组成部分。异常数据是指在数据集中偏离大部分数据或者与数据集中其他大部分数据不服从相同统计模型的小部分数据[1]。而异常检测就是要识别出异常数据

收稿日期：2017−06−21；**修回日期**：2017−09−26
基金项目：国家自然科学基金资助项目（No.61372157）；“电子科学与技术”浙江省一流学科 A 类基金资助项目（No.GK178800207001）
Foundation Items: The National Natural Science Foundation of China(No.61372157), Zhejiang Provincial First Class Disciplines: Class A-Electronic Science and Technology (No.GK178800207001)

从而消除不符合预期行为的模式问题。异常检测在信用卡欺诈、网络入侵、健康医疗监控等诸多生活领域中均有重要应用[2]。

在异常检测中，单分类支持向量机（one-class support vector machine，OCSVM）是常用的有效手段[3]。OCSVM 是对二分类支持向量机的一种细化，是在异常检测领域中的重要经典算法。当确定合适的参数配置时，OCSVM 对于异常数据的检测可以提供良好的泛化能力。在 OCSVM 中，有两种经典算法用于异常检测，分别为基于超平面支持向量机（plane based support vector machine，PSVM）和基于超球面的支持向量描述（support vector data description，SVDD）法。相比较而言，利用超球面分类的 SVDD 算法性能优于基于 PSVM 算法。因此，通常采用 SVDD 算法进行异常检测。

然而，随着互联网的快速发展和物联网的逐渐普及，数据的收集更加容易。这导致数据库的规模和数据的复杂性急剧增加，从而产生大量的高维数据。如证券交易数据、Web 用户数据、网络多媒体数据等。维度的迅速增长，使得传统的 OCSVM 方法对高维数据的异常检测效率逐渐下降，从而导致高维数据的异常检测成为数据挖掘的难点[4]。

高维数据存在的普遍性使得对高维数据挖掘的研究有着非常重要的意义。但“维度灾难”问题导致对高维数据挖掘变得异常困难。即在分析高维数据时，所需的空间样本数会随维数的增加而呈指数倍增长。对于高维数据的处理，传统的多元统计分析方法存在很多的局限性，同时高维数据空间中的稀疏性使得采用非参数方法的大样本理论也并不适用。因此，采用数据降维是处理高维数据的最主要的高效手段。

在机器学习领域中，所谓降维就是指采用某种映射方法，将原高维空间中的点映射到新的低维空间中[5]。经典的数据降维方法如主成分分析[6]（principle component analysis，PCA）法、局部线性嵌入[7]（locally linear embedding，LLE）法和典型相关分析[8]（canonical correlation analysis，CCA）法等在特征提取和数据降维方面有着广泛的应用。但这些降维方法均属于线性降维，只能提取数据间的线性关系，从而导致在处理高维数据时存在着统计特性的渐进性难以实现、算法顽健性低等问题。尽管对 PCA 和 CCA 基于核函数改进后的核主成分分析（kernel principle component analysis，KPCA）法和核典型相关分析[9]（kernel canonical correlation analysis，KCCA）法可以解决非线性降维的问题，但算法的复杂度较高、效率较低。

对于解决高维的异常检测问题，近几年有多种经典的方法被提出。参考文献[10]直接提出了 OCSVM 中的经典算法，即基于超球面的支持向量数据描述法。该算法虽然对当时的高维数据异常检测起了很大的推动作用，但算法的正确率偏低。参考文献[11]将 PCA 算法和 OCSVM 相结合，将数据利用经典的线性降维方法 PCA 进行降维，在 OCSVM 中采用非线性核函数进行异常检测。由于线性降维的局限性，其结果并没有很大的提升。参考文献[12]利用改进后的 KPCA 算法和 OCSVM 进行异常检测。检测结果虽有所提升，但由于非线性核函数计算量大，对数据进行训练和测试所需要的时间较长，导致该算法的效率不高。参考文献[13]利用自动编码器（autoencoder，AE），通过对比不同数据间的重构误差进行异常检测。其识别率虽有所提升，但测试效率依然不高。

本文提出利用深度信念网络（deep belief network，DBN）进行数据降维，再利用基于线性核函数的单分类支持向量机这种组合模型实现异常检测。深度信念网络本质上是一种概率生成模型，通过无监督的训练方法由底层至顶层逐层训练而成。与其他传统的线性降维方法相比，深度信念网络最大的特点就是利用其自身非线性的结构进行特征提取，将数据从高维空间映射至低维

空间，从而降低数据的维度。这种非线性降维方法可以在最大程度上保留原始数据的高维特征，并且算法的复杂度较低，相比于其他算法可以更有效地解决高维数据的异常检测问题。实验结果表明，本文提出的混合算法模型，即将深度信念网络和线性单分类支持向量机组合在一起解决高维数据的异常检测问题，在检测正确率和测试效率上都有很大提升。

2 算法设计

本文所提出的算法（DBN-OCSVM）模型如图 1 所示，该模型由两部分组成，即底层的 DBN 和顶层的 OCSVM。DBN 由 2 个限制玻尔兹曼机（restricted Boltzmann machine，RBM）堆叠而成。将原始数据首先输入 DBN 的输入层，经 RBM1 训练后，输入层数据被映射至隐藏层 1。隐藏层 1 的输出作为 RBM2 的输入继续训练后得到隐藏层 2。隐藏层 2 的数据即 DBN 的输出，并将其输入 OCSVM 中进行异常检测。

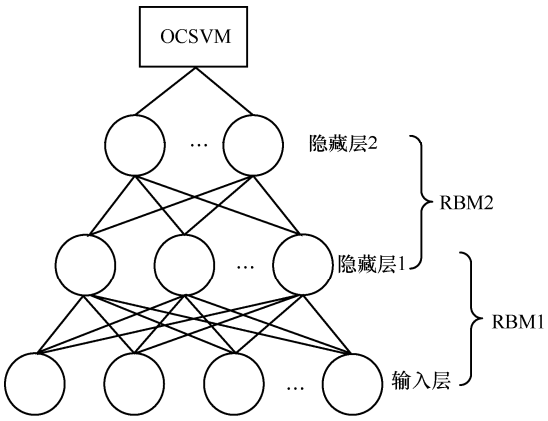


图 1 DBN-OCSVM 结构

在 OCSVM 中，使用 SVDD 算法进行异常检测。SVDD 为无监督训练算法，与有监督的二分类 SVM 相比，它并不是要寻找能够区分数据的最优超平面，而是寻找能够包含大多数正常数据的最优超球面。如图 2 所示，当输入空间的数据不可分时，构造一个映射函数，将输入空间中的数据映射到特征空间中。在特征空间中，寻找支持向量构造一个将绝大多数点包围在其中并具有最小半径的最优超球面。由支持向量确定的超球面即正常数据类的描述模型，超球面外的点被判断为离群类数据点，即异常数据。

在 SVDD 的核函数选取中，选择线性函数代替传统方法中的径向基函数（radical basis function，RBF）。在 SVM 中，核函数的选择对算法的性能起着重要的作用，利用核函数可以将线性不可分的输入空间映射到更高维的特征空间，从而将正常数据和异常数据进行完全分离。通常，相比较线性核函数而言，RBF 等非线性核函数可以将数据映射到更适于线性分类的特征空间，从而提高 SVM 的分类性能。但利用本文提出的模型，经 DBN 进行降维以及特征提取后的数据通过线性核函数依然可以进行优秀的分类，从而规避了线性核函数的缺点，反而突出了其优点。即降低了算法的时间复杂度和空间复杂度，提高了系统的运行速率。

3 算法原理

3.1 基于深度信念网络的高度降维

DBN 的实质是由一个高斯—伯努利型 RBM

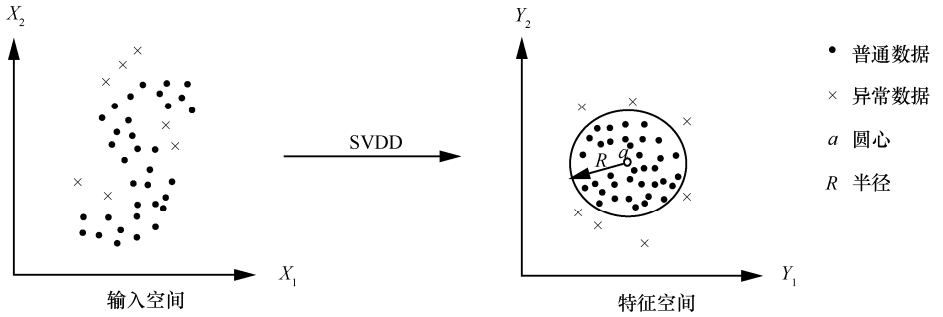


图 2 OCSVM-SVDD 算法示意

作为底层，上层接有多个伯努利—伯努利型 RBM，这样将多个 RBM 堆叠起来便得到了所需要的生成模型 DBN。将第一个 RBM 训练后得到的输出作为下一个 RBM 的输入继续训练，如此往复，经过训练后的各个 RBM 参数就是 DBN 的初始化参数。

限制玻尔兹曼机是玻尔兹曼机（Boltzmann machine）的一个特例。其本质上是由一层可见层神经元和一层隐藏层神经元所构成的双层无向神经网络模型。同时，可见层和隐藏层的各自每层网络之间的神经元是无连接的，二者构成一个二分图。对于给定的样本数为 m、维数为 n 的集合 $\boldsymbol{D}_{m\times n}$，经过深度信念网络后，得到样本数为 m、维数为 d 的集合 $\boldsymbol{X}_{m\times d}$。由于 $d \ll n$，从而达到数据降维的目的。

对于每一个 RBM，都有一个能力值与其对应。设可见层向量为 $\boldsymbol{v}$，隐藏层向量为 $\boldsymbol{h}$，当可见层神经元和隐藏层神经元取二进制值并服从伯努利分布时，即 $\boldsymbol{v}\in\{0,1\}$，$\boldsymbol{h}\in\{0,1\}$，此时的 RBM 为伯努利—伯努利型 RBM，其能量函数为：

$$E(\boldsymbol{v},\boldsymbol{h}) = -\sum_i v_i c_i - \sum_j h_j b_j - \sum_{i,j} w_{i,j} v_i h_j \tag{1}$$

其等价于：

$$E(\boldsymbol{v},\boldsymbol{h}) = -\boldsymbol{c}^{\mathrm{T}}\boldsymbol{v} - \boldsymbol{b}^{\mathrm{T}}\boldsymbol{h} - \boldsymbol{v}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{h} \tag{2}$$

其中，v_i 和 h_j 分别为可见层向量 $\boldsymbol{v}$ 和隐藏层向量 $\boldsymbol{h}$ 的第 i 个和第 j 个神经元。$\boldsymbol{W}$ 为连接可见层单元和隐藏层单元的权重矩阵，w_{ij} 为其中第 i 行第 j 列的元素。$\boldsymbol{c}$ 和 $\boldsymbol{b}$ 分别为可见层和隐藏层的偏置向量，c_i 和 b_j 为相应的第 i 个和第 j 个偏置单元。

当可见层单元取实数值，即 $\boldsymbol{v}\in\mathbb{R}$，并服从高斯分布时，此时的 RBM 为高斯—伯努利型 RBM。对于每一个配置 $(\boldsymbol{v},\boldsymbol{h})$，其能量函数为：

$$E(\boldsymbol{v},\boldsymbol{h}) = \frac{1}{2}(\boldsymbol{v}-\boldsymbol{c})^{\mathrm{T}}(\boldsymbol{v}-\boldsymbol{c}) - \boldsymbol{b}^{\mathrm{T}}\boldsymbol{h} - \boldsymbol{v}^{\mathrm{T}}\boldsymbol{w}\boldsymbol{h} \tag{3}$$

根据该能量配置函数，设定可见层和隐藏层的联合概率密度为：

$$P(\boldsymbol{v},\boldsymbol{h}) = \frac{\mathrm{e}^{-E(\boldsymbol{v},\boldsymbol{h})}}{Z} \tag{4}$$

其中，$Z=\sum_{\boldsymbol{v},\boldsymbol{h}}\mathrm{e}^{-E(\boldsymbol{v},\boldsymbol{h})}$ 为正规化因子，把 Z 称为配分函数。

目的是要找到 RBM 的势能最低点，使 RBM 达到稳态。即通过训练优化 RBM 的各项参数，使 $E(\boldsymbol{v},\boldsymbol{h})$ 达到最小值。而：

$$\begin{aligned} P(\boldsymbol{v}) &= \sum_{\boldsymbol{h}} P(\boldsymbol{v},\boldsymbol{h}) \\ &= \sum_{\boldsymbol{h}} \frac{\mathrm{e}^{-E(\boldsymbol{v},\boldsymbol{h})}}{Z} \end{aligned} \tag{5}$$

因此可以得出 $E(\boldsymbol{v},\boldsymbol{h})$ 的最小值即 $-P(\boldsymbol{v})$ 的最小值。采用随机梯度下降算法来极小化 $P(\boldsymbol{v})$ 的负对数似然度（negative log likelihood）：

$$\min J_{\mathrm{NLL}}(\boldsymbol{W},\boldsymbol{a},\boldsymbol{b};\boldsymbol{v}) = -\mathrm{lb}P(\boldsymbol{v}) \tag{6}$$

在 RBM 中负对数似然度对于任意一个模型参数的导数为：

$$\nabla_\theta J_{\mathrm{NLL}}(\boldsymbol{W},\boldsymbol{a},\boldsymbol{b};\boldsymbol{v}) = -\left[\left\langle\frac{\partial E(\boldsymbol{v},\boldsymbol{h})}{\partial\theta}\right\rangle_{\mathrm{data}} - \left\langle\frac{\partial E(\boldsymbol{v},\boldsymbol{h})}{\partial\theta}\right\rangle_{\mathrm{model}}\right] \tag{7}$$

其中，θ 为某个模型参数，$\langle x\rangle_{\mathrm{data}}$ 和 $\langle x\rangle_{\mathrm{model}}$ 分别是从数据和模型中估计的 x 的期望值。对于可见层—隐藏层神经元的权重，有：

$$\nabla_{w_{ij}} J_{\mathrm{NLL}}(\boldsymbol{W},\boldsymbol{a},\boldsymbol{b};\boldsymbol{v}) = -\left[\langle v_i h_j\rangle_{\mathrm{data}} - \langle v_i h_j\rangle_{\mathrm{model}}\right] \tag{8}$$

由于 $\langle\cdot\rangle_{\mathrm{model}}$ 这一项很难得到精确的计算，采用对比散度[14]（contrastive divergence，CD）算法估计此项。该算法利用 k 步吉布斯采样后得到的 $\langle v_i h_j\rangle^k$ 近似代表 $\langle v_i h_j\rangle_{\mathrm{model}}$ 项的值（通常 $k=1$）。则式（8）可以转化为：

$$\nabla_{w_{ij}} J_{\mathrm{NLL}}(\boldsymbol{W},\boldsymbol{a},\boldsymbol{b};\boldsymbol{v}) = -\left[\langle v_i h_j\rangle^0 - \langle v_i h_j\rangle^k\right] \tag{9}$$

其中，$\langle\cdot\rangle^I$ 代表 CD 算法迭代 I 次的平均值。在

一个 RBM 训练好后，可以将另一个 RBM 堆叠在其上层，并将第一个 RBM 的隐藏层向量作为下一个 RBM 的可见层单元继续训练。底层的 RBM 为高斯—伯努利型，由于其输出为二进制数，则上层的 RBM 均为伯努利—伯努利型。DBN 由多个 RBM 堆叠而成，堆叠的多个 RBM 可以看作不同层的非线性特征提取器。随着层数的递增，提取的特征更加抽象，更能代表高维数据中愈渐复杂的统计结构。因此，用 DBN 进行数据降维的效果要远远优于其他传统方法。

3.2 基于单分类支持向量机的异常检测

对于经 DBN 降维后的数据 $\boldsymbol{X}_{m\times d}$，将其输入 OCSVM 进行异常检测。在 OCSVM 中，采用 SVDD 算法对降维后的数据进行异常检测。对于给定的输入数据，SVDD 可以找到一个紧致的超球面来尽可能地包含绝大部分数据点。在超球面以外的数据点，则被认定为异常数据。超球面的中心表示为 a，R 为其半径。则进行异常检测便转化为优化以下二次规划问题：

$$\begin{aligned}&\min_{a,R,\zeta}\ R^2+\frac{1}{mv}\sum_{l}^{m}\zeta_l\\&\text{s.t. }\left\|\phi\left(\boldsymbol{y}_l\right)-a\right\|^2\leqslant R^2+\zeta_l\\&\forall\, l=1,\cdots,m,\ \zeta_l\geqslant 0\end{aligned}\tag{10}$$

其中，$\phi(\cdot)$ 是将数据映射到较高维空间的非线性函数。v 为正则化参数，用来调整超平面的大小，权衡球面内外数据点的分布，避免出现过拟合。ζ_l 和 $l=1,\cdots,m$ 是松弛变量，允许一些数据点位于超球面外部。令：

$$\boldsymbol{\alpha}=\left[\alpha_1,\cdots,\alpha_m\right]^{\mathrm{T}},\quad 0\leqslant\alpha_l\leqslant\frac{1}{mv}\tag{11}$$

则上述问题便可以转化为：

$$\begin{aligned}&\max_{\alpha}\ \sum_{l=1}^{m}\alpha_l\left(\boldsymbol{y}_l\cdot\boldsymbol{y}_l\right)-\sum_{l,t}\alpha_l\alpha_t\left(\boldsymbol{y}_l\cdot\boldsymbol{y}_t\right)\\&\forall\, 1\leqslant l,t\leqslant m\\&\text{s.t. }0\leqslant\alpha_l\leqslant\frac{1}{mv}\end{aligned}\tag{12}$$

通过求解上述问题，可以得出：$\left\|\phi\left(\boldsymbol{y}_l\right)-a\right\|^2<R^2+\zeta_l$，拉格朗日乘子 $\alpha_l=0$，数据在超球面内部，为非异常数据；$\left\|\phi\left(\boldsymbol{y}_l\right)-a\right\|^2=R^2+\zeta_l$，$0<\alpha_l<\frac{1}{mv}$，数据在超球面边界上，为非异常数据；$\left\|\phi\left(\boldsymbol{y}_l\right)-a\right\|^2>R^2+\zeta_l$，$\alpha_l=\frac{1}{mv}$，数据在超球面外部，为异常数据。

4 实验与分析

实验中将 DBN-SVDD 算法与 SVDD 算法、PCA-SVDD 算法和 AE 算法进行比较，从检测正确率和训练以及测试时间方面对比 3 种算法的性能。本实验采用的数据集来自 UCI 机器学习库，数据均采集于真实的生活。共选取 4 个高维数据集进行训练和测试，其分别为：森林覆盖集（forest covertype，FC）、基于传感器检测的气体种类集（gas senor array drift，GAS）、日常活动集（daily and sport activity，DSA）和基于智能设备穿戴的人类活动集（human activity recognition using smartphone，HAR）。其维数分别为：54、128、315 和 561 维。采用不同维度的数据集进行测试，从而更好地评估本文算法性能。

对于每个数据集，分别将其中的 70%用来训练，即作为训练集；将其余数据的 30%用作测试集。在训练集中混入 5%的异常数据，测试集中混入 20%的异常数据，异常数据的每一维由均匀分布 $U(0,1)$ 随机生成[15]。实验前将数据进行预处理，对于每一个维度的数据统一归一化至 $[0,1]$。数据的类别标签分为两种，1 代表正常数据，−1 代表异常数据。由于 3 种算法均采用无监督训练，数据的类别标签只在测试时使用。对于 PCA 算法，选取 95%的贡献率来确定降维后的数据维度[16]。对于 AE 算法，根据参考文献[17]设定重构误差的异常阈值。

在以下实验中，用 DBN 后所加的数字表示 DBN 的层数。例如：DBN1 和 DBN3 分别表示为具有 1 层和 3 层隐藏层的深度信念网络。在实验一

表 1　3 种算法在 RBF 核函数下的异常检测正确率

算法模型	数据集				平均值
	FC	GAS	DSA	HAR	
SVDD	95.45%	91.73%	82.66%	86.73%	89.14%
PCA-SVDD	96.55%	95.23%	91.48%	92.34%	93.90%
DBN-SVDD	97.92%	98.29%	96.76%	97.55%	97.63%

表 2　3 种算法在线性核函数下的异常检测正确率

算法模型	数据集				平均值
	FC	GAS	DSA	HAR	
SVDD	80.34%	83.54%	79.47%	78.62%	80.49%
PCA-SVDD	84.32%	91.45%	79.13%	79.85%	83.69%
DBN-SVDD	98.50%	96.52%	97.14%	98.45%	97.65%

和实验二中，默认的 DBN 为具有 2 层隐藏层的深度信念网络。对于 DBN 的每个隐藏层神经元个数，根据参考文献[18]的方法在最优性能下确定。

在 OCSVM 中，采用交叉验证法对数据进行训练，交叉训练重数设定为 3[19]。OCSVM 的参数选取采用“网格搜索”法，在规定的参数范围内寻找最优解。其参数范围分别为：RBF 系数 $g(2^{-15},2^{-9},\cdots,2^{3})$，惩罚参数 $C\left(2^{-5},2^{-4},\cdots,2^{15}\right)$ [20]。

（1）实验一

将 DBN-SVDD 算法与 SVDD、PCA-SVDD 两种经典算法分别在线性（linear）核函数和径向基函数（radical basis function，RBF）下进行实验对比。通过对以上 4 个数据集进行异常检测，其识别率见表 1、表 2（识别率保留百分号前小数点后两位），并将表 1、表 2 的数据绘制成图 3、图 4 的折线。

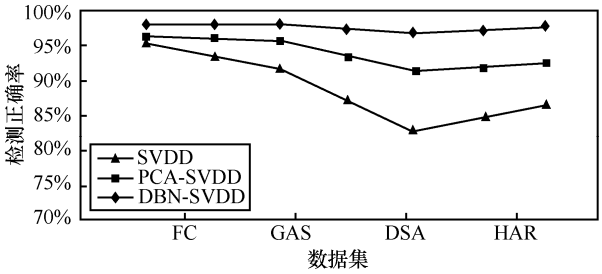


图 3　RBF 函数下 3 种算法对 4 个数据集的异常检测正确率

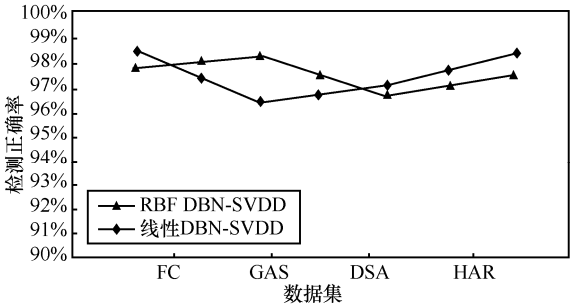


图 4　DBN-SVDD 算法下两种核函数的异常检测正确率对比

通过观察表 1、表 2 中的数据以及图 3、图 4，可以得出以下结论。

- 通过比较 4 个数据集异常检测正确率的平均值，可以得出 3 种模型算法的对比结果为：DBN-SVDD > PCA-SVDD > SVDD，其中，线性 DBN-SVDD 的检测正确率最高，为 97.65%，相比于 RBF-PCA-SVDD 和 RBF-SVDD 的 93.90%和 89.14%的检测结果，分别有 4.65%和 8.51%的提升。
- 对于 PCA 方法降维，当使用线性核函数时，对于低维数据集如 FC、GAS，异常检测的正确率有一定提升；当数据维度较高时，如 DSA、HAR，利用 PCA 降维相比于 SVDD 算法其测试结果几乎没有提升。

- 对于 SVDD 和 PCA-SVDD 这两种算法，无论使用线性核函数或者径向基函数，随着数据维度的增加，其异常检测的正确率逐渐下降。而使用 DBN-SVDD 算法其异常检测结果基本不受数据维度的影响，在各种维度的数据中，其检测结果都要优于另外两种算法。
- 对于 DBN-SVDD 算法，当使用线性核函数和径向基函数时，对实验结果基本不产生影响。这说明利用 DBN 更好地提取了高维数据中的特征，即使用线性核函数也有很好的检测结果。

（2）实验二

将 AE 算法与 DBN-SVDD 算法分别在检测正确率和检测效率上进行比较。对于 DBN-SVDD 混合模型，训练和测试的时间包括数据降维部分和降维后异常检测两部分的总和，训练和测试的时间为 SVDD 平均迭代 1 000 次的时间值。

首先将 DBN-SVDD 分别在线性和 RBF 两种核函数下的异常检测率与 AE 算法进行比较，实验结果见表 3。

由表 3 可以看出，AE 算法的平均异常检测正确率为 97.24%，与 DBN-SVDD 算法在 RBF 核下的 97.63%以及线性核下的 97.65%几乎没有差别。说明 AE 算法通过对比数据间的重构误差，在异常检测正确率上也可以达到很好的效果。再将两种算法的训练和测试时间进行对比，实验结果分别见表 4 和表 5。

由表 4 可以看出，DBN-SVDD 算法下两种核函数分别进行训练的时间基本一致，这也进一步表明 DBN 对高维数据进行特征提取的优良特性。对于 AE 算法，其训练时间平均为 0.772 1 s，分别为线性核 DBN-SVDD 的 5.5 倍和 RBF 核的 4.4 倍，进一步说明了 DBN-SVDD 算法的高效性。

由表 5 可以看出，AE 算法的测试时间平均时间为 3.993 0 ms，均大于线性核 DBN-SVDD 和 RBF 核 SVDD 算法。与 AE 算法相比，线性核函数的测试平均时间为 0.281 3 ms，时间缩短了近 13.2 倍；RBF 核函数的测试平均时间为 0.473 1 ms，时间缩短了近 7.4 倍。对于 DBN-SVDD 算法，其采用线性核函数所测试的时间小于采用 RBF 核函数进行测试的时间。这是由于 RBF 核函数具有更高的计算复杂度，因此需要花费更多的时间。由于采用线性核函数和 RBF 核函数，异常检测正确率几乎一致，而采用线性核函数进行测试的平均时间为 0.281 3 ms，相比于采用核函数的 0.473 1 ms，时间降低了 40.54%。因此，采用线性核函数在很大程度上缩短了进行数据测试的时间，提高异常检测效率。

表 3　DBN-SVDD 与 AE 算法的异常检测正确率对比

核函数	算法模型	数据集				平均值
		FC	GAS	DSA	HAR	
	AE	97.83%	97.54%	96.32%	97.25%	97.24%
RBF	DBN-SVDD	97.92%	98.29%	96.76%	97.55%	97.63%
线性	DBN-SVDD	98.50%	96.52%	97.14%	98.45%	97.65%

表 4　DBN-SVDD 与 AE 算法的训练时间对比（单位：s）

核函数	算法模型	数据集				平均值
		FC	GAS	DSA	HAR	
线性	DBN-SVDD	0.076 8	0.096 9	0.149 5	0.241 6	0.141 2
RBF	DBN-SVDD	0.090 9	0.107 5	0.188 1	0.318 6	0.176 3
	AE	0.436 8	0.987 4	0.632 5	1.031 3	0.772 1

表 5　DBN-SVDD 与 AE 算法的测试时间对比（单位：ms）

核函数	算法模型	数据集				平均值
		FC	GAS	DSA	HAR	
线性	DBN-SVDD	0.121 3	0.183 1	0.258 1	0.562 8	0.281 3
RBF	DBN-SVDD	0.182 4	0.359 7	0.473 8	0.876 7	0.473 1
	AE	1.537 2	2.243 6	5.569 0	6.622 2	3.993 0

表 6　线性核函数下不同 DBN 隐藏层数对实验结果的影响

算法模型	数据集				平均值
	FC	GAS	DSA	HAR	
DBN1-SVDD	97.35%	97.02%	95.54%	96.76%	96.67%
DBN2-SVDD	98.50%	96.52%	97.14%	98.45%	97.65%
DBN3-SVDD	97.60%	96.87%	97.32%	98.26%	97.54%

（3）实验三

在确定 DBN-SVDD 混合模型为最优算法的前提下，探究 DBN 隐藏层的层数对实验结果的影响。由于过多的层数会增加模型的复杂性和算法计算量，因此只讨论最多 3 层隐藏层对实验结果的影响。在实验 1 中，进行了具有 2 层隐藏层的 DBN 测试。接下再分别对 DBN1 和 DBN3 在线性核函数下进行实验测试，实验结果见表 6。

将表 5 中的实验结果绘制成图 5 后可以看出，具有 1 层隐藏层的 DBN1 属于“浅层模型”，导致其最终实验测试结果除了在 GAS 数据集为 97.02%，略高于其他两种算法外，在其余数据集的测试结果均低于另外两种“深层模型”。对于 DBN3，其实验结果与 DBN2 相比除了在 FC 数据集上有较大波动外（检测率降低了 0.90%），在其他数据集上的检测结果相差甚微，只在 0.18%~0.35%范围波动，基本相同。而对于 DBN3 而言，其网络模型的复杂度以及计算量均高于 DBN2。因此，确定具有 2 层隐藏层的 DBN2 为最佳网络模型。

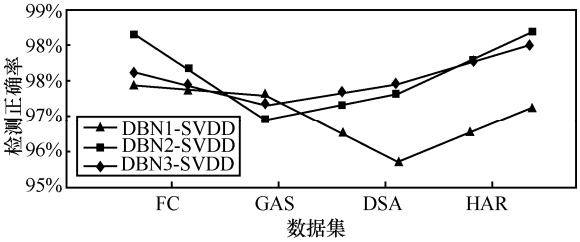


图 5　不同 DBN 隐藏层数下的异常检测正确率

5　结束语

本文通过将深度信念网络和单分类支持向量机组合到一起，提出 DBN-SVDD 算法模型。通过数据降维的方式，该算法很好地解决了高维数据的异常检测问题。利用 DBN 的非线性特性以及逐层递进的特征提取方式来获得高维数据中的低维特征，良好地解决了“维数灾难”问题。通过实验，确定了 DBN2 为最佳的降维网络模型。采用线性核的 DBN-SVDD 算法在测试时间上相比 RBF 核可以降低 34.9%。对比 PCA-SVDD 算法，其检测正确率最高提升了 4.65%；对比 AE 算法，其测试时间缩短到 1/13。

参考文献：

[1] 王忠伟, 陈叶芳, 肖四友, 等. 一种高维大数据全 k 近邻查询算法[J]. 电信科学, 2015, 31(7): 52-62.
WANG Z W, CHEN Y F, XIAO S Y, et al. An AkNN algorithm for high-dimensional big data[J]. Telecommunications Science, 2015, 31(7): 52-62.

[2] CHANDOLA V, BANERJEE A, KUMAR V. Anomaly detection:A survey[J]. ACM Computing Surveys, 2009, 41(3): 1-58.

[3] SHIN H J, EOM D H, KIM S S. One-class support vector machines—an application in machine fault detection and classification[J]. Computers & Industrial Engineering, 2005, 48(2): 395-408.

[4] 李昕, 钱旭, 王自强. 一种高效的高维异常数据挖掘算法[J]. 计算机工程, 2010, 36(21): 34-36.
LI X, QIAN X, WANG Z Q. Efficient data mining algorithm for high-dimensional outlier data[J]. Computer Engineering, 2010, 36(21): 34-36.

[5] TENENBAUM J B, DE S V, LANGFORD J C. A global geometric framework for nonlinear dimensionality reduction[J]. Science, 2000, 290(5500): 2319.

[6] POMERANTSEV A L. Principal component analysis(PCA)[M]. New York: John Wiley & Sons, Inc., 2014: 4229-4233.

[7] ROWEIS S T, SAUL L K. Nonlinear dimensionality reduction by locally linear embedding[J]. Science, 2000, 290(5500): 2323.

[8] GONZALEZ I, DÉJEAN S, MARTIN P G P, et al. CCA: an R package to extend canonical correlation analysis[J]. Journal of Statistical Software, 2008, 23(12).

[9] CHENOURI S, LIANG J, SMALL C G. Robust dimension reduction[J]. Wiley Interdisciplinary Reviews Computational Statistics, 2015, 7(1): 63-69.

[10] 程辉, 方景龙, 王大全, 等. 超平面支持向量机简化性能分析[J]. 电信科学, 2015, 31(8): 78-83.
CHENG H, FANG J L, WANG D Q, et al. Performance analysis of simplification of hyperplane support vector machine[J]. Telecommunications Science, 2015, 31(8): 78-83.

[11] GEORGE A. Anomaly detection based on machine learning dimensionality reduction using PCA and classification using SVM[J]. International Journal of Computer Applications, 2012, 47(21): 5-8.

[12] BAO S, ZHANG L, YANG G. Trajectory outlier detection method based on kernel principal component analysis[J]. Journal of Computer Applications, 2014, 34(7): 2107-2110.

[13] SAKURADA M, YAIRI T. Anomaly detection using autoencoders with nonlinear dimensionality reduction[C]//Mlsda Workshop on Machine Learning for Sensory Data Analysis, December 2, 2014, Gold Coast, Australia QLD, Australia. New York: ACM Press, 2014: 4-11.

[14] HINTON G E. Training products of experts by minimizing contrastive divergence[J]. Neural Computation, 2002, 14(8): 1771-1800.

[15] SUBRAMANIAM S, PALPANAS T, PAPADOPOULOS D, et al. Online outlier detection in sensor data using non-parametric models[C]//International Conference on Very Large Data Bases, September 12-15, 2006, Seoul, Korea. New York: ACM Press, 2006: 187-198.

[16] MOORE B. Principal component analysis in linear systems: controllability, observability, and model reduction[J]. IEEE Transactions on Automatic Control, 2003, 26(1): 17-32.

[17] HU C, HOU X, LU Y. Improving the architecture of an autoencoder for dimension reduction[C]//Ubiquitous Intelligence and Computing, 2014 IEEE, Intl Conf on and IEEE, Intl Conf on and Autonomic and Trusted Computing, and IEEE, Intl Conf on Scalable Computing and Communications and ITS Associated Workshops, Dec 9-12, 2014, Bali, Indonesia. Piscataway: IEEE Press, 2014: 855-858.

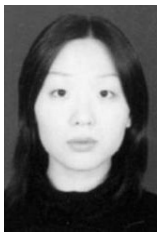
[18] HINTON G E. A practical guide to training restricted Boltzmann machines[M]. Berlin: Springer Berlin Heidelberg, 2012: 599-619.

[19] YANG J, DENG T, SUI R. An adaptive weighted one-class svm for robust outlier detection[M]. Berlin: Springer Berlin Heidelberg, 2016.

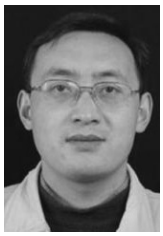
[20] LIN C J. A practical guide to support vector classification[EB/OL]. (2003-01-31)[2017-06-21]. http://www.researchgate.net/publication/200085999_A_Practical_Guide_to_Support_Vector_Classication.

[作者简介]

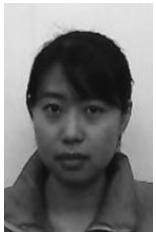
李昊奇（1992−），男，杭州电子科技大学硕士生，主要研究方向为深度学习与数据挖掘。



应娜（1978−），女，博士，杭州电子科技大学副教授、硕士生导师，主要研究方向为信号处理与人工智能。



郭春生（1971−），男，博士，杭州电子科技大学副教授、硕士生导师，主要研究方向为模式识别与人工智能。



王金华（1992−），女，杭州电子科技大学硕士生，主要研究方向为深度学习与自然语言处理。