基于有监督判别投影的网络安全数据降维算法

郭方方，吕宏武，任威霖，王瑞妮
（哈尔滨工程大学计算机科学与技术学院，黑龙江 哈尔滨 150001）

摘 要：针对传统流形学习在数据降维时不考虑原数据类别和聚类程度低的缺陷，提出了一种有监督判别投影（SDP）的流形学习降维算法来改善网络安全数据降维效果。在近邻矩阵基础上，利用数据集的类别标签信息，构建有监督判别矩阵，变无监督流形学习为有监督学习，寻找一个同时具有最大全局散度矩阵和最小局部散度矩阵的低维投影子空间，保证了降维投影后同类数据聚集而异类数据分散的特性。实验结果显示，与传统降维算法相比，所提算法可以较低的时间复杂度去除冗余数据，并且降维后的数据聚类效果更好，异类样本更分散，适用于实际的网络安全数据分析模型。

关键词：数据降维；流形学习；有监督学习；判别投影
中图分类号：TP309.2
文献标识码：A
DOI: 10.11959/j.issn.1000−436x.2021117

Reduction algorithm based on supervised discriminant projection for network security data

GUO Fangfang, LYU Hongwu, REN Weilin, WANG Ruini
College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China

Abstract: In response to the problem that for dimensionality reduction, traditional manifold learning algorithm did not consider the raw data category information, and the degree of clustering was generally at a low level, a manifold learning dimensionality reduction algorithm with supervised discriminant projection (SDP) was proposed to improve the dimensionality reduction effects of network security data. On the basis of the nearest neighbor matrix, the label information of the raw data category was exploited to construct a supervised discriminant matrix in order to translate unsupervised popular learning into supervised learning. The target was to find a low dimensional projective space with both maximum global divergence matrix and minimum local divergence matrix, ensuring that the same kind of data was concentrated and heterogeneous data was scattered after dimensionality reduction projection. The experimental results show that the SDP algorithm, compared with the traditional dimensionality reduction algorithms, can effectively remove redundant data with low time complexity. Meanwhile the data after dimensionality reduction is more concentrated, and the heterogeneous samples are more dispersed, suitable for the actual network security data analysis model.

Keywords: data dimension reduction, manifold learning, supervised learning, discriminant projection

1 引言

网络空间安全已成为互联网发展的核心挑战，从系统漏洞、隐私泄露到网络诈骗，各种安全威胁日益增多，网络安全分析中所需要收集和统计的网络安全数据量正在以指数级增长，所以优化分析处理网络安全数据的效率对于提高网络安全与服务质量有着非常重大的意义。然而，网络安全数据的

收稿日期：2020−12−03；**修回日期**：2021−03−01
通信作者：任威霖，ren.weilin@foxmail.com
基金项目：国家自然科学基金资助项目（No.61872104）；中央高校基础科研业务费专项资金资助项目（No.3072020CF0603）
Foundation Items: The National Natural Science Foundation of China (No.61872104), The Fundamental Research Fund for the Central Universities (No.3072020CF0603)

高维数据空间具备本征稀疏性，使多元密度估计问题更加复杂，难以直接对其进行求解。这一问题于1957 年在 Bellman 的著作序言中被提出，称作“维度灾难”。该问题导致在分析原始的高维网络安全数据时，会产生巨大的计算量，严重影响研究效率。为了更好地理解和处理这些高维复杂的网络安全数据，人们开始关注如何有效地降低数据的维度从而提高数据分析模型的性能。数据降维技术通过分析网络安全数据不同维度之间的内在联系，在高维空间中发掘出其隐藏的低维映射，且能够在一定程度上等效替代原有的高维结构，从而降低网络安全分析的时间复杂度[1]。为了提高网络安全分析能力，十分有必要对网络安全数据进行降维处理[2]。

近年来，流形学习方法的研究方兴未艾，国内外不断涌现出新的研究成果[3]。流形学习方法在特征空间内建立的映射能够将数据从高维度投影至低维度，有效去除了冗余信息，让人们能够更加直观、清晰地理解数据的含义。在网络安全分析领域，使用流形学习方法能够有效降低网络数据特征杂乱、冗余信息过多对模型带来的负面影响，使模型的性能得以突破[4]。传统流形学习方法为了保留数据的几何信息，便于观察，大多采用无监督的方式。这虽然增强了数据的可视性，但没有考虑原始数据的类别信息，会使降维后数据的聚类效果偏低，分类不明显，流量分析准确率变低，潜藏网络安全漏洞。近年来，随着网络面临的安全威胁日益增加，越来越多的学者开始将目光转向聚类效果更强的有监督流形学习方法。网络安全数据的有监督学习方法具备一定的理论基础，目前国内外的网络安全数据分析技术已较成熟，如美国麻省理工学院林肯实验室的 DARPA 98、DARPA 99、DARPA 2000 数据分析项目，加利福尼亚大学网络安全实验室[5]、斯坦福大学计算机安全实验室[6]等团队提出的较全面的网络安全数据集和网络安全数据分析方法，对于网络数据的绝大部分安全特征均有所研究。因此，即使对于未知的网络安全数据，也能够通过对现有数据特征的掌握和一定的数据分析技术手段，初步获取必要的类别信息，从而实现有监督学习。因此，本文针对网络安全数据分析中尚存的问题，提出了一种有监督判别投影（SDP, supervised discriminant projection）降维算法，在局部保留投影（LPP, locality preserving projection）等传统方法的基础上，根据高维数据的欧氏距离建立有监督判别矩阵，并根据矩阵对局部近邻图赋值，建立有监督全局散度矩阵和局部散度矩阵来寻找最佳投影子空间，挖掘高维数据的几何结构信息来对数据进行降维。实验结果表明，与原有算法相比，经该算法降维后的数据聚类程度和算法效率均有所提高。

2　相关工作

近年来，数据降维技术的研究已取得很大进展。这些研究主要分为线性降维方法和非线性降维方法，其主要区别在于分别适用于不同结构类型的数据。本节将对二者分别说明，并详细介绍非线性降维方法中的流形学习方法。

在数据降维技术发展早期，主流的研究方向是全局线性数据的降维方法，如主成分分析（PCA, principal component analysis）、线性判别分析（LDA, linear discriminant analysis）以及多维尺度分析（MDS, multiple dimensional scaling）等。文献[7]提出了基于 PCA 的分布式并行数据降维算法。作为最具代表性的线性算法之一，PCA 算法不需要先验知识，而是寻找一个高维特征空间和低维特征空间之间的特殊映射，因此在降维后保持了原始数据的样本模式。文献[8]使用 LDA 方法，通过寻找一个同时拥有最小局部散度和最大全局散度的降维投影来实现数据降维。这些算法都通过线性转换矩阵建立了高维数据和低维数据之间的联系。文献[9]提出了一种 MDS 方法，将 PCA 与局部保留投影相结合，不再同等处理所有的数据点，而是保留了关键数据点的局部邻域结构和全局方差。这一类方法虽然并没有使用线性转换矩阵，但其本质仍是线性的，也均广泛应用在诸多领域。

线性降维方法固然有其局限性，但在线性结构的数据集上，依然能够获得不错的效果。然而，近年来互联网的发展使数据规模呈指数级增加，复杂度也日渐提高，很多数据并不符合线性的分布规律，对于这些数据，线性降维方法的实际效果十分有限。为了弥补这方面的不足，研究者将目光转向了非线性降维方法，其中具有代表性的一类是基于循环迭代求解的方法。这类方法大多借助了人工神经网络（ANN, artificial neural network）的思想，如典型的自组织映射（SOM, self-organizing map）方法。SOM 具有理想的拓扑保存特性，保留了输入空

间神经元间的距离，被广泛应用于多元数据的投影、密度近似等问题的研究中。文献[10]利用现代计算机硬件优势引入了高分辨率 SOM 的概念，并证明了其作为集成学习模型的预处理器，在网络垃圾邮件、网络入侵和恶意软件检测等领域的适用性。另一种典型的循环迭代降维方法是主曲线（PC, principal curve）方法，文献[11]对主曲线方法的理论基础以及发展脉络进行了详细的介绍。基于循环迭代的方法能够在一定程度上弥补线性降维方法的不足，但仍存在一些问题：1) 在迭代求解过程中容易陷入局部最优解；2) 迭代会造成误差积累；3) 在处理大型样本集时计算代价过于高昂。

另一类常见的非线性降维方法是基于特征值或广义特征值的方法，其计算方式与基于循环迭代的方法完全不同，主要包括核变换方法和流形学习方法。核变换方法构建一个核空间，通过在空间中寻找源数据的一个线性可分的投影来实现非线性数据的降维。文献[12]提出了一种分布式环境下进行核主成分分析（KPCA, kernel PCA）的高效通信算法，结合子空间嵌入和自适应采样技术，能够根据任意配置的分布式数据集计算出一组全局核主成分，并保证其相对误差与特征空间维数和数据点数目无关。文献[13]提出了一种基于自适应局部核 Fisher 判别分析（KFDA, kernel Fisher discriminant analysis）的欺骗干扰识别方法，能够应用核技巧来减少非线性维数状态，当信噪比大于 4 dB 时，该方法在距离门拖引（RGPO, range gate pull off）欺骗干扰算法下的识别精度大于 90%。然而这类算法也有不足之处，核函数的引入使这类方法的计算通常较复杂，可能升高数据的维度；另外，方法的参数调优没有统一的标准，依赖专家的先验知识，普适性较差。

流形学习方法由于能够探索低维流形的内在结构，并根据拓扑学等原理分析其本征维度，因此常被用于处理在高维空间中内嵌的非线性低维流形数据。不过流形降维技术对于高维几何数学原理具有天然的高依赖性，这导致其模型建构通常十分复杂，使用成本较高。为解决这一困境，Tenenbaum 和 Roweis 对流形学习方法进行了长久深入的研究，最终提出了两大经典流形学习算法：局部线性嵌入（LLE, locally linear embedding）[14]和等度规映射（ISOMAP, isomatric mapping）[15]。之后，出现了越来越多的流形学习算法。文献[16]在拉普拉斯特征映射（LE, Laplacian eigenmap）和 John- Lindenstrauss 引理的基础上，提出了一种稀疏低秩近似等距线性嵌入方法，用于对高光谱图像进行降维和特征提取。文献[17]提出了一种基于局部切空间排列（LTSA, local tangent space alignment）的微阵列数据降维方法，证明了流形学习方法在医疗领域微阵列数据分析上的有效性。文献[18]提出了一个统一的图像复原−流形近似变换框架，在训练过程中流形学习方法会导致沿着低维数据流形的域变换稀疏的表示，极大地提升了抗噪性并减少了处理痕迹。为解决最大差异展开（MVU, maximum variance unfolding）和最小体积嵌入（MVE, minimum volume embedding）等理论模型产生的流形结构质量无法保证的问题，文献[19]提出了一种欧氏距离矩阵的凸优化模型，并证明了当均匀样本大小的排序使低秩矩阵的自由度达到对数因子时，该模型能够产生高精度的矩阵估计值。与线性降维方法相比，这些方法通过保留输入数据的局部结构来提供更强大的非线性降维性能，为探索非线性分布数据的内在拓扑结构提供了更优的路径。

原始的流形学习方法绝大多数都是无监督学习过程，这导致降维后数据的聚类程度偏低，不利于后续的数据处理。而有监督学习则从已知的类别信息出发，更注重降维后数据的分类效果。因此，近年来监督和半监督流形学习方法受到了越来越多的重视，也出现了一些新的方法，其中具有代表性的是对 LPP 方法进行监督学习的改进算法——局部判别投影（LDP, locality discriminant projection）算法[20]。文献[21]提出了有监督流形学习分类器，对于满足条件的有监督嵌入数据，其分类误差随着训练样本集的扩大而呈指数级衰减，证明了以保持数据低维几何结构为目标的有监督非线性嵌入数据的可分性。文献[22]提出了基于图嵌入概率半监督判别分析维数化简的早期故障辨识方法，在利用局部几何结构搜索分类的最优映射子空间的同时，半监督的训练方式还能使其充分利用原始数据的类别信息作为参考，因此即使在规模较小、数据量不充分的情况下依然能够发挥一定的作用。上述成果都对流形学习方法的有监督改良起到了重要的推进作用，但从领域整体发展进程来看，对于有监督流形学习方法的研究仍处于起步阶段，依然存在聚类效果不足、效率过低的缺陷。而在网络安全数据分析领域，由于数据集规模大、维度高、

样本稀疏的特点，尤其看重降维后数据的聚类效果，因此目前的算法无法较好地满足需求。为解决上述问题，本文对有监督流形学习降维算法进行了更深入的研究，将有监督学习和判别投影算法相结合，提出了一种有监督的判别投影降维算法。

3　有监督判别投影的流形学习降维算法

为解决上述问题，使流形学习降维方法更加贴合网络安全数据处理需求，本节基于原始数据类别信息，对无监督判别投影方法进行改造，提出了一种适用于网络安全数据的有监督判别投影降维算法（简称为 SDP 算法）。

3.1　有监督判别矩阵的建立

大部分经典的流形学习方法，如 LE、LDP 等，在建立近邻图时权值只能设置为 0/1 或热核函数值，但是这些权值并不能较好地体现数据的分类信息。SDP 算法在建立近邻图时，结合原始数据的类别信息建立有监督判别矩阵，能够更好地体现样本数据的类别特征。

有监督判别矩阵方法的具体分析过程如下。

给定 m 个训练样本 $x_1,x_2,x_3,\cdots,x_m$，首先根据数据集上高维空间数据的样本点的局部近邻关系，建立近邻矩阵 $\boldsymbol{H}$，如式(1)所示。

$$\boldsymbol{H}=\begin{cases}0,\ i\in N_s(j)\text{且}j\in N_s(i)\\ 1,\text{其他}\end{cases} \tag{1}$$

其中，$i\in N_s(j)$且 $j\in N_s(i)$代表样本 x_i 是样本 x_j 的近邻且样本 x_j 是样本 x_i 的近邻。

对于近邻矩阵 $\boldsymbol{H}$ 的任意元素 $h_{i,j}$，当 h_{ij}=0 时，说明 x_i 与 x_j 为近邻关系；当 $h_{i,j}$=1 时，说明 x_i 与 x_j 为非近邻关系。由于任意元素为 0 或 1 时，对于数据分类而言没有判别性，因而利用数据集的类别标签信息，并结合近邻矩阵 $\boldsymbol{H}$ 的近邻关系，变流形无监督学习为有监督，并构造有监督判别矩阵 $\boldsymbol{S}$，如式(2)所示。

$$\boldsymbol{S}=\begin{cases}\exp\left(-\|x_i-x_j\|p\right),x_i\text{与}x_j\text{为同类近邻点}\\ \exp\left(-\|x_i-x_j\|p\right)\left(1-\exp\left(-\|x_i-x_j\|p\right)\right),x_i\text{与}x_j\text{为异类近邻点}\\ 0,\text{其他}\end{cases} \tag{2}$$

其中，$\|x_i-x_j\|$是两点之间的欧氏距离，p 是一个可以调节的常数。

3.2　降维算法原理

SDP 算法能够有效消除原始数据产生的冗余干扰，缩减网络安全数据的规模，使降维投影后同类的数据距离更近，表现出明显的集簇效果；异类的簇之间彼此远离，界限较清晰。这一现象能够显著降低后续数据处理工作的难度。具体降维方法如下。

1) 根据近邻点数量 K 建立局部近邻图，利用有监督判别矩阵对局部近邻图的边进行赋值从而建立近邻图，再根据近邻图构建局部散度矩阵 $\boldsymbol{S}_{\mathrm{L}}$，如式(3)所示。

$$\boldsymbol{S}_{\mathrm{L}}=\frac{1}{2}\cdot\frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}S_{i,j}\left(x_i-x_j\right)\left(x_i-x_j\right)^{\mathrm{T}}=\frac{1}{m^2}\boldsymbol{XL} \tag{3}$$

其中，$\boldsymbol{L}$ 为拉普拉斯矩阵，$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{H}$，矩阵 $\boldsymbol{D}$ 如式(4)所示。

$$\boldsymbol{D}=\begin{bmatrix}\sum_{k=1}^{i}H_{k,1} & 0 & \cdots & 0\\ 0 & \sum_{k=1}^{i}H_{k,2} & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & 0 & \sum_{k=1}^{i}H_{k,j}\end{bmatrix} \tag{4}$$

2) 构建全局散度矩阵 $\boldsymbol{S}_{\mathrm{N}}$，如式(5)所示。

$$\boldsymbol{S}_{\mathrm{N}}=\frac{\sum_{i=1}^{m}\sum_{j=1}^{m}\left(1-S_{i,j}\right)\left(x_i-x_j\right)\left(x_i-x_j\right)^{\mathrm{T}}}{2m^2} \tag{5}$$

3) 为了寻找一个变换矩阵 $\boldsymbol{A}=[\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_r]$，使经过判别向量 $\boldsymbol{a}$ 转化后的低维投影子空间能够同时具有最大全局散度矩阵 $\boldsymbol{S}_{\mathrm{N}}$ 和最小局部散度矩阵 $\boldsymbol{S}_{\mathrm{L}}$，建立一个关于 $\boldsymbol{A}$ 的函数模型 $J(\boldsymbol{A})$，如式(6)所示。

$$J(\boldsymbol{A})=\max\frac{J_{\mathrm{N}}(\boldsymbol{A})}{J_{\mathrm{L}}(\boldsymbol{A})}=\max\frac{\mathrm{tr}\left\{\boldsymbol{A}^{\mathrm{T}}\boldsymbol{S}_{\mathrm{N}}\boldsymbol{A}\right\}}{\mathrm{tr}\left\{\boldsymbol{A}^{\mathrm{T}}\boldsymbol{S}_{\mathrm{L}}\boldsymbol{A}\right\}} \tag{6}$$

在建立函数模型 $J(\boldsymbol{A})$的基础上，增加正交化约束，求解正交基向量$\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_r$，并构建约束目标函数模型。

4) 计算正交基函数。正交基为 $\boldsymbol{A}=[\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_r]$，令 $\boldsymbol{A}^{r-1}=[\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_{r-1}]$，根据广义特征方程 $\boldsymbol{XLX}^{\mathrm{T}}\boldsymbol{a}=\lambda\boldsymbol{XLX}^{\mathrm{T}}\boldsymbol{a}$，通过求解使式(7)取得最小值的向量 $\boldsymbol{a}_1$，计算得到正交矩阵 $\boldsymbol{A}$ 的一个特征向量为

$$\boldsymbol{a}_1=\arg\min\frac{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{XLX}^{\mathrm{T}}\boldsymbol{a}}{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{XDX}^{\mathrm{T}}\boldsymbol{a}} \tag{7}$$

5) 求解在约束条件下使式(8)取得最小值的向量 $\boldsymbol{a}_m$，得到第 m 个特征值对应的特征向量为

$$\begin{aligned}&\boldsymbol{a}_m=\arg\min\frac{\boldsymbol{a}_m^{\mathrm{T}}\boldsymbol{XLX}^{\mathrm{T}}}{\boldsymbol{a}_m^{\mathrm{T}}\boldsymbol{XDX}^{\mathrm{T}}}\\&\text{s.t. }\boldsymbol{a}_m^{\mathrm{T}}\boldsymbol{a}_1=\boldsymbol{a}_m^{\mathrm{T}}\boldsymbol{a}_2=\cdots=\boldsymbol{a}_m^{\mathrm{T}}\boldsymbol{a}_{m-1}\\&\quad\boldsymbol{a}_m^{\mathrm{T}}\boldsymbol{XDX}^{\mathrm{T}}\boldsymbol{a}_m=\boldsymbol{I}\end{aligned} \tag{8}$$

其中，$\boldsymbol{I}$ 为单位矩阵。

通过求解以上方程获得正交基向量$\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_m$。

6) 在线性投影矩阵满足正交化的约束下，构建约束目标函数模型为

$$\begin{aligned}&\max\frac{\mathrm{tr}\left\{\boldsymbol{A}^{\mathrm{T}}\boldsymbol{S}_{\mathrm{N}}\boldsymbol{A}\right\}}{\mathrm{tr}\left\{\boldsymbol{A}^{\mathrm{T}}\boldsymbol{S}_{\mathrm{L}}\boldsymbol{A}\right\}}\\&\text{s.t. }\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}=\boldsymbol{I}\end{aligned} \tag{9}$$

根据以上步骤构建约束目标函数模型 $J(\boldsymbol{A})$，利用特征分解获得约束目标函数的解，并输出高维数据在低维空间的投影。

以上模型的构建方式与传统流形算法 LPP 以及 UDP 对于降维过程中邻接矩阵权值的处理方式不同，但其模型数学原理基本一致，在实际计算过程中通常使用拉格朗日乘数法构建辅助函数以加快计算速度，因此计算复杂度上，SDP 算法与经典流形学习算法并没有明显差异。

3.3 降维算法流程

在 3.2 节提出的有监督判别投影算法中，首先根据输入样本点的近邻关系，在考虑类别信息的基础上构建有监督判别矩阵，增加条件正交化约束，并寻找一个同时具有最大全局散度矩阵和最小局部散度矩阵的低维投影子空间，经过有监督判别降维后，数据的特征维度得到缩减，且异类数据之间的界限明显清晰。SDP 算法的实现过程如算法 1 所示。

算法 1 SDP 算法

输入 高维数据 $\boldsymbol{x}=[x_1,x_2,\cdots,x_m]\in\boldsymbol{R}^{D\times n}$，类别信息 $\boldsymbol{C}=[C_1,C_2,\cdots,C_n]$

输出 线性变换 $\boldsymbol{A}\in\boldsymbol{R}^{D\times d}$ 和低维投影 $\boldsymbol{Y}=\boldsymbol{A}^{\mathrm{T}}\boldsymbol{X}\in\boldsymbol{R}^{D\times d}$

步骤 1 建立近邻图。

步骤 1.1 根据近邻点数量 k，建立局部近邻图。

步骤 1.2 结合局部近邻图的近邻关系，利用有监督判别矩阵 $\boldsymbol{S}$ 计算 x_i 与 x_j 间的权值，并使用权值对近邻图的边进行赋值。

步骤 2 特征分解。

步骤 2.1 根据近邻图，计算局部散度矩阵 $\boldsymbol{S}_{\mathrm{L}}$。

步骤 2.2 根据近邻图，求得全局散度矩阵 $\boldsymbol{S}_{\mathrm{N}}$。

步骤 2.3 根据所计算的局部散度矩阵和全局散度矩阵，增加正交化约束，构建约束目标函数模型。

步骤 2.4 利用特征分解求得约束目标函数的解。

步骤 3 低维投影。输出高维数据在低维空间的投影 $\boldsymbol{Y}_{\mathrm{t}}=\boldsymbol{A}^{\mathrm{T}}\boldsymbol{X}_{\mathrm{t}}$，其中，下角标 t 表示低维空间。

4 仿真实验

4.1 实验目的及实验环境设置

降维算法的性能优劣主要体现在其降维的效果和运行算法所消耗的时间方面。研究者普遍认为，在有效降低数据维度的前提下，如果经过某种降维方法处理后的数据能够保留更多的原有信息，并且产生更明显的聚类效果，那么就可以说这种降维方法的效果是更优秀的。而时间复杂度同样是十分重要的评估标准，消耗时间过多的方法不适用于现实的网络安全实践。因此，本节将围绕这 2 个评价指标，对 SDP 算法和其他经典的数据降维算法进行对比实验，以评估 SDP 算法的有效性。

本文中的实验依托于 Hadoop 云环境，环境结构如图 1 所示。

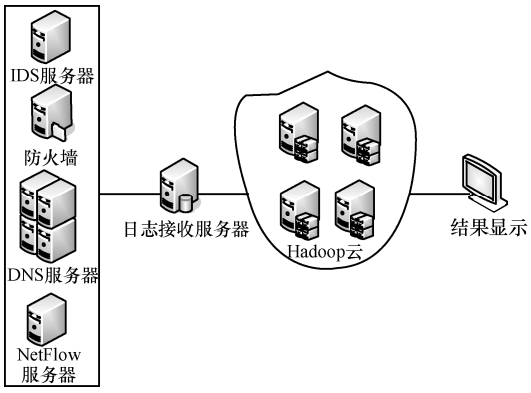


图 1 实验环境结构

实验采用 NSL KDD 异常入侵检测数据集[23]，该数据集于 2009 年由新布伦瑞克大学提出。与其前身 KDD Cup 99 数据集相比，该数据集无冗余，无重复记录，复杂度更低。NSL KDD 是关于网络事件的公共数据集，包含一组完整的被标记入侵事件，其实例和特征数量非常庞大，提供了事件分布和特性之间的依赖关系，这些特点使它更适合作为网络安全分析研究的基准。NSL KDD 的训练集包含 21 种不同的网络攻击类型，而测试集在此基础上额外添加了 17 种新的攻击类型。这些攻击大体上可以分为 4 类：拒绝服务器（DoS, denial of service）、PROBE、R2L（remote-to-login）以及 U2R（user-to-root），而非攻击类型的正常数据被标记为 Normal。实验数据集类别分布如表 2 所示。

4.2　对比实验

为了对于 SDP 算法的性能进行充分测试，本文选择了降维算法 PCA、LE、LDP 作为对照组。其中 PCA 和 LDP 分别为线性降维算法和有监督流形学习算法中最具代表性的算法之一；LE 的最终目的是使高维空间中邻近的点在低维嵌入中依然邻近，这一思想与 SDP 算法较相近，因此作为无监督流形学习算法的代表。实验将从降维效果、时间消耗和综合性能 3 个方面来分析 SDP 算法的性能。

表 1　　实验数据集类别分布

标记类型	训练集数目	测试集数目
Normal	67 343	13 449
DoS	45 927	9 234
PROBE	11 656	2 289
R2L	995	209
U2R	52	11

1) 降维效果分析

分别使用 PCA、LE、LDP 和 SDP 算法对 NSL KDD 数据集进行降维，降维后的数据可视化投影如图 2 所示。由图 2 可以看出，通过 PCA 降维后的数据，不同类之间混杂在一起，结构较混乱，这是

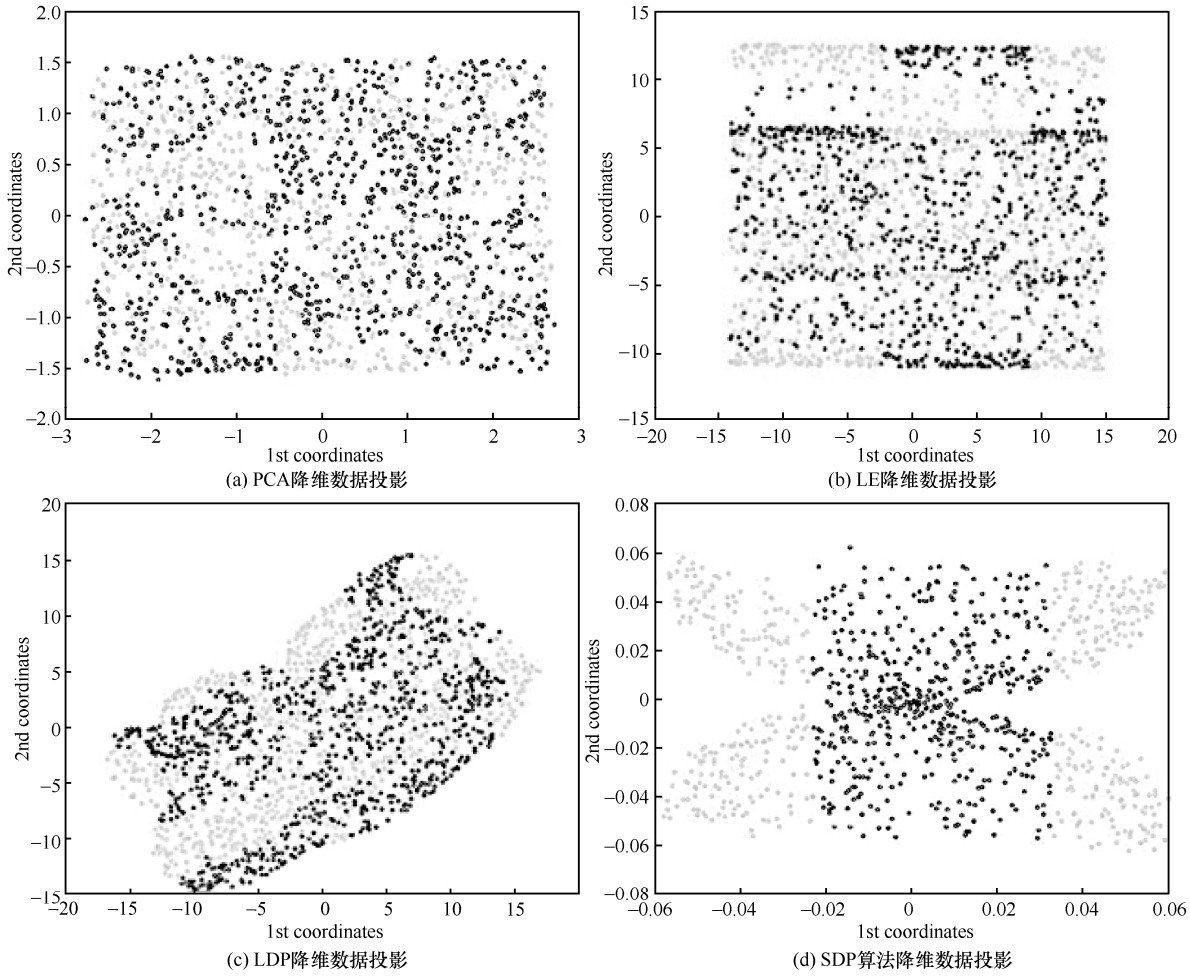


图 2　不同算法降维数据投影

由于线性降维算法自身的缺陷会导致处理后的数据维度丢失，拓扑结构遭到破坏。LE 降维后的数据结构较分明，但大量边缘数据混淆，部分区域数据十分密集。LDP 降维后的数据整体结构清晰，不同类别区分更加明显，不过数据分布仍显分散，聚类程度较低。经 SDP 算法降维后的数据，不同类别之间轮廓清晰，视觉效果上明显优于另外 3 种算法。这是由于 LDP 虽然与 SDP 算法同为有监督降维算法，但此类算法在构建近邻图时仅使用了热核函数等手段作为聚类的权值，这种方法对于样本类间距的描述能力不足。SDP 算法则构建了完整的样本距离判别矩阵，因此降维后的类别间距更加精准、清晰。

为了更具体化地证明所观察到的结论，本文引入“轮廓系数”的概念对 4 种算法的降维效果进行评估。轮廓系数是聚类效果好坏的一种评价方式，由 Rousseeuw 于 1986 年提出。对于已经处理过的数据，其轮廓系数可以表示为

$$S(x_i)=\frac{b(x_i)-a(x_i)}{\max\{a(x_i),b(x_i)\}} \tag{10}$$

其中，$a(x_i)$为样本点 x_i 到所有它属于的簇中其他点的平均距离；$b(x_i)$为样本点 x_i 到与它相距最近的一个异类簇内的所有点的平均距离，具体到本文的二分类聚类问题，则是样本点 x_i 到数据集中所有与其异类的样本点的平均距离；数据集整体的轮廓系数 S 为所有样本轮廓系数的均值，即

$$S=\frac{1}{m}\sum_{i=1}^{m}S(x_i) \tag{11}$$

可以看出，轮廓系数 S 的值为[−1, 1]，越接近 1 则证明数据的聚类程度越高。

分别对上述 4 种算法降维后的数据计算轮廓系数，结果如图 3 所示。

由图 3 可知，计算得出的轮廓系数基本和上文对于数据的视觉观测保持一致，其中线性算法 PCA 的效果最差，仅为 0.007 7，这表明经其降维后的数据基本丢失了原有的类别信息。LE 和 LDP 虽同为流形学习算法，但由于 LE 为无监督算法，LDP 为有监督算法，因此轮廓系数相差较大，LE 的轮廓系数仅为 0.016 3，而 LDP 的轮廓系数却达到了 0.093 2。改良后的 SDP 算法在降维后的数据类别信息完整度方面不仅远超过 PCA 和 LE，和同为有监督算法的 LDP 相比也表现出了一定的优势，其轮廓系数达到了 0.140 8，证明了 SDP 在降维后数据聚类效果的优势。

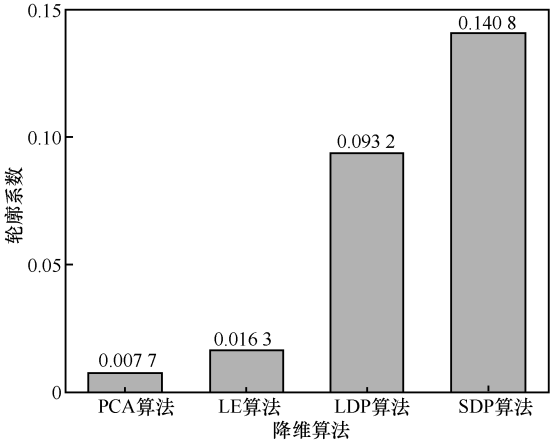


图 3 4 种算法降维数据的轮廓系数

为了测试 SDP 算法在网络安全分析领域的适用性，本文针对 NSL KDD 中 4 种不同的攻击方式：DoS、PROBE、U2R 和 R2L，分别在这些训练集上使用 SDP 算法对其降维，实验结果如图 4 所示。

从图 4 可以看出，对于 4 种攻击方式数据，SDP 算法的降维效果都较理想。降维后的数据基本保留了原本的类别属性，正常流量数据和异常流量数据在视觉效果上有着显著的区分，且异常数据的聚类效果明显。由此可见，SDP 算法在网络安全分析领域具有较强的适用性。

2) 时间消耗分析

SDP 算法为了强化降维的效果，使用了有监督的学习方式，上文的实验数据表明这一改动是成功的。但在实际的网络安全分析实践中，算法的效率也同样重要，如果这项改动带来了不可接受的时间消耗，那么也无法称之为成功的降维算法。因此，本节对 SDP 算法的时间消耗进行对比实验，对比算法仍然选择 PCA、LE 和 LDP。为了保证数据的准确性，降维时间测试设置了 7 组不同数据规模的对照组，其样本数分别为 300、600、1 200、2 400、4 800、9 600、19 200 测试，以验证 SDP 算法在不同规模的安全数据集下的时间消耗量。实验结果如表 2 和图 5 所示。

通过分析数据可以得知，与线性算法相比，流形学习算法消耗的时间明显更多，这是由于流形学习算法为非线性算法，需要寻找高维空间的局部结构，并利用 K 近邻运算进行判断，每一步都会显著增加算法的时间复杂度，但这也让流形学习算法能

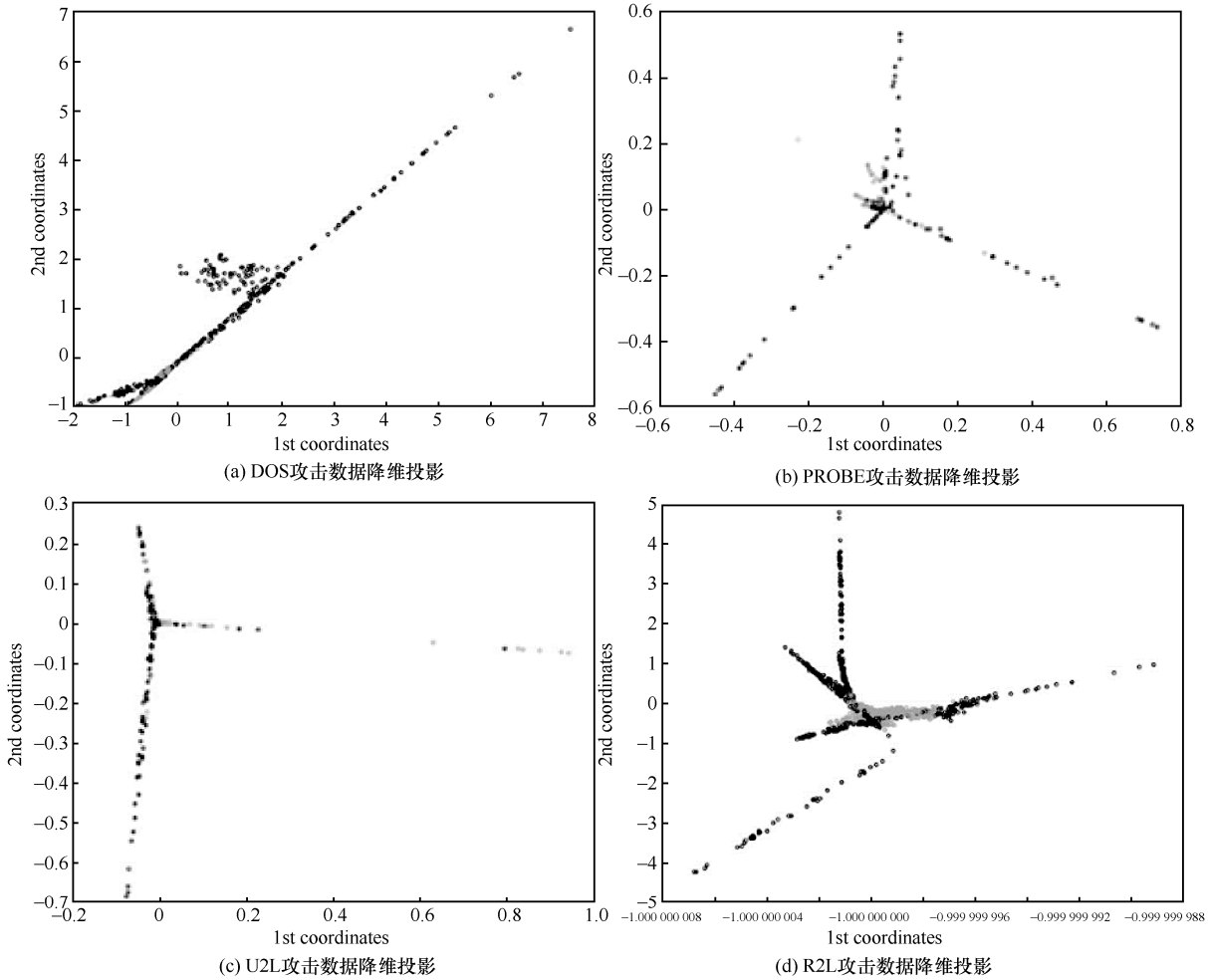


图 4　不同攻击数据集的数据降维投影

提供线性算法无法比拟的降维效果。在 3 种流形学习算法中，LDP 虽然在降维效果上优于 LE，但消耗时间却是 LE 的 5~10 倍。SDP 在降维效果上明显领先于其他算法，但消耗时间与 LDP 基本持平，而且在较大规模的数据集上，消耗时间甚至少于 LDP。出现这种现象是由于 SDP 算法在定义邻接图权值时，采用的有监督判别矩阵计算方式较稳定，只需在求解降维变换函数之前计算一次即可满足后续使用；LDP 在求解过程中使用的热核函数计算方式虽然在单项复杂度上基本与 SDP 算法持平，但在算法运行过程中可能会出现变化，导致需要多次重复计算，因此计算量偏高。

这项实验表明，SDP 算法在时间消耗方面并没有超出原有流形学习算法的范畴，并且在某些特定的情况下体现了一定的优势。

3) 综合性能分析

为了综合考量上述测试的结果，本文定义了综合性能指数 P 作为评估降维算法综合性能（效费比）的标准，进一步验证 SDP 算法在降维效果和时间消耗 2 个方面的表现，即验证算法能否在可接受的时间消耗内取得性能上的优势。P 的定义为

表 2　4 种算法在不同数据规模下的时间消耗

数据规模	运行时间/s			
	线性算法	流形学习算法		
	PCA 算法	LE 算法	LDP 算法	SDP 算法
300	0.031	0.08	0.37	0.38
600	0.072	0.17	1.21	1.31
1 200	0.075	0.61	2.93	3.27
2 400	0.075	1.63	9.21	9.72
4 800	0.075	4.19	32.51	33.21
9 600	0.138	12.41	110.31	90.53
19 200	0.138	30.36	200.24	130.48

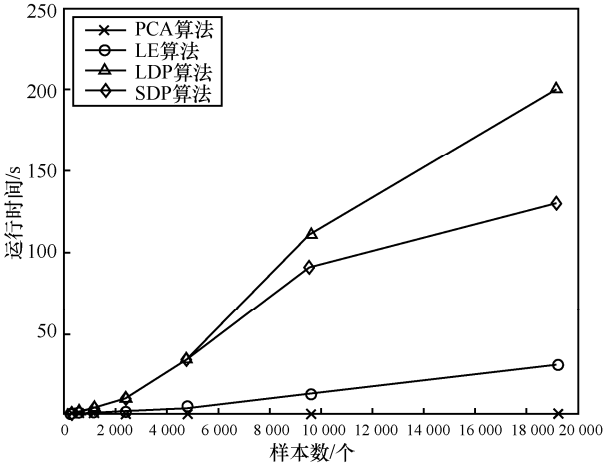


图 5 4 种算法在不同数据规模下的时间消耗曲线

$$P=\sum_{i=1}^{n}\frac{S(x_i)}{T} \tag{12}$$

其中，n 为测试数据的规模，T 为算法运行的时间消耗。结合上文中 3 种流形学习算法的轮廓系数和时间消耗数据，得到的综合性能指数如表 3 所示。

表 3 3 种流形学习算法的综合性能指数

数据规模	综合性能指数		
	LE	LDP	SDP
300	61.13	75.57	111.16
600	57.53	46.21	64.49
1 200	42.52	38.17	51.67
2 400	24	24.29	34.77
4 800	18.67	13.76	20.35
9 600	12.61	8.11	14.93
19 200	10.31	8.94	20.72

实验结果如图 6 所示。实验结果表明，与 LE 相比，SDP 算法虽然在时间开销上占据劣势，但由于在以轮廓系数为代表的降维效果上显著优于 LE，因此依然能够保持领先地位。在与 LDP 的对比中，SDP 算法不仅降维效果较优，而且在小规模数据集上的时间开销也和 LSP 基本保持一致，甚至在较大规模数据集上的时间开销小于 LSP，因此在综合性能上取得了稳定的优势。

5 结束语

本文针对网络安全数据降维领域的算法聚类效果差、效率低的问题，在传统数据降维技术的基础上，提出了一种有监督判别投影的流形学习降维算法——SDP 算法。SDP 算法利用一个有监督判别矩阵，找到同时具有最大全局散度矩阵和最小局部散度矩阵的低维投影子空间，最终实现数据的降维。实验证明，SDP 算法仅需消耗与传统流形学习算法接近的时间，但降维后数据的聚类效果显著优于线性降维算法和其他流形学习算法，且对于网络安全数据有较强的适应性，因此很适合被用于网络安全分析领域的数据降维工作中。

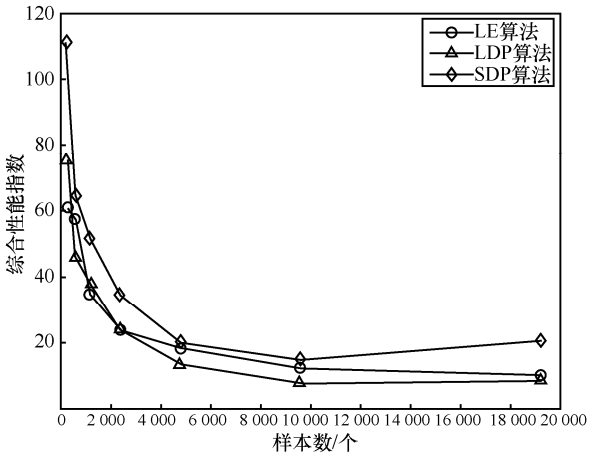


图 6 3 种流形学习算法的综合性能指数曲线

由于篇幅和时间的限制，本文仅讨论了如何在降维中保留更多的原始数据类别信息，未能深入研究如何进一步提高算法的效率，也没有涉及如何进一步提高后续的网络入侵检测精度。这些问题都有待于在未来工作中探索。

参考文献：

[1] WANG Z, PARKINSON T, LI P X, et al. The squeaky wheel: machine learning for anomaly detection in subjective thermal comfort votes[J]. Building and Environment, 2019, 151: 219-227.

[2] VIKRAM M, PAVAN R, DINESHBHAI N D, et al. Performance evaluation of dimensionality reduction techniques on high dimensional data[C]//2019 3rd International Conference on Trends in Electronics and Informatics. Piscataway: IEEE Press, 2019: 1169-1174.

[3] BYRNE J J, MORGAN J L, TWICKLER D M, et al. Utility of follow-up standard sonography for fetal anomaly detection[J]. American Journal of Obstetrics and Gynecology, 2020, 222(6): 615.e1-615.e9.

[4] NAKAZAWA T, KULKARNI D V. Anomaly detection and segmentation for wafer defect patterns using deep convolutional encoder–decoder neural network architectures in semiconductor manufacturing[J]. IEEE Transactions on Semiconductor Manufacturing, 2019, 32(2): 250-256.

[5] TERAUCHI T, AIKEN A. Secure information flow as a safety problem[M]. Berlin: Springer, 2005: 352-367.

[6] DURUMERIC Z, MA Z N, SPRINGALL D, et al. The security impact of HTTPS interception[C]//2017 Network and Distributed System Security Symposium. Virginia: the Internet Society, 2017: 1-5.

[7] WANG L L. Research on distributed parallel dimensionality reduction algorithm based on PCA algorithm[C]//2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference. Piscataway: IEEE Press, 2019: 1363-1367.

[8] GHOSH J, SHUVO S B. Improving classification model's performance using linear discriminant analysis on linear data[C]//2019 10th International Conference on Computing, Communication and Networking Technologies. Piscataway: IEEE Press, 2019: 1-5.

[9] WU D, XIONG N X, HE J R, et al. Critical data points-based unsupervised linear dimension reduction technology for science data[J]. The Journal of Supercomputing, 2016, 72(8): 2962-2976.

[10] SARASWATI A, NGUYEN V T, HAGENBUCHNER M, et al. High-resolution self-organizing maps for advanced visualization and dimension reduction[J]. Neural Networks, 2018, 105: 166-184.

[11] 张军平, 王珏. 主曲线研究综述[J]. 计算机学报, 2003, 26(2): 129-146.
ZHANG J P, WANG J. An overview of principal curves[J]. Chinese Journal of Computers, 2003, 26(2): 129-146.

[12] BALCAN M F, LIANG Y, SONG L. Communication efficient distributed kernel principal component analysis[J]. Computer Science, 2016, 27(4): 555-559.

[13] NOURI M, MIVEHCHY M, AGHDAM S A. Adaptive time-frequency kernel local fisher discriminant analysis to distinguish range deception jamming[C]//2015 6th International Conference on Computing, Communication and Networking Technologies. Piscataway: IEEE Press, 2015: 1-5.

[14] CAO Z Y, JI G L, TAN C. Improvement of algorithm multi-manifold LLE learning[J]. Computer Engineering and Applications, 2018, 54(24): 156-163.

[15] 石陆魁, 郭林林, 房子哲, 等. 基于 Spark 的并行 ISOMAP 算法[J]. 中国科学技术大学学报, 2019, 49(10): 842-850.
SHI L K, GUO L L, FANG Z Z, et al. Parallel ISOMAP algorithm based on Spark[J]. Journal of University of Science and Technology of China, 2019, 49(10): 842-850.

[16] SUN W W, YANG G, DU B, et al. A sparse and low-rank near-isometric linear embedding method for feature extraction in hyperspectral imagery classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(7): 4032-4046.

[17] CHEN X Z. LTSA algorithm for dimension reduction of microarray data[J]. Advanced Materials Research, 2013, 645: 192-195.

[18] ZHU B, LIU J Z, CAULEY S F, et al. Image reconstruction by domain-transform manifold learning[J]. Nature, 2018, 555(7697): 487-492.

[19] DING C, QI H D. Convex optimization learning of faithful Euclidean distance representations in nonlinear dimensionality reduction[J]. Mathematical Programming, 2017, 164(1/2): 341-381.

[20] NING X, LI W J, TANG B, et al. BULDP: biomimetic uncorrelated locality discriminant projection for feature extraction in face recognition[J]. IEEE Transactions on Image Processing, 2018, 27(5): 2575-2586.

[21] VURAL E, GUILLEMOT C. A study of the classification of low-dimensional data with supervised manifold learning[J]. The Journal of Machine Learning Research, 2017, 18(1): 5741-5795.

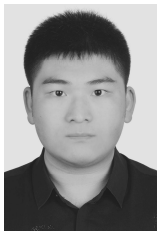
[22] 李锋, 汤宝平, 王家序, 等. 基于图嵌入概率半监督判别分析的故障辨识[J]. 机械工程学报, 2017, 53(9): 92-100.
LI F, TANG B P, WANG J X, et al. Fault identification method based on graph-implanted probability-based semi-supervised discriminant analysis[J]. Journal of Mechanical Engineering, 2017, 53(9): 92-100.

[23] GURUNG S, GHOSE M K, SUBEDI A. Deep learning approach on network intrusion detection system using NSL-KDD dataset[J]. International Journal of Computer Network and Information Security, 2019, 11(3): 8-14.

[作者简介]

郭方方（1973− ），男，黑龙江哈尔滨人，博士，哈尔滨工程大学副教授、硕士生导师，主要研究方向为计算机网络应用、新型网络体系结构、网络安全态势感知、云监控等。

吕宏武（1983− ），男，山东日照人，博士，哈尔滨工程大学副教授、博士生导师，主要研究方向为网络安全、移动云计算与移动边缘计算、形式化建模与性能评价等。



任威霖（1997− ），男，黑龙江哈尔滨人，哈尔滨工程大学硕士生，主要研究方向为网络数据预处理、网络态势感知预测等。

王瑞妮（1994− ），女，山西运城人，哈尔滨工程大学硕士生，主要研究方向为流形学习、网络数据异常处理等。