文章编号: 1001-0920(2019)05-1032-05 DOI: 10.13195/j.kzyjc.2017.1412

基于局部线性嵌入的免疫检测器优化生成算法

席 亮†, 蒋 涛, 张凤斌

(哈尔滨理工大学 计算机科学与技术学院，哈尔滨 150080)

摘 要: 网络安全已上升到国家安全战略层面,入侵检测技术是其重要的组成部分,已得到广泛关注. 在基于免疫的入侵检测研究中,针对传统实值否定选择算法不利于高效分析数据而造成的检测器生成速度慢、检测效率低等问题,引入局部线性嵌入算法,借鉴其能对高维数据进行映射降维的特点,提出一种基于局部线性嵌入的免疫检测器优化生成算法,利用局部线性嵌入对高维数据预处理优化降维,并结合实值否定选择算法生成检测器. 将该算法用于检测模型,从而提升检测器的生成速率,并可保证生成的检测器高效地处理高维数据. 该算法在降维前后可保证样本的局部线性结构不变,具有可变参数少、计算时间短的特点. 实验结果表明,所提出算法在显著提高检测器生成速率和对数据检测效率的基础上,检测性能也表现出很好的水平.

关键词: 人工免疫系统；入侵检测；局部线性嵌入算法；实值否定选择算法；检测器；降维

中图分类号: TP393.08 **文献标志码:** A

Immune detector optimized generation algorithm based on locally linear embedding

XI Liang†, JIANG Tao, ZHANG Feng-bin

(School of computer Science and Technology，Harbin University of Science and Technology, Harbin 150080，China)

Abstract: Nowadays, network security has risen to the national security strategy level. As a significant part of network security, the intrusion detection technology has aroused general concern. Based on the research of the immune mechanism intrusion detection, aiming at the problems concerning the slow generation of the detector and the low detection efficiency caused by the traditional real-valued negation selection algorithm is not conducive to the efficient analysis of the data, this paper introduces the local linear embedding algorithm which can be applied to reduce the high dimensional data preprocessing optimization dimension due to the characteristic of map the dimensionality of high-dimensional data, and combines with the real-valued negative selection algorithm to generate the detectors. Then, using this algorithm to detection model can enhance the generation velocity of the detectors and ensure the generated detector to process the high-dimensional data efficiently. The algorithm can ensure the local linear structure of the sample is the same after the dimensionality reduction, and it also has the characteristics of less variable parameters and shorter computation time. The experimental results show that this algorithm can significantly improves the detector generation velocity and the detection efficiency of the data, and it is also outstanding in the detection performance.

Keywords: artificial immune system；intrusion detection；local linear embedding algorithm；real-valued negation selection algorithm；detector；dimension reduction

0 引 言

入侵检测系统(intrusion detection system, IDS)是一种以主动防御为特色的网络安全技术,通过对网络中的各个状况信息进行实时监视以发现各种异常行为并进行必要的控制. 为了应对日益多变的攻击方式,国内外相关学者将目光集中到了基于数据挖掘、人工免疫等智能技术的IDS研究中. 基于免疫的IDS模拟生物体的各类免疫过程,通过自体耐受训练构造检测器以检测不断升级变化的安全威胁特征,特别在防御未知异常方面,表现出了很好的检测性能,成为

收稿日期: 2017-10-24；**修回日期:** 2017-12-25.

基金项目: 国家自然科学基金项目(61172168)；黑龙江省教育厅科学技术研究项目(12541130)；黑龙江省自然科学项目(F2018019)

责任编委: 魏秀琨.

作者简介: 席亮(1983−), 男, 博士, 副教授, 从事人工智能与应用、网络与信息安全等研究；张凤斌(1965−), 男, 教授, 博士生导师, 从事网络与信息安全等研究.

†通讯作者. E-mail: xiliang@hrbust.edu.cn.

本领域的一个研究热点.

为了更准确地描述各种网络行为和最大化异常识别范围,IDS 的问题空间维度和检测器规模通常过大,造成检测器的生成与优化周期长、检测效率低、难以得到准确的结果等问题. 因此,需要在保证足够信息量的前提下将问题空间进行降维优化处理. 数据优化后能保持数据间关系不变的关键是使用的映射方法,目前所用到的方法大致可分为线性降维与非线性降维两类. 前者包括主成分分析(principal component analysis, PCA)、多维尺度变换(multi-dimensional scaling transformation, MDS)等;后者包括等距离映射(equidistance mapping, ISOMAP)、局部线性嵌入(locally linear embedding, LLE)等.

在线性降维方法中,PCA 的基本思想是寻求最优投影模型[1]. MDS 可以在数据点相似的基础上保持它们之间的差异进行降维. 但这两种方法处理包含非线性结构的数据集时,难以取得理想效果,这是所有线性降维方法最大的局限性[2]. 在非线性降维方法中,ISOMAP 通过保距映射将数据集降维到低维空间中. 但当数据集存在较大噪声时,内在结构的恢复难以进行. LLE 算法通过全局线性拟合得到其内在全局线性结构[3],在高维数据的可视化和统计描述等方面非常有效.

鉴于此,本文引入 LLE 算法,将其运用到人工免疫检测器降维优化生成中,并运用于检测中以提升系统的检测效率. 基本过程为:使用 LLE 对数据集进行特征分析与降维,并结合实值否定选择算法(real-valued negative selection algorithm, RNSA),提出基于 LLE 的检测器优化生成算法,并设计检测模型,从而提升检测器的生成速率和检测效率. 实验分析表明,所提出算法能够显著提高检测器生成速率和检测效率,而且检测性能表现出很好的水平.

1　相关工作

1.1　人工免疫算法

生物免疫系统(biological immune system, BIS)与 IDS 有许多类似的地方,确保“自体”免受“非自体”的攻击. BIS 关键是抗体生成和对抗原进行检测. 同理,IDS 的核心是检测器的生成和对异常行为的检测,二者的基本过程相似. RNSA 是人工免疫理论的核心方法,基本思想为:对自体与候选检测器进行亲和力计算,若大于等于给定的阈值 r 则匹配,舍弃该候选样本,否则该候选检测器通过耐受训练,将其作为成熟检测器[4].

基于免疫的 IDS 的核心是检测器. 在进行检测时,如果亲和力计算速率高,则检测率也可维系在较高水平. 但当待检测数据处于高维空间时,难以解决检测时间代价过高和检测效率低等问题. 在保证检测率的前提下,通常需要对待检测的高维数据进行降维操作,从而提高检测效率. 因此,本文将基于 LLE 的检测器优化生成算法用于对待检测事件进行检测以提高系统的检测效率.

1.2　LLE 算法

LLE 算法在特征提取中需要使用数据的类别信息,目前主要有半监督和全监督两种方式,已取得了满意的结果[5].

在数据分类研究方面,文献[6]提出了基于 LLE 的故障分类器训练方法,通过减少提取特征的维数使样品更加分离,提升了诊断振动故障的能力. 文献[7]为了解决奇异矩阵无法求解和近邻点不合适的问题,通过引入正交匹配追踪算法(orthogonal matching pursuit, OMP)到 LLE,提出基于 OMP 的 LLE,将其中的最小二乘法算子替换成 OMP,改善了分类效果.

在降维数据准确性研究方面,主要通过优化参数和算子实现. 文献[8]提出一个新颖的流形学习技术,能明智地选择一个合适数量(即参数 k)的邻近点,最小化空间相关索引,从而提高降维性能. 文献[9]提出了一种修正的 LLE 来解决使用欧氏距离无法准确测量高维样本间距的问题,从而保留更多的原始数据信息.

在非线性数据降维研究方面,文献[10]提出了基于 LLE 改进的稀疏散度指数修正算法以解决高光谱中非线性数据的降维问题,提高了降维后数据的准确性. 文献[11]提出一种基于 LLE 最大色散矩阵的不相关统计算法并应用于消除冗余信息,将其用于对人脸库进行识别验证获得了很好的效果,表明该方法可有效提高算法的识别率.

LLE 算法的降维过程简述如下.

Step 1: 每次随机选取一个样本点,并寻找其 k 个近邻点.

Step 2: 对这 k 个近邻点采用 Euclidean 距离算法构建误差函数,求得局部协方差矩阵,从而得到局部重建权值矩阵.

Step 3: 每个样本点的输出值由 Step 1 和 Step 2 中构建的矩阵得到.

在 Step 2 中,误差函数定义如下:

$$\min\ \varepsilon(w^i)=\sum_{i-1}^{N}\left|x_i-\sum_{j-1}^{k}w_j^i x_{ij}\right|^2. \tag{1}$$

其中:$x_{ij}(j=1,2,\cdots,k)$ 为 x_i 的 k 个近邻点,w_j^i 为 x_i

与x_{ij}的权值. $\min \varepsilon(w^i)$值越小,局部重建权值矩阵的效果越好.

式(1)满足$\sum_{i-1}^{k} w_j^i = 1$,构造一个局部协方差矩阵Q^i,有

$$Q_{jm}^i = (x_i - x_{ij})^{\mathrm{T}}(x_i - x_{im}). \tag{2}$$

Q^i与$\sum_{i-1}^{k} w_j^i = 1$相结合,并采用Lagrange乘数法,求出局部优化重建权值矩阵W,其中每个元素w_j^i为

$$w_j^i = \frac{\sum_{m-1}^{k} (Q^i)_{jm}^{-1}}{\sum_{p-1}^{k}\sum_{q-1}^{k} (Q^i)_{pq}^{-1}}. \tag{3}$$

另外,Q^i可能会是一个奇异矩阵,此时需对Q^i进行正则化,有

$$Q^i = Q^i + rI. \tag{4}$$

其中:r为正则化参数,I为$k \times k$的单位矩阵.

2) 在Step 3中,对所有样本点降维的条件为

$$\min \varepsilon(Y) = \sum_{i-1}^{n} \left| y_i - \sum_{j-1}^{k} w_j^i y_{ij} \right|^2. \tag{5}$$

其中:$\varepsilon(Y)$为损失函数,y_i是x_i的输出向量,$y_{ij}(j = 1,2,\cdots,k)$为y_i的k个近邻点. 式(5)需要满足条件

$$\sum_{i-1}^{N} y_i = 0,\ \frac{1}{N}\sum_{i-1}^{N} y_i y_i^{\mathrm{T}} = I,$$

其中I为$m \times m$的单位矩阵. 在$N \times N$的矩阵W中,存储$w_j^i(i = 1,2,\cdots,N)$. 只有当x_j和x_i是近邻点时,$W_{i,j} = w_j^i$,如果不是,则$W_{i,j} = 0$. 可以将损失函数重写为

$$\min \varepsilon(Y) = \sum_{i-1}^{N}\sum_{j-1}^{N} M_{i,j} y_i^{\mathrm{T}} y_i. \tag{6}$$

其中M是一个$N \times N$的对称矩阵,即

$$M = (I - W)^{\mathrm{T}}(I - W). \tag{7}$$

算法要求M的值最小,这便需要升序排列M的特征值,取M中最小的m个非零特征值的特征向量. 另外,第1个特征值接近于零,可舍去,从而得到由第$2 \sim m+1$个特征值的特征向量组成的矩阵M'.

2 基于LLE的检测器优化生成与检测

2.1 算法思想和过程

根据以上分析,引入LLE算法对IDS的问题域进行降维处理,并基于此设计检测器优化生成算法并利用其进行检测,从而改善免疫检测器生成与检测速率慢、检测实时性较差等问题.

算法流程如下.

Step 1: 将训练数据(自体)放入矩阵$X_{N\times D}$.

Step 2 为选中的样本点寻找其k个近邻点.

Step 3: 对这k个近邻点采用Euclidean距离算法构建并得到局部重建权值矩阵.

Step 4: 由Step 2中寻找出的k个近邻点和Step3构建的局部重建权值矩阵求得第$2 \sim d+1$间特征值的特征向量,构建矩阵M'.

Step 5: 利用初始数据集乘以M'T,得到$X^{N\times d}$的数据集.

Step 6: 将降维后的$X_{N\times d}$标准化为

$$X^* = \frac{X - \min(X)}{\max(X) - \min(X)}, \tag{8}$$

其中$\max(X)$和$\min(X)$为所在维度取值的最大值和最小值.

Step 7: 随机生成候选检测器样本,使其与降维和标准化后的自体集计算亲和力大小,保留不匹配个体,作为成熟检测器放入检测器集合.

Step 8: 记录新生成检测器数量:若数量小于数量阈值n,则转至Step 7继续执行,否则结束过程.

在检测器对数据集进行检测时,由于网络中的数据是持续进入系统,每捕获到一定量的数据就将其整理成矩阵$X_{N\times D}$. 使用LLE算法得到$X_{N\times d}$,再对其进行检测操作. 若有异常则发出警告.

2.2 算法分析

首先估计算法的时间复杂度,分析如下:

1) 在求近邻样本点的过程中,任意两个样本需要做一次内积计算,每次计算的时间复杂度为$O(D)$. 由于集合中共有$N(N-1)$个点对,求近邻点的总时间复杂度为$O(DN^2)$.

2) 计算每一个x_i的权值w_j^i的时间复杂度为$O(kND)$,由式(3)可得局部优化重建权值矩阵W中,求每个元素所消耗的时间为$O(k^2)$,则W求解的总时间复杂度为$O(k^3ND)$.

3) 在高维数据降维到d维空间的过程中,需要使损失函数最小以得到M',由式(1)和(5)可得其复杂度为$O(dN^2)$.

综上,由于$k, d \ll N, d < D$,算法的总复杂度为$O(DN^2) + O(k3DN) + O(dN^2) = O(DN^2)$.

通过算法描述和时间复杂度分析可以看出,基于LLE的检测器优化生成算法具有以下优点:

1) 算法参数较少,对实验结果的影响小,且都能进行特征参数优化.

2) 数据降维处理过程中没有迭代过程,能够大幅

减小算法的时间复杂度.

3) 样本点在降维前后的局部结构能保持不变.

3 实验分析

为了对算法进行全面测试,实验从不同角度展开: 1) 确定主要参数取值; 2) 降维优化生成检测器的检测性能; 3) 与同类型方法的对比实验.

实验采用本领域知名的KDD CUP99数据集,并进行数据预处理,需要将离散型属性转换成连续型属性,如协议属性,实验设定的变换规则为: $TCP \rightarrow 1, UDP \rightarrow 2, ICMP \rightarrow 3$等. 在各个实验中,每次随机选取一定数量($\geqslant 10^4$)的数据子集进行.

3.1 参数设置

3.1.1 参数k的设定

k指某样本的邻接点数,是一个经验参数,通过大量的实验论证得到,文献[12]建议选取为12.

3.1.2 参数r的设定

r为检测器的检测半径,主要通过大量的实验确定其取值. 本实验选取数据集样本量为10^4,并设定r的初始值为0.1,步长为0.1. 然后进行检测器优化生成,设定降维空间的维度为21(由下一实验得出,该维度较为合理). 当生成成熟检测器数量满足100时,记录所消耗的时间,且利用其对数据集进行检测. 反复进行5次实验,选取耗时最短、检测率较高、误报率较低的结果更于表1. 由表1可见,当r选取为0.5时,检测器优化生成所耗费的时间代价最小,而且检测率和误报率结果也较为理想,因此本实验环境下r取0.5最为合适.

表1　r 参数的确定

r	时间/s	检测	
		检测率	误报率
0.1	18.0	0.45	0.42
0.2	12.5	0.52	0.34
0.3	10.0	0.55	0.33
0.4	9.5	0.68	0.10
0.5	5.5	0.75	0.12
0.6	6.0	0.62	0.25
0.7	8.5	0.50	0.21
0.8	12.0	0.35	0.22
0.9	19.0	0.38	0.25

3.1.3 维数d的设定

d为降维后数据维度. 实验选取数据集中标记类型的10%数据进行不同维度的降维优化生成检测器操作并利用其进行测试,每个维度生成检测器450个,测试3次取平均值,如图1所示. 由图1可见,在维度为21时,所生成的检测器综合检测性能较为理想,因此本文实验选取的降维维度为21维. 另外,在操作不同数据集时,LLE算法都可通过对样本数据进行特征值计算得到相应合理的d值.

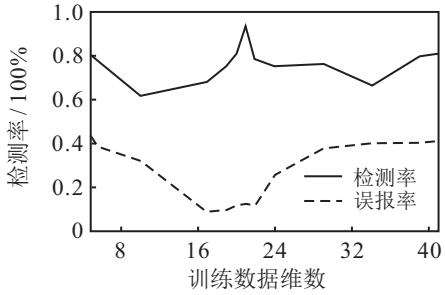

图1　不同降维维数训练数据集所生成检测器的检测性能

3.2 检测性能测试

3.2.1 检测稳定性测试

IDS对数据检测的稳定性十分重要. 本实验过程同上,反复进行5次,汇总检测率和误报率,计算均值和标准差如表2所示. 总体上看,通过LLE算法降维优化并生成的检测器的检测性能在大幅度削减信息量的前提下依然具有较好的稳定性.

表2　不同维度下检测率和误报率实验结果

维数	检测率		误报率	
	均值	标准差	均值	标准差
5	0.798 95	0.000 06	0.397 513	0.000 011
10	0.618 647	0.002 89	0.323 872	0
17	0.685 98	0.009 881	0.086 951	0
19	0.749 988	0.000 289	0.096 655	0.005 961
20	0.809 99	0	0.116 951	0.000 017
21	0.937 744	0.009 767 8	0.123 397	0.0137 784
22	0.78	0.000 013 5	0.116 951	0.000 017
24	0.750 001	0.000 002 3	0.253 575	0.056 325
29	0.760 75	0.021 109 3 8	0.376 746	0.018 767
34	0.661 704	0	0.397 877	0.000 068
39	0.788 492	0.019 440 8	0.401 393	0.002 857 8

3.2.2 对比实验测试

人工免疫降维处理中用过很多方法,其中以PCA方法居多. 本实验选取基于PCA的检测器优化生成算法进行算法对比实验. 实验设定相同的实验环境,结果如图2和图3所示. 由图2可见: PCA算法所耗费的时间随着数据量的增加而增速逐渐加快; 本文算法随着测试数据量的增加,所消耗的时间增长曲线较为缓慢,从而可以推断其可以在较大范围内保证系统能够对测试数据进行实时处理. 由图3可见,本文算法优化生成的检测器在保证检测速率的基础上,随着成熟检测器数量的不断增多,检测性能也好于基于PCA的检测器生成算法所产生的检测器.

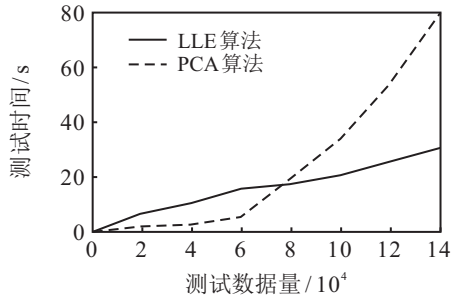

图 2 两种算法所生成检测器对测试数据量处理时间对比

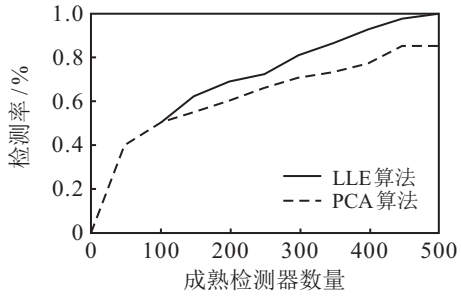

图 3 两种算法生成检测器的检测性能对比

通过以上实验的对比观察可以得出以下的结论: 1) 应用LLE算法可以较大地提高检测器的生成速率和对检测数据的检测速度; 2) 由于基于本文算法的检测器在维度上降低可以有效地为系统减负, 降低了系统在检测阶段由于处理速率过慢而造成的丢包率, 拥有较高的检测效率, 同时也具有较好的检测性能. 另外, 在操作不同维度、不同数据规模的数据集时, 算法可根据前期的数据预处理得到相应合理的参数取值, 从而满足不同数据集的实际应用需要. 前期预处理计算中, 应随机选取一定规模的数据样本(建议选取经验样本总量的10 %以上, 且数据量至少为10^4)进行特征值等计算, 从而可以得到较为合理的参数取值.

4 结 论

本文通过在免疫入侵检测中引入局部线性嵌入算法, 对数据集进行预处理, 使其能够更好地适应免疫入侵检测的实际需要. 算法通过计算合适特征值所对应的特征向量对数据集进行降维操作, 利用降维后的自体集训练检测器, 达到优化检测器的目的. 使用优化后的检测器检测待检测样本, 可显著提升检测速度, 满足系统的实时性需求. 且该算法保证数据降维前后的内在流行结构不发生变化, 依然保留数据的本质特征, 从而可以在提高检测效率的同时保证检测性能. 这对于IDS及其他安全技术实时有效地检测与防御入侵行为意义重大.

参考文献(References)

[1] Aouabdi S, Taibi M, Bouras S, et al. Using multi-scale entropy and principal component analysis to monitor gears degradation via the motor current signature analysis[J]. Mechanical Systems and Signal Processing, 2017, 90: 298-316.

[2] Wang Y L, Wu Y, Yi S C, et al. Complex multidimensional scaling algorithm for time-of-arrival-based mobile location: A unified framework[J]. Circuits, Systems, and Signal Processing, 2016, 36(4): 1-15.

[3] Ji R R, Liu H C, L J. Toward optimal manifold hashing via discrete locally linear embedding[J]. IEEE Signal Processing Society, 2017, 26(11): 5411-5420.

[4] 柴争义, 王献荣, 王亮. 用于异常检测的实值否定选择算法[J]. 吉林大学学报: 工学版, 2012, 42(1): 176-181. (Chai Z Y, Wang X R, Wang L. Real-valued negative selection algorithm for abnormal detection[J]. J of Jilin University: Engineering Edition, 2012, 42(1): 176-181.)

[5] Jain V K, Tapaswi S, Shukla A. RSS fingerprints based distributed semi-supervised locally linear embedding (DSSLLE) location estimation system for indoor WLAN[J]. Wireless Personal Communications, 2013, 71(2): 1175-1192.

[6] Yang Y, Jiang D. Casing vibration fault diagnosis based on variational mode decomposition, local linear embedding, and support vector machine[J]. Shock and Vibration, 2017, 1971: 1-14.

[7] Zhang L, Leng Y, Yang J, et al. Supervised locally linear embedding algorithm based on orthogonal matching pursuit[J]. IET Image Processing, 2015, 9(8): 626-633.

[8] Peng T, Yang X. Optimize parameter K in locally linear embedding based on spatial distributions[C]. The 12th Int Conf on Fuzzy Systems and Knowledge Discovery. Zhangjia jie IEEE, 2015: 1588-1595.

[9] Lipinski P. Training complex decision support systems with differential evolution enhanced by locally linear embedding[C]. European Conf on the Applications of Evolutionary Computation. Berlin Springer. Heidelberg, 2014: 125-137.

[10] Zhang L, Zhao C. Sparsity divergence index based on locally linear embedding for hyperspectral anomaly detection[J]. J of Applied Remote Sensing, 2016, 10(2): 1-24.

[11] Deng T, Deng Y, Shi Y, et al. Research on improved locally linear embedding algorithm[C]. Bio-Inspired Computing-Theories and Applications. Berlin Springer Heidelberg, 2014: 88-92.

[12] Roweis S T, Saul L K. Nonlinear dimensionality reduction by locally linear embedding[J]. Science, 2000, 290(5500): 2323-2326.

(责任编辑: 郑晓蕾)