有监督 S-kv-Isomap 在入侵检测中的应用

郑凯梅[1,2],钱 旭[1]
ZHENG Kai-mei[1,2],QIAN Xu[1]

1.中国矿业大学 机电与信息工程学院,北京 100083
2.中国防卫科技学院 信息工程系,北京 101601
1.School of Mechanical Electronic & Information Engineering,China University of Mining & Technology Beijing,Beijing 100083,China
2.Information Engineering Department,China Institute of Defense Science and Technology,Beijing 101601,China
E-mail:zhengkaimei@126.com

ZHENG Kai-mei,QIAN Xu.Intrusion detection based on supervised S-kv-Isomap algorithm.Computer Engineering and Applications,2010,46(3):20-22.

Abstract: Intrusion detection is still a hot area in computer security.When performing visualization and classification,the problem should be confronted is the high dimensionality.As one of the manifold learning algorithms Isomap is an effective nonlinear dimension reduction tool.However,when Isomap is applied to the real-world data,it shows some limitations,such as failing to guarantee connectedness of the constructed neighborhood graphs and not using the class labels of the data.An improved version of Isomap,namely S-kv-Isomap,is proposed.S-kv-Isomap utilizes class information to guide the dimension reduction procedure and k-variable method to build connected neighborhood graphs so as to enhance the robustness.The new scheme is evaluated with KDD CUP 1999 datasets in visualization and classification on four kinds of intrusion types: Dos,R2L,Probe,and U2R.Experiment results show that S-kv-Isomap performs best compared with kv-Isomap in visualization.In the classification test,S-kv-Isomap is compared with kv-Isomap,SVM,and k-NN.The results show that S-kv-Isomap performs higher detection rate and very low false positive rate.

Key words: supervised learning;dimension reduction;manifold learning;Isomap;visualization;classification;intrusion detection

摘 要:入侵检测是计算机安全研究方面的热点领域,在入侵检测数据可视化和分类方面面临的问题是其高维特性。流形学习算法 Isomap 是有效的非线性降维工具。但是 Isomap 算法在实际应用中存在不能保证构造连通的邻接图和没有利用样本已知类别标记的缺点,针对上述缺陷提出了健壮的有监督 S-kv-Isomap 算法。该算法利用类别标记来指导降维,并且利用 k-variable 算法构造联通的邻接图。实验选用 KDDCUP1999 数据集,对四类入侵数据即 Dos、R2L、Probe、U2R 进行了可视化和分类研究。可视化中比较了 S-kv-Isomap 算法与 kv-Isomap 算法, 前者具有更好的可视化效果。在分类研究中比较了 S-kv-Isomap、kv-Isomap、SVM 和 k-NN 算法,实验结果表明,S-kv-Isomap 方法在入侵检测中不仅保持较高的入侵检测率,而且误警率很低。

关键词:有监督学习;维数约简;流形学习;Isomap;可视化;分类;入侵检测
DOI:10.3778/j.issn.1002-8331.2010.03.006 **文章编号:**1002-8331(2010)03-0020-03 **文献标识码:**A **中图分类号:**TP393

1 问题的提出

随着计算机技术和通讯技术的迅猛发展, 各种基于 Internet/Intranet 的网络服务如电子商务、电子政务等得到广泛应用。越来越多的来自于网络内部职员或外部黑客发动的攻击事件层出不穷,给企业或社会造成极大损失。由 Anderson[1]最初提出的入侵检测系统仍然是网络安全方面的研究重点。

入侵检测可以分为误用检测和异常检测。误用检测首先提取异常模式的特征,然后通过比照来判断是否入侵。这种方法的优点是能有效地检测到已知攻击,缺点是不能识别新的攻击类型。而异常检测的核心思想是利用预定行为来发现背离正常行为的异常问题。该方法的优点是能够识别新的攻击行为,缺点是误警率较高。

目前很多数据挖掘或机器学习算法如人工神经网络、支持向量机、聚类、决策树、人工免疫和遗传算法等已经用于入侵检测系统中进行异常检测,但是绝大多数算法使用了数据的所有特征。然而,现实中采集到的数据属性存在冗余,因此特征抽取或维数约简技术成为异常检测技术中的重要步骤。

流形学习是近年来发展迅速的特征提取或维数约简技术。Tenenbaum 于 2000 年提出的非线性维数约简算法 Isomap[2]受到了广泛的关注。已经成功应用于图像检索,人脸识别等应用

基金项目:07 教育部科学技术研究重点(重大)项目资助(the Key(Key Grant) Project of Chinese Ministry of Education,No.107021)。
作者简介:郑凯梅(1970-),女,博士研究生,副教授,主要研究领域为机器学习,信息融合,网络安全;钱旭(1962-),男,博士生导师,教授,主要研究领域为知识工程,信息融合。
收稿日期:2009-10-14 **修回日期:**2009-12-13

领域,并取得了较好的结果。将 Isomap 用于入侵检测系统中进行异常检测也取得了良好的效果[3]。文献[3]首先利用 Isomap 对样本数据和测试数据进行特征提取,然后对只有较少特征的测试数据利用支持向量机进行异常检测,有效降低了误警率。

然而,上述 Isomap 算法在入侵检测中存在不足,首先是未使用样本中的类别标记,其次由于 Isomap 算法不能保证所构造的邻接图是连通的,从而出现个别数据不能嵌入到低维空间中的现象,给可视化和分类研究带来一定的影响。

针对上述问题,提出了 S-kv-Isomap 算法,不仅实现了入侵检测数据的可视化,而且结合径向基函数网络 RBF[4]和 kNN[5]实现了异常检测分类,在保持高检测率的同时极大地降低了误警率。

2　Isomap 简介

Isomap[2]算法是在多维尺度变换(MDS)的基础上发展起来的,MDS 的基本思想是低维空间中任意两点之间的距离应该与其原始空间中的距离相同。Isomap 算法的核心思想是通过计算最邻近图中任意两点之间的测地线距离作为两点间距离的度量,然后利用多维尺度变换(MDS)得到高维数据在低维空间中的相应嵌入坐标。

Isomap 算法的输入是高维空间中所有样本点对 i,j 之间的距离 $d_x(i,j)$,可以是欧式距离或其他距离,输出是最能代表样本数据内部结构的 d-维欧式空间中的坐标向量 $\boldsymbol{y}$。Isomap 算法步骤如下:

步骤 1　构造邻接图。任意两点 x_i,x_j 之间的距离用欧式距离 $d_x(x_i,x_j)$表示,邻接关系为 k-邻近或 ε-球。

步骤 2　计算最短路径。通过计算图的最短路径距离来估计所有的点对在流形 M 上的测地距离。

步骤 3　对得到的图形距离矩阵通过经典 MDS 变换,构造在 d-维欧式空间中最能代表流形内部结构的低维嵌入坐标。

3　S-kv-Isomap 算法

Isomap 算法由于从全局入手,能更好地反映数据的流形结构。但是,在实际应用中,存在邻接图不连通的情况,即不能把所有的数据全部嵌入在低维欧式空间中。文献[6]提出的 k-variable 算法可以有效地解决邻接图不连通的问题。

k-variable 算法的核心思想是设置一个初始邻居数目 k,得到当前 k 值下的邻接图,然后计算该邻接图的连通性,如果有数据点没有落在连通的邻接图上,则在一个阈值控制下增加邻居 k 的数目,再次计算邻接图的连通性,此时必有部分落在旧连通图上的结点经邻居数目增加后落在新的连通图上,如此反复,直到所有结点都位于一个连通图上。

事实上,Isomap 是一种非监督学习算法,在学习过程中不考虑样本的类别标签,而样本的类别标签在可视化和分类中是有用的。文献[7]提出了有监督的 S-Isomap 算法用于数据可视化和分类。针对上述两类问题,提出了有监督的 S-kv-Isomap 算法。算法步骤如下:

步骤 1　用相异度距离 $D(x_i,x_j)$代替 Isomap 第一步中的距离 $d_x(x_i,x_j)$[7]。

假设样本为(x_i,y_j),i=1,2,…,N,其中 x_i 为样本观测值,y_j 是样本 x_i 的类别标签,$d_x(i,j)$表示任意两点 x_i,x_j 之间的欧式距离。相异度距离 $D(x_i,x_j)$公式如下所示:

$$D(x_i,x_j)=\begin{cases}\sqrt{1-\mathrm{e}^{\frac{-d^2(x_i,x_j)}{\beta}}} & y_i=y_j\\ \sqrt{\mathrm{e}^{\frac{-d^2(x_i,x_j)}{\beta}}}-\alpha & y_i\neq y_j\end{cases}\tag{1}$$

公式(1)中的 β 用来防止当 $d_x(x_i,x_j)$相对较大时 $D(x_i,x_j)$增长过快。一般设置为所有点对之间欧式距离的平均值。公式(1)中的 α 用来调节不同类别标签的点对之间的相似程度。

步骤 2　用上一步得到的相异度距离 $D(x_i,x_j)$和输入的初始邻居数目 k 计算最短路径。

步骤 3　判别邻接图连通性并记录不在连通图中的数据。

步骤 4　如果存在没有位于连通图中的数据,利用公式 $k=k+\Delta$ 或 $k=k*(1+\Delta)$来迭代增加邻居数目。否则转步骤 7。

步骤 5　利用增加后的邻居数目 k 和距离 $D(x_i,x_j)$计算最短路径。对所有不在连通图中的数据,用新计算的最短路径数据代替旧的数据。

步骤 6　转到步骤 3。

步骤 7　对得到的图形距离矩阵通过经典 MDS 变换,构造在 d-维欧式空间中最能代表流形内部结构的低维嵌入坐标。

把上述步骤 1 中的相异度距离替换为原始 Isomap 中的欧几里德距离,执行步骤 2 到步骤 7 即 kv-Isomap 算法[6]。

4　实验

4.1　S-kv-Isomap 用于入侵检测数据可视化

入侵检测数据可视化的目标是把高维数据投影到二维或三维欧式空间中以便可以有效分离不同类别的数据。尽管 I-somap 可以实现数据的可视化,但是当有噪声存在时可视化效果会受到很大影响,并且 Isomap 算法不能保证构造包含所有数据的连通邻接图。提出的 S-kv-Isomap 不仅消除了邻接图不连通的缺点,使所有的数据点都可以嵌入到低维特征空间中,而且,由于利用了样本中的类别信息,使得同一类别的数据更加紧密而不同类别的数据离得更远,从而提高可视化效果。

可视化实验采用 KDD CUP 1999[8]入侵检测数据集。该数据集收集了美国空军局域网仿真环境下的 TCP/IP dump 文件,全部由 TCP/IP 连接数据组成。每个 TCP/IP 连接包含 41 个属性以及一个标定为正常或特定攻击类型的属性。数据集中的攻击类型大致可以归为 4 类:(1)拒绝服务攻击 Dos(Denial of Service);(2)远端未授权访问 R2L (Remote to Local);(3)超级用户未授权访问 U2R(User to Root);(4)系统漏洞探测 Probe。

实验中随机抽取正常样本和异常样本,表 1 所示为实验数据集组成结构。

表 1　实验数据集结构

	正常样本	异常样本	攻击类型
数据集 1	2 000	1 000	Dos
数据集 2	2 000	1 000	R2L
数据集 3	2 000	1 000	Probe
数据集 4	2 000	200	U2R

S-kv-Isomap 可以有效地对上述 4 个数据集构造包含所有样本数据的连通邻接图,图 1 显示了 S-kv-Isomap 算法执行过程中前 4 个数据集的迭代次数和当前连通邻接图上的样本点数目,第 4 个数据集中由于样本少,因此不需要额外迭代就得到了全部样本的低维坐标。实验结果表明,Isomap 算法不能保证构造包含所有数据在内的连通邻接图,而应用 S-kv-Isomap 算法后,经过若干次迭代,所有入侵检测数据集中的数据都可以得到其低维空间中的坐标,从而便于实现数据的可视化。

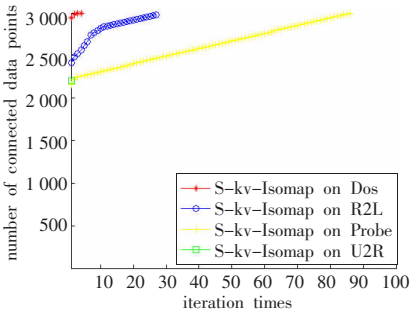


图 1　迭代次数与连通邻接图样本数目

对 S-kv-Isomap 和 kv-Isomap 这两种算法在入侵检测数据的可视化中的应用进行了比较。图 2 和图 3 分别显示了上述 4 个数据集利用上述两类方法进行可视化的比较结果。实验结果表明由于利用了数据的类别标记,S-kv-Isomap 比 kv-Isomap 具有更好的可视化效果。从图 2 中可以看出 Probe 和 Dos 攻击与正常连接有明确的边界，而 R2L 和 U2R 攻击和部分正常连接数据稍微混杂在一起,但是仍然是可以区分的。后面的分类实验结果也证明了经过 S-kv-Isomap 算法进行维数约简后获得了很高的分类性能。从图 3 中可以看出 Dos 攻击与正常连接有明确的边界，而 Probe、U2R、R2L 攻击和部分正常连接数据不同程度地混杂在一起。从而验证了很多分类方法对 Dos 攻击都有很好的识别效果,而对 Probe,R2L、U2R 攻击的分类效果则因算法不同而检测性能差异较大。

4.2　S-kv-Isomap 用于入侵检测分类

4.2.1　S-kv-Isomap 用于分类

Isoamp 算法把当前所有数据当成一个整体来进行维数约简,当有新的数据加入时,必须把新数据加入后重新计算,这是相当消耗资源和时间的。为避免这样的问题,提出了将 S-kv-Isomap 与径向基函数网络 RBF[4]结合进行维数约简,首先对包含有正常样本和异常样本的训练数据使用 S-kv-Isomap 进行维数约简,然后利用径向基函数网络 RBF 进行训练,最后使用训练好的 RBF 网络得到测试数据在低维空间中的嵌入坐标。

RBF 神经网络是一个三层网络。第一层是输入层,由源结点组成。第二层是隐层,由许多隐结点构成,第三层是输出层。从输入层到隐结点层的变换是非线性变换,而从隐结点层到输出层的变换属于线性变换。经过实验，使用的 RBF 网络由 41 个源结点、56 个隐结点和 9 个输出结点构成。

在低维特征空间中分类使用的算法是 k-NN 算法[5]。

S-kv-Isomap 用于分类的步骤如下：

第一步　对于由正常样本和异常样本组成的训练数据,结合其类别标记,应用 S-kv-Isomap 算法把高维特征空间数据映射到低维特征空间。

第二步　训练 RBF 网络,其中输入层数据来源于原始训练数据,输入层结点个数等同于训练数据的原始特征维数,输出层数据来源于第一步得到的低维空间坐标,输出层结点个数为低维空间坐标的维数。

第三步　对于测试样本数据，利用第二步得到的训练好的 RBF 网络,计算所有测试样本数据的低维空间坐标。

第四步　对于所得到的已经映射到低维空间中的训练数据和测试数据,使用 k-NN 算法对样本进行分类。

4.2.2　S-kv-Isomap 入侵检测分类实验

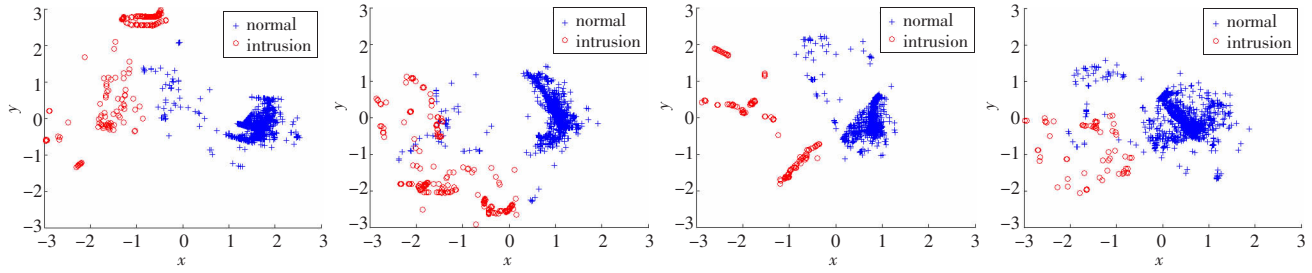
选择 KDD CUP 1999 数据集作为实验数据源，训练数据集和测试数据集如表 2 所示,所有数据随机抽取。

表 2　实验数据集结构

	训练集		测试集		攻击类型
	正常样本	异常样本	正常样本	异常样本	
数据集 1	2 000	1 000	1 000	500	Dos
数据集 2	2 000	1 000	1 000	500	R2L
数据集 3	2 000	1 000	1 000	500	Probe
数据集 4	2 000	100	1 000	100	U2R

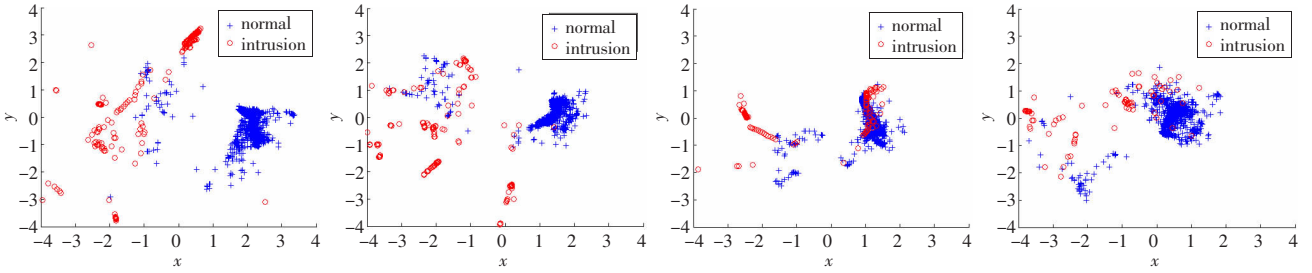
分类算法的性能通过两个指标来衡量:检测率 *TP*,即被检测出来的异常样本数占异常样本总数的比例,以及误警率 *FP*,即正常样本数误报为异常样本的数目占正常样本总数的比例。

比较了提出的 S-kv-Isomap 分类算法和 kv-Isomap+RBF+k-NN,SVM[9]以及 k-NN 算法在入侵检测分类中的性能。检测率 *TP* 和误警率 *FP* 实验结果如表 3 所示。kv-Isomap 方法对



(a)正常连接与 Dos 攻击可视化　(b)正常连接与 R2L 攻击可视化　(c)正常连接与 Probe 攻击可视化　(d)正常连接与 U2R 攻击可视化

图 2　S-kv-Isoamp 用于入侵检测数据可视化



(a)正常连接与 Dos 攻击可视化　(b)正常连接与 R2L 攻击可视化　(c)正常连接与 Probe 攻击可视化　(d)正常连接与 U2R 攻击可视化

图 3　kv-Isoamp 用于入侵检测数据可视化

(下转 66 页)