Journal of Computer Applications
计算机应用,2012,32(9):2606 - 2608
ISSN 1001-9081
CODEN JYIIDU
2012-09-01
http://www.joca.cn

文章编号:1001 - 9081(2012)09 - 2606 - 03
doi:10.3724/SP.J.1087.2012.02606

基于 LISOMAP 的相关向量机入侵检测模型

唐朝伟,李超群*,燕　凯,严　鸣
(重庆大学 通信工程学院,重庆 400030)
(*通信作者电子邮箱 ly1202033@sina.com)

摘　要:针对现有入侵检测模型分类检测精度低、误报率高的问题,提出一种基于地标等距映射(LISOMAP)的相关向量机(RVM)入侵检测分类模型。首先采用 LISOMAP 对训练样本中的数据进行非线性降维,结合深度优先搜索(DFS)参数优化的 RVM 进行分类检测。结果表明,该模型与基于主成分分析(PCA)法的支持向量机(SVM)、基于 LISOMAP 的 SVM 模型相比,在保证一定检测率的情况下,误报率有了明显下降。

关键词:入侵检测;主成分分析;支持向量机;地标等距映射;相关向量机;深度优先搜索
中图分类号:TP393.083　　**文献标志码**:A

Intrusion detection model based on LISOMAP relevant vector machine

TANG Chao-wei, LI Chao-qun*, YAN Kai, YAN Ming
(*College of Communication Engineering, Chongqing University, Chongqing* 400030, *China*)

Abstract: Concerning low classification accuracy and high false alarm rate of current intrusion detection models, an intrusion detection classification model based on Landmark ISOmetric MAPping (LISOMAP) and Deep First Search (DFS) Relevant Vector Machine (RVM) was proposed. The LISOMAP was adopted to reduce the dimension of the training data, and RVM based on the DFS was used for classification detection. Compared with the Principal Components Analysis (PCA)-Supported Vector Machine (SVM), the experimental results indicate that the LISOMAP-DFSRVM model has lower false alarm rate with almost the same detection rate.

Key words: intrusion detection; Principal Component Analysis (PCA); Support Vector Machine (SVM); Landmark ISOmetric MAPping (LISOMAP); Relevant Vector Machine (RVM); Deep First Search (DFS)

0　引言

入侵检测本质上是一个分类问题:区分待检测数据是攻击还是正常。但在处理检测数据时,待检测样本数目的增加,对分类速度影响很大[1]。因此在建立模型时应考虑特维数约减和特征分类两个方面。

特征抽取的目的是找出隐藏在高维数据的低维结构。主成分分析(Principal Components Analysis, PCA)法广泛应用于入侵检测测试数据的高维维数约减[2-3],然而该方法受限于网络实际数据的非线性情况,性能效果并不理想。文献[4]中引入的地标等距映射(Landmark ISOmetric MAPping, LISOMAP)对高维检测数据进行非线性降维,取得更优的漏报率效果。

在特征分类方面,支持向量机(Support Vector Machine, SVM)以其良好的一般化特性以及解决维数分类能力得到广泛应用,其中基于遗传算法的 SVM[5-6]、核二叉树多分类 SVM[7]、中间分类超平面 SVM[8]、基于结构风险最小化的 SVM[9]、基于因子分析降维的 SVM[10]都取得了很好的分类效果。但 SVM 惩罚参数的设置以及 Mercer 定理的限制,对分类结果存在着一定的不利影响。基于遗传相关向量机(Relevant Vector Machine, RVM)[11]算法进行图像分类检测,实验结果表明,相关向量机检测算法与传统 SVM 分类检测算法相比不但克服了 SVM 的一些惩罚参数和 Mercer 定理限制,而且有着更高的分类精度。

针对传统入侵检测模型中非线性数据降维和 SVM 分类检测中存在的问题,结合 ISOMAP 和相关向量机的优点,提出了基于 LISOMAP 的相关向量机入侵检测分类模型并通过深度优先搜索(Deep First Search, DFS)方法用于确定相关向量机的核参数。该方法选取分类错误率作为评估标准,采取交叉验证方法对核参数进行评估,通过启发式方法缩小参数的选择范围,从而获得较佳的核参数。

1　模型相关算法

1.1　LISOMAP 算法

等距映射[12](ISOmetric MAPping, ISOMAP)方法是一种非线性维数约减方法,其原理是将高维点的邻域映射到低维的线性投影。不同于其他非线性维数约减方法,它强调单算法的执行且避免了非线性规划的局部最小化。

ISOMAP 算法是基于多维标度(MultiDimensional Scaling, MDS)。MDS 基本思想是约简后在低维空间中任意两点间的距离应该与它们在原高维空间中的距离相同。ISOMAP 算法的思路是首先构建一个图,图的顶点就是所有数据点。当两个点互为邻域点时,连接这两点并将这条边的权重设置为两点的欧氏距离。在图中计算顶点间的最短路径,得到所有数据点间的距离矩阵,从而得到所有数据点之间的差异度矩阵。最后利用 metric MDS 算法进行降维。

为了克服 ISOMAP 算法的最短路径矩阵计算复杂度和 MDS 特征集计算复杂度较高的瓶颈,本文采取 LISOMAP 算法[13],其具体步骤如下:1)采用 metric MDS 算法计算只与 Landmark 有关的距离矩阵 $\boldsymbol{D}$;2)采用地标 MDS 算法嵌入其

收稿日期:2012-02-29;**修回日期**:2012-05-13。　**基金项目**:国家科技重大专项基金资助项目(2009ZX03004-002)。
作者简介:唐朝伟(1966 -),男,四川达州人,教授,博士,CCF 会员,主要研究方向:宽带移动多媒体通信、互联网内容识别及处理、分布式网络安全及抗毁路由;　李超群(1985 -),男,河北唐山人,硕士研究生,主要研究方向:分布式防火墙及路由抗毁;　燕凯(1989 -),男,山东淄博人,硕士研究生,主要研究方向:分布式网络安全;　严鸣(1988 -),男,陕西安康人,硕士研究生,主要研究方向:分布式网络安全。

他数据点;3)利用 PCA 算法对齐数据和坐标。

1.2　RVM 分类检测方法

1.2.1　RVM 理论

RVM 训练源于 Tipping 建立的稀疏贝叶斯学习理论，能在概率意义下进行合理划分,使得分类函数对于训练集似然函数值最大。设样本集$\{\boldsymbol{x}_i,t_i\}_{i=1}^N$,输入输出相互独立,目标值$\{t_i\}_{i=1}^N$带有均值 0 方差$\delta^2$的高斯噪声$\varepsilon_n$。存在表达式:

$$y(\boldsymbol{x};\boldsymbol{w}) = \sum_{i=1}^{N} w_i K(\boldsymbol{x},\boldsymbol{x}_i) + \varepsilon_n \tag{1}$$

$$p(t_i) = \mathrm{N}(t_i \mid y(\boldsymbol{x}_i;\boldsymbol{w}),\delta^2) \tag{2}$$

其中:$K(\boldsymbol{x},\boldsymbol{x}_i)$是核函数,$N$为样本集长度,$\boldsymbol{x}_i$表示输入向量,$\boldsymbol{w}$代表权重向量,N(·)表示正态分布密度函数。N(·)的定义如下:

$$\mathrm{N}(x \mid \boldsymbol{\mu},\delta^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{x^2}{2\sigma^2}\right) \tag{3}$$

$$p(\boldsymbol{t} \mid \boldsymbol{w},\sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{-\frac{N}{2}}\exp\left(-\frac{\|\boldsymbol{t}-\boldsymbol{\Phi w}\|^2}{2\sigma^2}\right) \tag{4}$$

其中$\boldsymbol{\Phi}$表示各特征向量代入核函数得到的设计矩阵。

通过为$\boldsymbol{w}$加入均值为 0 的高斯先验分布变量$\boldsymbol{\alpha}$来避免最大似然估计的过调试,如下所示:

$$p(\boldsymbol{w} \mid \boldsymbol{\alpha}) = \prod_{j=0}^{N}\sqrt{\frac{\alpha_j}{2\pi}}\exp\left(-\frac{\alpha_j}{2}w_j^2\right) \tag{5}$$

其中:$\boldsymbol{\alpha}$称为超参数,$\boldsymbol{\alpha} = [\alpha_1,\alpha_2,\cdots,\alpha_N]^{\mathrm{T}}$,每个超参数与权值独立。

由贝叶斯公式可计算权重的后验分布:

$$p(\boldsymbol{w} \mid \boldsymbol{t},\boldsymbol{\alpha},\sigma^2) = (2\pi)^{-\frac{N+1}{2}} \mid \boldsymbol{\Sigma} \mid^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\times (\boldsymbol{w}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{w}-\boldsymbol{\mu})\right\} \tag{6}$$

其中:

$$\boldsymbol{\Sigma} = (\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{\Phi})^{-1}$$

$$\boldsymbol{\mu} = \boldsymbol{\Sigma}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{t}$$

$$\boldsymbol{A} = \mathrm{diag}(a_0,a_2,\cdots,a_N),\boldsymbol{B} = \sigma^2\boldsymbol{I}_N$$

σ^2作为超参数处理,可从数据中得到。对权重求积分,可得超参数边界似然为:

$$p(\boldsymbol{t} \mid \boldsymbol{\alpha},\sigma^2) = (2\pi)^{-\frac{N}{2}} \mid \boldsymbol{B}^{-1}+\boldsymbol{\Phi}\boldsymbol{A}^{-1}\boldsymbol{\Phi}^{\mathrm{T}} \mid^{-\frac{1}{2}}\times \exp\left\{-\frac{1}{2}\boldsymbol{t}^{\mathrm{T}}(\boldsymbol{B}^{-1}+\boldsymbol{\Phi}\boldsymbol{A}^{-1}\boldsymbol{\Phi}^{\mathrm{T}})^{-1}\boldsymbol{t}\right\} \tag{7}$$

求解$\boldsymbol{\alpha}$和σ^2可通过迭代方法求得:

$$\alpha_i^{\mathrm{new}} = \frac{\gamma_i}{\mu_i^2} \tag{8}$$

$$(\sigma^2)^{\mathrm{new}} = \frac{\|\boldsymbol{t}-\boldsymbol{\Phi}\boldsymbol{\mu}\|^2}{N-\sum_{i=0}^{N}\gamma_i} \tag{9}$$

$$\gamma_i = 1-\alpha_i\Sigma_{i,i} \tag{10}$$

其中$\Sigma_{i,i}$是$\boldsymbol{\Sigma}$中第i项对角线上的元素。

相关向量机的分类[14]需要采用分类问题常用的 S 函数,在观测值为独立事件的前提下,得到观测结果为$\boldsymbol{t}$的概率为:

$$P(\boldsymbol{t} \mid \boldsymbol{w}) = \prod_{i=1}^{N}\sigma[y(\boldsymbol{x}_i;\boldsymbol{w})]^{t_i}\{1-\sigma[y(\boldsymbol{x}_i;\boldsymbol{w})]\}^{1-t_i} \tag{11}$$

分类算法的流程为:

1)对于已知的$\boldsymbol{\alpha}$值通过迭代法找到最大值近似权重$\boldsymbol{\omega}_{\mathrm{MP}}$。

2)计算协方差矩阵:

$$\boldsymbol{\Sigma} = (-\boldsymbol{H} \mid \boldsymbol{\omega}_{\mathrm{MP}})^{-1} = (\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{\Phi}+\boldsymbol{A})^{-1} \tag{12}$$

3)使用$\boldsymbol{\omega}_{\mathrm{MP}}$代替$\boldsymbol{\mu}$,配合上一步$\boldsymbol{\Sigma}$更新$\boldsymbol{\alpha}$,具体方法见式(8)。

1.2.2　RVM 核参数的深度优先搜索优化

在利用 RVM 进行样本集分类之前,首先需要进行核函数的选取,确定核函数的参数。本文选取径向基高斯函数作为核函数,因此需要确定的参数为径向基函数高斯核的宽度σ。本文采用深度优先搜索对 RVM 核参数进行优化,具体步骤如下:

1)对 RVM 参数σ按照参考值进行初始化设置,按照步长为均匀分布的离散数组表示:σ的范围设定为[0, 10],步长设定为 0.5,则离散数组表示为[0,0.5,1,…,9.5,10]。

2)采用交叉验证方法对σ进行评估,选取分类错误率作为评估标准进行比较,并选取具有最小分类错误率的σ作为本轮最优σ。

3)将本轮最优σ的分类错误率与上一轮最优σ的分类错误率作对比,如果小于上一轮就跳转 4),否则就跳转 5)。

4)根据上一轮最优的σ值在其附近采取启发式方法进行参数范围选择。具体方法为:原步数按((*stepnum* − 2) >> 1) + 2 减少,*stepnum* 表示上一轮的步数,参数范围上下限为最优参数 ∓ 步长,然后转 1),重新进行递归运算。

5)参数优化结束,退出运算,得到相应的最优σ。

2　入侵检测分类模型

本文的入侵检测系统中,网络数据先通过 ISOMAP 模块进行特征提取,接着通过 RVM 分类模块进行分类,根据分类情况提出异常行为告警。

RVM 分类模块包括 RVM-DoS、RVM-Probe、RVM-U2R、RVM-R2L、RVM-Normal 5 个子模块。每个子模块分为训练和检测模块。RVM 入侵检测分类模型如图 1 所示。

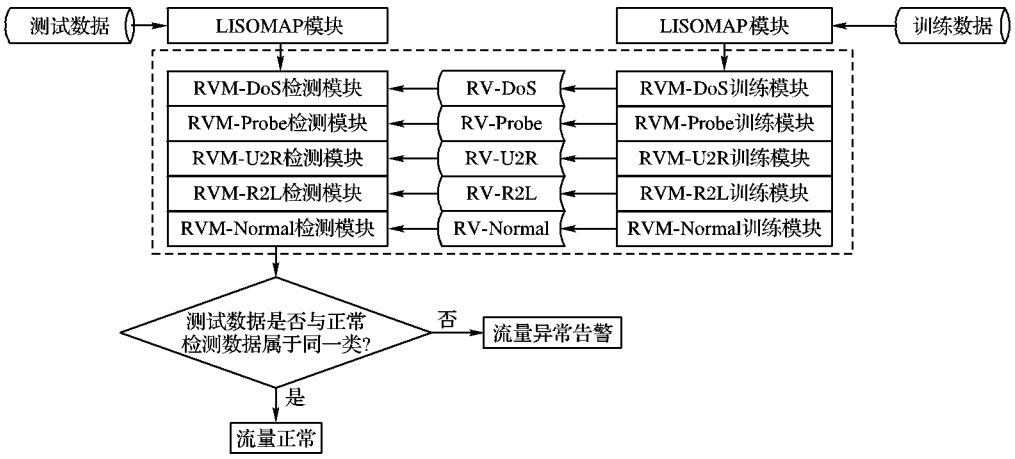

图 1　RVM 入侵检测分类模型

图 1 中的相关向量机分类器采用的分类函数为：

$$f(\boldsymbol{x}) = \boldsymbol{\Phi}^{\mathrm{T}}(\boldsymbol{x})\left[\sum_{i=1}^{n} a_i \boldsymbol{\Phi}(\boldsymbol{x}_i)\right] \tag{13}$$

采用径向基函数高斯核函数：

$$K(\boldsymbol{x},\boldsymbol{x}_i) = \exp\left(\frac{\|\boldsymbol{x}-\boldsymbol{x}_i\|^2}{2\sigma^2}\right) \tag{14}$$

LISOMAP-DFSRVM 分类器处理流程如图 2 所示。

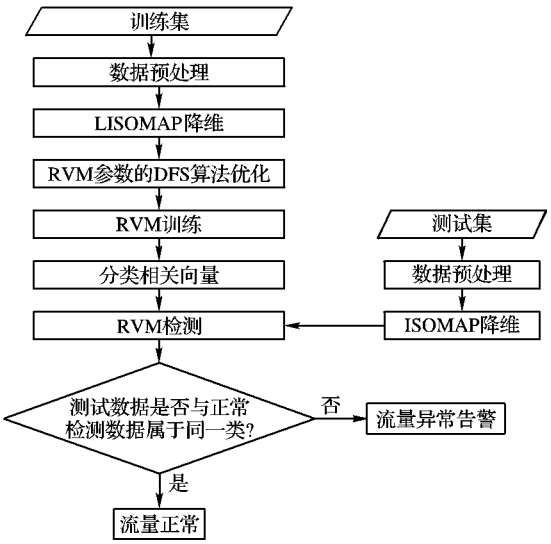

图 2 LISOMAP-DFSRVM 模型处理流程

3 实验结果与分析

为了验证本文提出的入侵检测模型对于高维非线性样本集的有效性，实验选用 MIT 林肯实验室提供的 KDD99 数据集。该数据集来自模拟真实网络环境，包含多种类型的攻击，其中代表性的 4 类为 DoS 攻击、R2L 攻击、U2R 攻击、Probe 攻击。本文采用 KDD99 数据的 10% 作为基础样本，从中提取 DoS，R2L 攻击样本集进行样本训练和分类检测。

在两种攻击分类检测实验中，提取 KDD99 的 10% correct 数据集中 2 000 个作为训练样本，另选取 2 000 个作为测试样本，其中包含 1 900 个正常数据和 100 个 DoS/R2L 数据。通过 PCA-SVM[2]，LISOMAP-SVM[4] 与本文提出的 LISOMAP-DFSRVM 对比，验证本文方法的优越性和有效性。

LISOMAP 选取样本集维数 $d = 15$，随机地标节点数目为 100，邻居数目为 18。在降维过程中，会出现少量数据点无法映射到低维空间的情况，这些数据会被视为异常数据。最后通过深度优先搜索优化得到相关向量机入侵检测分类模型。

3 种入侵检测模型的检测结果表 1 所示。

表 1 3 类入侵检测模型检测率和误报率实验结果对比 %

入侵检测模型	DoS		R2L	
	检测率	误报率	检测率	误报率
PCA-SVM	99	2.05	98	2.68
LISOMAP-SVM	97	1.10	97	1.58
LISOMAP-DFSRVM	98	0.95	97	1.26

表 1 中的检测率和误报率的计算公式如下：

$$\text{检测率} = \frac{\text{检验出的异常样本数目}}{\text{检测集中的异常样本总数}} \times 100\%$$

$$\text{误报率} = \frac{\text{正常样本被检验为异常样本的数目}}{\text{检测集中的正常样本总数}} \times 100\%$$

在 DoS 攻击检测实验中，检测率方面，LISOMAP-DFSRVM 略低于 PCA-SVM，但高于 LISOMAP-SVM；但误报率方面相比 PCA-SVM 下降了 53.7%，相比 ISOMAP-RVM 也有 15.8% 的下降。R2L 攻击检测实验中，检测率方面，LISOMAP-DFSRVM 略低于 PCA-SVM，等同于 LISOMAP-SVM；但误报率方面相比 PCA-SVM 下降了 52.9%，相比 ISOMAP-RVM 下降了 20.3%。可看出，LISOMAP-DFSRVM 在检测率变化不大的情况下，在误报率上相比 PCA-SVM 和 LISOMAP-SVM 取得了更为理想的效果。

4 结语

本文提出了基于 LISOMAP 的 DFS 参数优化 RVM 入侵检测分类模型，对样本数据集采用 LISOMAP 算法进行高位数据非线性降维。相关向量机分类器采用经过 DFS 优化参数的核函数线性组合加权的函数进行分类，通过设定权重满足均值为零方差不同的高斯概率分布，进行概率意义上的合理划分，使分类函数对训练集的似然函数值最大，避免了 SVM 的核参数设置和 Mercer 定理的限制。采用 KDD99 数据作为样本集进行测试，实验结果表明：LISOMAP-DFSRVM 入侵检测分类模型相比传统的 PCA-SVM 和 LISOMAP-SVM，在保证检测率指标不降低的情况下，误报率指标有很大改善。

参考文献：

[1] 陈友，沈华伟，李洋，等. 一种高效的面向轻量级入侵检测系统的特征选择算法[J]. 计算机学报，2007，30(8)：1398－1408.

[2] 高海华，杨辉华，王行愚. 基于 PCA 和 KPCA 特征抽取的 SVM 网络入侵检测方法[J]. 华东理工大学学报：自然科学版，2006，32(3)：321－326.

[3] 谷雨，徐宗本，孙剑，等. 基于 PCA 与 ICA 特征提取的入侵检测集成分类系统[J]. 计算机研究与发展，2006，43(4)：633－638.

[4] ZHENG K M，QIAN X，ZHOU Y，*et al*. Intrusion detection using ISOMAP and support vector machine[C]// Proceedings of 2009 International Conference on Artificial Intelligence and Computational Intelligence. Piscataway，NJ：IEEE Press，2009：235－239.

[5] 赵军. 基于 CECA-SVM 的网络入侵检测算法[J]. 计算机工程，2009，35(23)：166－167，171.

[6] 钱鹏江，王士同，徐华，等. 基于 KCCA 优化的网络入侵检测算法[J]. 计算机工程，2009，35(23)：118－119.

[7] 牟琦，毕儒孝，龚尚福，等. 基于中间分类超平面的 SVM 入侵检测[J]. 计算机工程，2011，37(16)：117－119.

[8] 阚媛，刘以安，薛潇，等. 结合 SVM 的交互式遗传算法在入侵检测中的应用[J]. 计算机工程与应用，2010，46(29)：200－202，210.

[9] 张雪芹，顾春华，吴吉义. 异常检测中支持向量机最优模型选择方法[J]. 电子科技大学学报，2011，40(4)：559－563.

[10] 杨宏宇，李春林. 采用 FA 和 SVDFRM 的 SVM 入侵检测分类模型[J]. 电子科技大学学报，2009，38(2)：240－244.

[11] 张昱，谢小鹏. 基于遗传相关向量机的图像分类技术[J]. 计算机仿真，2011，28(5)：283－286.

[12] TENENBAUM J B，de SILVA V，LANGFORD J C. A global geometric framework for nonlinear dimensionality reduction [J]. Science，2000，290(5500)：2319－2323.

[13] ZHENG K M，QIAN X，WANG P C. Dimension reduction in intrusion detection using manifold learning [C]// Proceedings of 2009 International Conference on Computational Intelligence and Security. Piscataway，NJ：IEEE Press，2009：464－468.

[14] 李刚，邢书宝. 基于 RBF 核的 SVM 和 RVM 性能模式分析性能比较[J]. 计算机应用研究，2009，26(5)：1782－1784.