文章编号: 1000-6788(2010)01-0126-05 中图分类号: TP393.0 文献标识码: A

基于改进 NN-SVM 算法的网络入侵检测

于秋玲

(河南省电力公司郑州供电公司信息中心, 郑州 450052)

摘 要 在网络入侵检测中, 引入类归属度对 NN-SVM 算法进行改进. 综合距离与同异类点个数因素, 通过计算样本点对最近 T 个样本点的类别归属程度来决定取舍, 以此对样本集进行修剪, 从而降低正反类的混淆程度, 以降低 SVM 的学习代价, 提高泛化能力. 试验表明: 与 SVM 算法相比, 改进的 NN-SVM 算法能有效地减少学习样本数, 解决小样本的机器学习问题, 提高系统检测性能.

关键词 入侵检测; 改进 NN-SVM; 类归属度

Internet intrusion detection system based on improved NN-SVM

YU Qiu-ling

(Information Center in Zhengzhou Branch of Henan Power Company, Zhengzhou 450052, China)

Abstract Introducing the Degree of Class Ownership would improve NN-SVM algorithm in internet intrusion detection. According to the distance and the number of the same class or the different class, calculating the degree of class ownership of the sample point to its T nearest neighbors decided whether the sample point should be reserved or deleted. Based on this, the improved NN-SVM algorithm pruned the training sample set to reduce the confusion degree of the positive and negative categories. As a result, it could effectively reduce the cost of the learning and improve the generalization. The experiment shows that the improve NN-SVM algorithm, contrasting to traditional SVM algorithm, can effectively reduce the size of the training sample. So the improved NN-SVM algorithm can solve the machine-learning problem with a small sample set and improve the performance of the system detecting.

Keywords intrusion detection; improved NN-SVM (Nearest Neighbor-Support Vector Machine); degree of class ownership

1 引言

随着计算机和网络技术应用的日益普及, 计算机网络安全形势也越来越严峻. 入侵检测作为网络安全研究的重要内容, 更是成为国内外学者的研究热点. 入侵检测 (Intrusion detection) 通过检查有关的审计数据, 来判断系统中是否有违背系统安全或安全策略的行为. 入侵检测技术中的异常检测方法能够检测出未知的攻击, 但是目前的异常检测方法在学习阶段, 都要求大量的、完备的训练样本集以期达到较理想的检测性能, 同时也意味着较长的系统训练时间. 但是, 在现实的网络环境中, 完备训练样本集是很难获取的. 于是, 入侵检测系统的研究就聚焦在如何实现在小样本的情况下获得较高的检测精度. 针对这个问题, 本文提出一种基于改进 NN-SVM 算法的网络入侵检测系统. 改进 NN-SVM 算法能首先进行样本集的有效修剪, 从而能有效减少所需评价样本的数量, 降低正反类的混淆程度, 同时缩短训练时间. 支持向量机 (SVM) 能较好地解决小样本学习问题, 同时具有很强的泛化能力 [1]. 将基于改进 NN-SVM 算法应用于网络入侵检测, 可以实现在小样本的情况下, 保证系统的分类精度, 从而达到提高训练速度和降低构建训练样本集代价的目的, 以提高整个入侵检测系统的性能.

收稿日期: 2008-10-06

作者简介: 于秋玲 (1979–), 女, 硕士, 工程师, 研究方向: 数据挖掘, 知识管理, 调度自动化, E-mail: yql_mail@yahoo.com.cn.

2 SVM 分类原理

支持向量机 (SVM) 是建立在统计学基础上的 VC 维理论和结构风险最小 (SRM) 原理基础上的[2], 是适用于小样本训练的大边缘分类器. 我们要寻找一个分类规则, 使其能对未知类别的新样本 (新样本与训练样本独立同分布) 做尽可能正确的划分. 支持向量机用于分类问题其实就是寻找一个最优分类超平面, 把此平面作为分类决策面, 它不但可以将给定的输入样本正确地划分为正常和异常两类, 而且使得被分成的两类数据间的分类间隔尽可能大[3].

对于数据为线性可分情况, 分类超平面的描述为

$$w \cdot x + b = 0 \tag{1}$$

式 (1) 中向量 w 为分类超平面的权系数, b 是分类阈值. 对其进行归一化, 使得满足

$$y_i \left[(w \cdot x_i) + b\right] - 1 \geq 0, \quad i = 1, 2, \cdots n \tag{2}$$

此时分类间隔等于 $2/\|w\|$, 使间隔最大等价于使 $\|w\|^2$ 最小. 满足 (2) 使 $\|w\|^2/2$ 最小的分类面就是最优分类面. 于是得到最优分类超平面的分类判别函数为

$$f(x) = \operatorname{sgn}\left[\sum_{i=1}^{n} \alpha_i y_i \left(x_i \cdot x\right) + b\right] \tag{3}$$

对于数据为非线性不可分的情况, 就要将其转化为线性可分的情况, 首先通过函数 ϕ 将输入空间 X 中的样本映射到高维特征空间 F, 使这些样本在高维空间内线性可分. 根据泛函的有关理论, 若核函数 $K(x, x_i)$ 满足 Mercer 条件, 它就对应某一变换空间中的内积 $\langle\varphi(x_i) \cdot \varphi(x)\rangle$, 函数 $\phi : X \rightarrow F$ 是一个从非线性输入空间 X 到高维特征空间 F 的映射, 所以求映射 $\phi : X \rightarrow F$ 只要知道如何由输入 x, x_i 计算内积 $\langle\varphi(x_i) \cdot \varphi(x)\rangle$ 即可, 由 $K\left(x_i \cdot x\right) = \varphi(x_i) \cdot \varphi(x)$ 将式 (3) 重写, 即可得到对应高维空间的分类函数为

$$f(x) = \operatorname{sgn}\left[\sum_{i=1}^{n} \alpha_i y_i K\left(x_i \cdot x\right) + b\right] \tag{4}$$

这样, 利用 Lagrange 优化方法可以把最优分类问题转化为其对偶问题, 分类函数类型为式 (4) 的学习机称为支持向量机[4−5].

3 改进的 NN-SVM

SVM 以较强的泛化能力著称, 但当两类训练样本集混叠情况比较严重时, SVM 往往由于过学习使其泛化能力降低, 现有的解决方法有 NN-SVM 或者 KNN-SVM 分类器等.

对 SVM 分类时错分样本的分布进行分析发现, 其出错样本大都位于分界面附近, 所以我们应该着眼于分界面附近的样本分类来提高分类性能. 由 SVM 理论知道, 支持向量机中支持向量都出现在两类样本集间隔以内的正确划分区, 分界面附近的样本基本上都是支持向量, 同时 SVM 可以看成每类只有一个代表点的最近邻 (Nearest neighbor, 即 NN) 分类器. 所以结合 SVM 和 NN, 对不同空间分布的样本使用不同的的分类法. 具体地, 当样本和 SVM 最优超平面的距离小于给定的阈值, 即样本离分界面较近, 则用 NN 分类, 反之用 SVM 对样本分类. 在使用 NN 时以每类的所有的支持向量作为代表点组, 这样增加的运算量很少, 实验证明了使用支持向量机结合最近邻的分类器分类比单独使用支持向量机分类具有更高的分类准确率[6].

然而, NN-SVM[7] 针对训练集中的每一个样本点找出其最近邻, 根据类标的异同来判断该样本点与其最近邻是否属于同一类, 若相同则保留该样本点, 否则就将该样本点删去. 这时如果一个同类点周围密集分布着大量的异类点, 在进行删减的时候很可能将这些异类点删去, 这样就会影响分类精确度. 所以仅仅选择一个最近样本点进行考察而决定取舍, 误删的可能性就会比较大. 而对于 KNN-SVM[8] 若一个同类点的周围 K 个点中大多数是异类点, 但是这些异类点离得都很远, 个数占少数的同类点却比较密集的分布在这个样本点周围, 此时 KNN 就不够精确. 所以仅仅考虑最近 K 个样本点中同异类样本点的个数多少来决定取舍, 没有考虑这些样本点与考察点的距离, 这种做法也有失偏颇.

改进算法为了解决上述问题, 就要在保证对训练样本高效地删减的基础上, 又要在两类训练样本集混叠严重时精确地进行分类, 使 SVM 保持良好的泛化性, 引入类归属度的计算从 T 个最近邻中同异类点个数和距离两个方面来衡量考察样本的类别归属情况.

针对每个样本 $x_i(m$ 维向量) 计算其与距离最近 T 个样本之间的距离, 设该样本到这 T 个样本的距离为 $D_1, D_2, \cdots, D_t$, 采用欧氏距作为两个向量之间的距离, 即

$$D(x_i, x_j) = \sqrt{\sum_{k=1}^{m} (x_i^k - x_j^k)^2} \tag{5}$$

而用 $1/D_i$ 来表示第 i 个点从距离因素上对考察样本的类别归属的影响因子.

若这 T 个样本中有 r 个与考察样本都是同一个类的 (假设距离为 $D_1, D_2, \cdots, D_r$), 而剩余的 $T-r$ 个与考察样本都不是同一个类的 (假设距离为 $D_{r+1}, D_{r+2}, \cdots, D_T$), 用类归属度综合距离影响因子与同异类个数来说明样本的类别归属程度, 类归属度为:

$$E_i = \left(\sum_{i=1}^{r} 1/D_i\right) \Big/ r - \left(\sum_{i=r+1}^{T} 1/D_i\right) \Big/ (T-r) \tag{6}$$

下面以程序的方式给出上述方法的实现算法:

给定一个训练集 $(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)$, $x_i \in R^n$, $y_i \in \{1, -1\}$, $i = 1, 2, \cdots, m$. 将训练集表示为矩阵 $TR_{m\times(n+1)} = [X\ Y]$, 其中 $X = [x_1, x_2, \cdots, x_m]^{\mathrm{T}}$, $Y = [y_1, y_2, \cdots, y_m]^{\mathrm{T}}$.

修剪算法如下:

1. 找到每个样本的 T 个最近邻;

1) 求出每个点与其它各点的距离, 与自身的距离定义为 ∞;

For $p = 1$ to m

{ $Z_{1\times m}$=(z_{ij}), z_{ij}=∞, $i = 1$;

$j = 1, 2, \cdots, m$;

For $q = 1$ to m

{if $q \neq p$, z_{1q} =$D(x_p, x_q)$;}

}

2) 找出 T 个最近邻;

$NN_{m\times 1}$ =(nn_{ij}), nn_{ij}=1,

$i = 1, 2, \cdots, m$; $j = 1$

$D_{T\times 1}$ =(d_{i1}), $i = 1, 2, \cdots, T$

$s = 1$; value=z_{11};

For $k = 1$ to T

{For $q = 1$ to m

{ if $z_{1q} <$ value {value=z_{1q}; $s = q$;}

nn_{p1} =s;}

d_{T1} =value;}

2. 判断样本的类归属度 E_i

$E_{p\times 1} = (e_{i1})$, $i = 1, 2, \cdots, n+1$

For $p = 1$ to T

{if $y_p \neq y_{nn}$, l_{p1}= -1; }

e_{i1}=0, $v1$=0, $v2$=0, $r = 0$;

For $j = 1$ to T

{ if l_{j1}=1

$v1 = (v1 + 1/z_{ij})$, r++;}

For $j = 1$ to T

{if l_{j1}= -1

$v2 = (v2 + 1/z_{ij})$;}

e_{i1}=$v1/r - v2/(T-r)$;

3. 将样本的类归属度 E_i 与设定阀值 ε 作比较, 删除类归属度低的向量

for $i = 1$ to m

{if $e_{i1} < \varepsilon$

删除矩阵 TR 及 L 的第 i 行, 新矩阵仍设为 TR 及 L.}

4 系统实现

系统逻辑结构如图 1 所示, 系统由数据采集、特征抽取、向量化处理、NN-SVM 训练、入侵检测、系统响应六个模块组成.

数据采集模块从网络数据流截包, 采用 Tcpdump 实现[9].

特征抽取模块对每一次的网络连接, 抽取 41 个特征: duration, protocol-type, service ··· hot, num-failed-logins, logged-in ··· count, srv-count, serror-rate ··· dst-host-count, dst-host-srv-count, dst-host-same-srv-rate ···

向量化处理模块将复杂的网络连接记录信息格式转换成 SVM 能够处理的向量形式, 表征测试数据

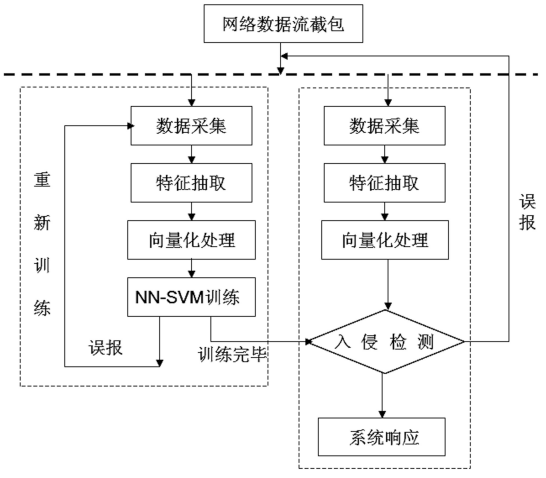

图 1 系统逻辑结构图

的结构. 对于值为 1 或 0 的逻辑型属性值和取值范围在 [0.0, 1.0] 内的连续数据, 不需进行处理. 处理的重点在于字符型的数据数值化, 每个字符属性包含不同个数的符号, 如 protocol_type 包含 3 个不同的符号、flag 包含 11 个不同的符号、service 包含 70 个不同的符号. 为每个符号编制一个标号, 将字符属性的不同的符号映射到 0 到 $N-1$ 之间 (N 即符号个数), 例如: 对于 protocol_type 属性把 tcp 标记为 0, udp 标记为 1, icmp 标记为 2. 然后再将这些标号数值映射到 [0.0, 1.0] 之间[10], 转换成二进制的形式. 对于数值类型的属性值, 取值范围非常大的情况, 如 src_bytes [0, 1.3 billion], 通过取以 10 为底的对数映射到 [0.0, 9.14] 之间, 然后再映射到 [0.0, 1.0] 之间; 取值范围在 [0.0, 58329.0] 整数范围内的, 直接映射到 [0.0, 1.0] 之间. 最终将所有属性值都规范到相同量级上, 这样就避免了取值范围大的属性支配取值范围小的属性[11].

NN-SVM 训练模块针对训练数据集, 首先采用改进 NN 算法 (类归属度) 对数据集进行删减, 然后采用 SVM 算法对删减后的数据集进行分类. 入侵检测模块针对检测数据集, 采用上述训练模块训练成熟的分类器进行数据分类. 系统响应模块针对入侵检测模块检测出的入侵数据调用系统策略进行处理.

系统的入侵检测分为两个阶段: 第一阶段进行训练, 训练数据抽取特征并向量化之后, 输入 NN-SVM 训练模块采用改进 NN-SVM 算法反复训练得到分类器; 第二阶段进行检测, 用第一阶段训练好的分类器对抽取特征并向量化的网络数据进行检测, 如果发现入侵, 就调用系统响应模块采取相应的处理策略, 另外在实际检测过程中如果发现误报, 则进行误差分析之后仍可返回重新训练分类器. 整个过程就是一个不断循环、不断完善的过程, 以达到更高的系统性能.

5　试验结果及分析

5.1　数据源

本实验采用 KDD'99 入侵检测评测数据[12] 对所设计的入侵检测模型进行测试. 在该数据集中大约有 5 亿条训练数据记录和 0.3 亿条测试数据记录. KDD'99 入侵检测数据集中含有四种类型的攻击行为:

1) DoS (Denial of Service): 拒绝服务攻击类型, 如 ping-of-death, SYN flood, land 等.

2) Probe: 各种端口扫描和漏洞扫描, 如 port-scan, ping-sweep 等.

3) R2L (Remote to Local): 远程非法登陆, 如 guessing password 等.

4) U2R (Unauthorized access to Root): 非授权超级用户存取, 如 buffer-overflow 攻击等.

5.2　参数选择

利用基于 MATLAB 的 SVM 工具箱以及编制的数据修剪程序进行实验, SVM 核函数使用高斯核: $\exp\left[-\gamma\ (x_i-x)^2\right]$, 其中 γ =0.5, 惩罚参数 C=100. 采用检测精度作为衡量系统性能的指标.

试验 1 参数 T 的选择

采用改进 NN-SVM 分类器进行试验, 针对编制修建程序取 $\varepsilon=-2$, 针对不同的 T 取值进行检测精度的试验.

试验结果显示如图 2, 当 T=7 时, 检测精度最高.

试验 2 参数 ε 的选择

采用改进 NN-SVM 分类器进行试验, 针对编制修建程序取 T=7, 针对不同的 ε 取值进行检测精度的试验. 试验结果显示如图 3, 当 $\varepsilon=-3$ 时, 检测精度最高.

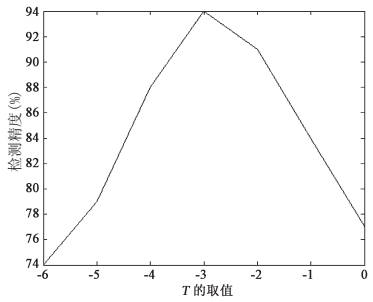

图 2　参数 T 的选择

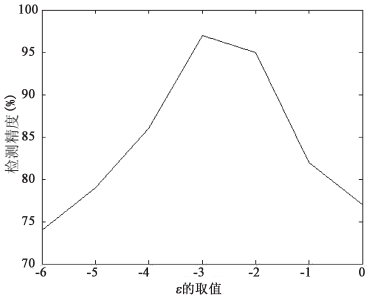

图 3　参数ε的选择

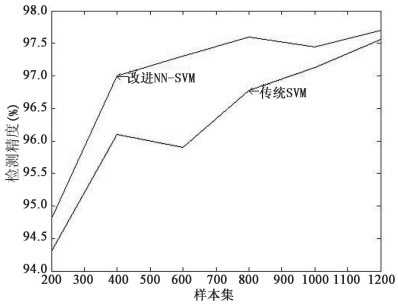

图 4　改进 NN-SVM 算法与传统 SVM 算法的检测精度比较

5.3 基于改进 NN-SVM 算法与传统 SVM 算法的比较

为了测试改进 NN-SVM 算法性能, 将其与传统 SVM 算法进行比较. 参数采用试验 1 与试验 2 结果, 数据集修剪程序中取 T=7, $\varepsilon = -3$. 候选样本集为 1200. 试验结果显示如图 4, 可以看出改进 NN-SVM 算法明显优于传统 SVM 算法, 使用改进 NN-SVM 算法, 当样本集等于 400 时, 系统检测精度就达到了 97% 以上, 而使用传统 SVM 算法, 当样本集逼近 1000 时, 系统检测精度才达到了 97% 以上. 由此可见, 改进 NN-SVM 算法可以有效地减轻正反类的混淆程度, 从而减少学习样本数, 提高系统检测速度和精度.

6 结语

本文提出了一种基于类归属度的 NN-SVM 算法应用于网络入侵检测系统, 其泛化性比传统 SVM 算法有明显提高, 提高了检测速度与精度. 解决了网络异常入侵检测中训练样本集构建代价较大且处理时间过长的问题, 可以大幅度地降低学习代价, 而且可以在小样本的情况下获得较高检测精度, 而且针对不同的训练集可以进行参数调节, 增强了系统的灵活性, 对于提高入侵检测系统的性能有较大研究意义.

参考文献

[1] Vapnik V N. 统计学习理论的本质 [M]. 张学工, 译. 北京: 清华大学出版社, 2000.
[2] Nello C, John S T. 支持向量机导论 [M]. 李国正, 王猛, 曾华军, 译. 北京: 电子工业出版社, 2004.
[3] Vapnik V N. An overview of statistical learning theory[J]. IEEE Transactions on Neura Networks, 1999, 10(5): 988–999.
[4] Vapnik V N. Statistical Learning Theory[M]. 2nd ed. New York: Springer Verlag, 1999.
[5] Klaus R M, Sebastian M, Gunnar R, et al. An introduction to kernel-based learning algorithms[J]. IEEE Transactions on Neural Networks, 2001, 12(2): 181–201.
[6] 李程雄, 丁月华, 文贵华. SVM-KNN 组和改进算法在专利文本分类中的应用 [J]. 计算机工程与应用, 2006, 20: 193–195.
Li C X, Ding Y H, Wen G H. Application of SVM-KNN combination improvement algorithm on patent text classification[J]. Computer Engineering and Application, 2006, 20: 193–195.
[7] 李红莲, 王春花, 袁保宗. 一种改进的支持向量机 NN-SVM[J]. 计算机学报, 2003, 26(8): 1015–1020.
Li H L, Wang C H, Yuan B Z. An improved SVM: NN-SVM[J]. Chinese Journal of Computers, 2003, 26(8): 1015–1020.
[8] 李蓉, 叶世伟, 史忠植. SVM-KNN 分类器 —— 一种提高 SVM 分类精度的新方法 [J]. 电子学报, 2005, 30(5): 745–748.
Li R, Ye S W, Shi Z Z. SVM-KNN classifier—A new method of improving the accuracy of SVM classifier[J]. Acta Electronica Sinica, 2005, 30(5): 745–748.
[9] 段丹青, 陈松乔, 杨卫平. 基于 SVM 主动学习的入侵检测系统 [J]. 计算机工程, 2007, 33(1): 153–155.
Duan D Q, Chen S Q, Yang W P. Intrusion detection system based on support vector machine active learning[J]. Computer Engineering, 2007, 33(1): 153–155.
[10] Sabhnani M, Serpen G. Application of machine learning algorithms to KDD intrusion detection dataset within misuse detection context[C]//International Conference on Machine Learning; Models, Technologies and Applications, (MLMTA), Las Vegas, Nevada, USA, 2005: 209–215.
[11] 张桂玲. 基于软计算理论的入侵检测技术研究 [D]. 天津: 天津大学, 2006.
Zhang G L. Research on intrusion detection based on soft computering theories[D]. Tianjin: Tianjin University, 2006.
[12] http: //www.kdd.ics.uci.edu/databases/kddcup99/kddcup99.htm.