

# 基于 TSVM 分类器和混合型特征选择方法的入侵检测研究

陈 鑫, 梁海洁, 廖腾峰

(广西工业职业技术学院, 广西南宁 5300031)

**摘 要:** 基于 TSVM 分类算法和混合型特征选择方法, 提出了一种网络入侵检测的新方法, 能够高效地检测网络入侵. 通过大量基于著名的 KDD Cup 1999 数据集的实验, 表明其相对于传统的入侵检测方法在保证较高检测率的前提下, 有效地降低了误报率.

**关键词:** 入侵检测; TSVM; 混合型特征选择; 信息增益

**中图分类号:** TP309

**文献标识码:** A

**文章编号:** 1000-7180(2010)08-0242-03

## Intrusion Detection Based on TSVM and Feature Selection

CHEN Xin, LIANG Hai-jie, LIAO Teng-feng

(Guangxi Vocational & Technical Institute of Industry, Nanning 530031, China)

**Abstract:** In this paper, we propose a new network intrusion detection method based on TSVM (Transductive Support Vector Machines) classification algorithm and hybrid feature selection approach. It can effectively detect anomalies with high detection rate and low false positives. Experimental results on the well-known KDD Cup 1999 dataset demonstrate the proposed method is robust and more effective than the state-of-the-art intrusion detection method.

**Key words:** intrusion detection; TSVM; hybrid feature selection; information gain

### 1 引言

入侵检测系统是网络安全防御体系的一个重要组成部分<sup>[1]</sup>. 传统的入侵检测方法大多基于数据挖掘及机器学习方法, 并且, 他们都可以转化为相应的分类问题来解决. 如: 基于 SVM 算法的入侵检测模型<sup>[2]</sup>、基于多分类支持向量机的网络入侵检测方法<sup>[3-4]</sup>等. 如何在现实环境下保证入侵检测的效率(高检测率和低误报率), 成为当前入侵检测领域的经典难题. 文中提出了一种入侵检测新方法, 它基于 TSVM(Transductive Support Vector Machines)算法和特征选择机制. 通过一系列基于著名的 KDD Cup 1999 数据集的实验测试, 验证了其有效性.

### 2 TSVM 算法原理

支持向量机(SVM)近年来受到了广泛的关注, 它在多种分类问题表现出了优异的推广性能, 其基

本思想是基于统计学习理论的结构风险最小化<sup>[5]</sup>.

TSVM 是 Joachims 等对成熟的 SVM 文本分类方法进行进一步改进形成的一种经典的直推式学习方法<sup>[6]</sup>, 特别适合小样本训练集<sup>[7]</sup>. 这种方法在学习训练过程中, 把测试集也一起进行考虑, 然后尽量减少测试集(未标记集合)的分类错误. 通过研究这些测试集的位置, 获得信息, 建立一个基于训练集和测试集的分类间隔的结构.

TSVM 的算法原理如下<sup>[8]</sup>: 给定一组独立同分布的有标签训练样本点:  $(x_1, y_1), \dots, (x_n, y_n)$ ,  $x_i \in R^m$ ,  $y_i \in \{-1, +1\}$  和另一组来自同一分布的无标签样本点  $x_1^*, x_2^*, \dots, x_k^*$ . 在一般的线性不可分条件下, T. Joachims 的直推式向量机的训练过程可以描述为如式(1)所示的优化问题. 其中参数  $C$  和  $C^*$  为用户指定和调节的参数. 参数  $C^*$  称为无标签样本在训练过程中的影响因子,  $C^* \xi_j^*$  称为无标签样本  $j$  在目标函数中的影响项.

对于文中所考虑的人侵检测过程,训练算法大致可以分为以下几个步骤:

(1) 指定参数  $C$  和  $C^*$ , 使用归纳式学习对有标签样本进行一次初始学习, 得到一个初始分类器. 并按照某个规则指定一个无标签样本中的正标签样本数.

(2) 用初始分类器对无标签样本进行分类, 根据对每一个无标签样本的判别函数输出, 对输出值最大的  $N$  个无标签样本暂时赋正标签值, 其余的赋负标签值. 并指定一个临时影响因子  $C_{tmp}^*$ .

(3) 对所有样本重新训练, 对新得到的分类器, 按一定的规则交换一对标签值不同的测试样本的标签符号, 使得优化问题中的目标函数值获得最大下降. 这一步骤反复执行, 直到找不出符合交换条件的样本对为止.

(4) 均匀增加临时影响因子  $C_{tmp}^*$  的值并返回到步骤(3), 当  $C_{tmp}^* \geq C^*$  时, 算法结束, 并输出结果.

Minimize over  $(y_1^*, \dots, y_k^*, w, b, \xi_1, \dots, \xi_n, \xi_1^*, \dots, \xi_k^*)$  (1)

$$\begin{aligned} & \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i + C^* \sum_{j=1}^k \xi_j^* \\ \text{s.t. : } & \forall_{i=1}^n: y_i [w \cdot x_i + b] \geq 1 - \xi_i \\ & \forall_{j=1}^k: y_j [w \cdot x_j^* + b] \geq 1 - \xi_j^* \\ & \forall_{i=1}^n: \xi_i \geq 0 \\ & \forall_{j=1}^k: \xi_j^* \geq 0 \end{aligned}$$

### 3 混合型特征选择方法

特征选择是用于消除冗余与无关特征的重要分类预处理步骤. 利用特征选择技术不仅能够提高计算性能, 而且可以消减冗余和无关特征的干扰, 提高分类准确性. 文中提出了一种混合型特征选择算法, 该算法基于信息增益与 TSVM 分类算法.

#### 3.1 信息增益

信息增益 (information gain, IG) 又称为互信息, 它可以用来衡量某个特征含有的信息量大小. 特征的信息增益越大, 代表其包含的分类信息量也越多, 对分类的作用也越大. 对于特征  $A$ , 它的信息增益计算公式<sup>[9]</sup>为:

$$\text{Gain}(A) = I(s_1, s_2, \dots, s_m) E(A) \quad (2)$$

其中,  $I(s_1, s_2, \dots, s_m)$  是样本的总信息熵, 其定义为

$$I(s_1, s_2, \dots, s_m) = - \sum_{i=1}^m P(C_i) \log_2 P(C_i) \quad (3)$$

式中,  $P(C_i)$  是任意样本属于  $C_i$  的概率,  $P(C_i) =$

$s_i/s$ ;  $m$  表示样本类别数;  $s_i$  是属于类  $C_i$  的样本数;  $s$  是总的样本数. 式(2)中的  $E(A)$  是特征的信息熵, 其定义如下:

$$E(A) = \sum_{j=1}^v \frac{s_{1j} + s_{2j} + \dots + s_{mj}}{s} I(s_{1j}, s_{2j}, \dots, s_{mj}) \quad (4)$$

通过计算信息增益, 能够对特征在分类中的重要性进行评估排名.

#### 3.2 混合型特征选择方法

为了降低搜索的盲目性, 文中基于上述策略提出了一种变异的混合型特征选择算法. 一般混合型特征算法可以分为两个阶段: 首先根据数据自身的特性使用过滤型特征选择算法选取候选特征子集; 然后在候选特征子集中使用特定的学习算法进行交叉验证, 选取优化特征子集. 文中所提出的 IG-TSVM 算法与传统的混合型特征选择算法不同, 该算法在第一阶段不是利用过滤型算法产生候选特征子集, 而是通过计算信息增益对所有特征进行排序, 在第二阶段使用贪心前向搜索法 (greedy forward search, GFS) 作为搜索策略, TSVM 分类器作为评估器进行封装型特征选择. 为了方便表示, 使用  $IG_{i,c}$  表示特征  $A_i$  和类别  $C$  之间的信息增益, 算法的具体过程如下:

(1) 利用式(2)计算每个特征与类别的信息增益  $IG_{i,c}$ , 根据信息增益值从大到小对特征进行排序.

(2) 特征排序完成后, 进入封装型特征选择循环. 在循环中, 使用 GFS 作为搜索策略产生特征子集, TSVM 分类器作评估器, 选取优化特征子集.

算法使用 TSVM 分类器 5 折交叉验证获得的分类错误率均方差与分类错误率均值的比值作为特征选择循环的终止条件, 该值的计算方法如下:

$$f_{\text{error}} = \frac{\text{MSE}(R)}{\text{MEAN}(R)} \quad (5)$$

其中,  $R$  是 5 折交叉验证获取的 5 次分类错误率,  $\text{MSE}(R)$  用于计算分类错误率的均方差,  $\text{MEAN}(R)$  用于计算分类错误率的均值. 当  $f_{\text{error}}$  值较小时, 表明交叉验证的分类错误率接近, 分类结果稳定. 设定阈值  $\epsilon$ , 当该值小于  $\epsilon$  时, 终止特征选择循环. 文中经过多次实验设定  $\epsilon$  值为 1%.

基于 IG-TSVM 的人侵检测利用 IG-TSVM 选出优化特征子集, 并在此基础上建立 TSVM 入侵检测器. 它的详细流程如下所示:

(1) 将扩展特征集合和训练集作为输入, 计算

特征的信息增益,选取排名前  $\delta$  内的特征构成新的特征集合;

(2) 然后进入封装型特征选择的迭代循环,在每一次循环过程中,使用 GFS 搜索策略产生新的特征子集  $S$ ,以 TSVM 分类算法作评估器,通过 5 折交叉验证计算特征子集的分类正确率作为性能评估值  $F(S)$ ,比较  $F(S)$  与  $F(S_{\text{best}})$ ,如果  $F(S)$  小于  $F(S_{\text{best}})$ ,则令  $F(S_{\text{best}}) = F(S)$ ,  $S_{\text{best}} = S$ . 当特征子集的评估值达到了预先的标准或者最大迭代次数已经达到,就终止循环.

(3) 最后,在特征子集  $S_{\text{best}}$  上建立 TSVM 分类器,并将它应用于入侵检测.

#### 4 实验结果及讨论

实验采用基准评测数据集 KDD Cup 1999 数据集<sup>[10]</sup>进行测试.在实验中,首先将文中所述方法与著名的人工神经网络(ANN)方法、SVM(SVM)方法和 K 近邻(KNN)方法的检测效果进行了比较.实验所用的比较算法均采用机器学习领域著名的 WEKA 软件<sup>[11]</sup>进行实验,IG-TSVM 算法则采用 SVMlight 实现的版本<sup>[12]</sup>进行修改实验.

实验前,从数据集中随机提取了 31652 条正常数据(标记为“0”)和 7510 条攻击数据(标记为“1”).采用了十折交叉验证(ten-fold cross validation)的方法,取检测率(TP)和误报率(FP)的平均值对几个方法进行了对比.实验结果如表 1 所示.结果显示,TSVM 方法相比包括 SVM 在内的其他三种方法,在小样本环境下的检测效果要好得多,具体体现在其检测率高,且误报率很低,而其他方法检测率下降,误报率明显上升.

表 1 TSVM 实验比较结果

|      | TP (%) | FP (%) |
|------|--------|--------|
| SVM  | 98.3   | 1.4    |
| ANN  | 98.5   | 1.6    |
| KNN  | 96.7   | 3.6    |
| TSVM | 99.8   | 0.4    |

另外,对特征选择进行了测试.通过 IG-TSVM 特征选择后,进行 Rank 排序保留下来的 6 个重要特征为:dst\_host\_error\_rate,src\_bytes,dst\_bytes,count,num\_compromised,dst\_host\_srv\_error\_rate.表 2 给出了特征选取前后 TSVM 方法检测率和误报率的比较结果.结果表明:通过特征

选择的降维处理后,TSVM 方法的检测率和误报率都与未处理前基本一致(检测率有非常微弱的下降,但仍然是可以接受的),充分说明了该方法用于入侵检测方法的可扩展性很强.在实践中可以通过相应的优化工作来减少计算量,而保证性能没有减退.

表 2 经过特征选择后的实验结果

|        | 原始数据集 | 混合特征选择后数据集 |
|--------|-------|------------|
| TP (%) | 99.8% | 99.7%      |
| FP (%) | 0.2%  | 0.1%       |

#### 5 结束语

文中提出了一种基于 TSVM 分类方法和混合型特征选择方法的入侵检测技术.通过对比实验表明其相对于传统的入侵检测方法来说,具有检测率高、误报率低的优点.我们下一步工作主要是将该方法应用于实际的入侵检测中,通过在骨干级的网络上收集真实的网络数据使用 TSVM 进行训练和测试,并调试其相应的参数,并根据实际情况不断优化其检测性能.

#### 参考文献:

- [1] Bykova M, Ostermann S, Tjaden B. Detecting network intrusions via a statistical analysis of network packet characteristics[C]// Proc. of the 33rd Southeastern Symp. on System Theory. USA: Columbia, 2001: 309-314.
- [2] 饶鲜,董春曦,杨绍全.基于支持向量机的入侵检测系统[J].软件学报,2003,14(4):798-803.
- [3] 李昆仑,黄厚宽,田盛丰,等.模糊多类支持向量机及其在入侵检测中的应用[J].计算机学报,2005,28(2):274-280.
- [4] 肖云,韩崇昭,郑庆华,等.一种基于多分类支持向量机的网络入侵检测方法[J].西安交通大学学报,2005,39(6):562-565.
- [5] Angiulli F, Pizzuti C. Outlier mining in large high-dimensional data sets[J]. IEEE Transactions on Knowledge and Data Engineering, 2005,17(2):203-215.
- [6] Vapnik V. Statistical learning theory[M]. New York: Wiley, 1998.
- [7] Joachims T. Transductive inference for text classifications using support vector machines[C]// Proceeding of the 16th international conference on machine learning. San Francisco: Morgan Kaufmann, 1999: 200-209.
- [8] 陈毅松,汪国平,董士海.基于支持向量机的渐进直推式分类学习算法[J].软件学报,2003,14(3):451-460.

(下转第 247 页)

应客户左移要求,转 Step4;否则,转 Step2. 这里整个地图格元划分的右边缘列号为 6,  $A0(1,3)$  为 5.

Step2: 计算:  $A1 = A0 \times \text{left\_B}$

计算结果说明:  $A0$  矩阵从第 2 列到原最右第 3 列,依次向左移动 1 列.  $A0$  原第 1 列置换  $A0$  最右的第 3 列.

Step3:  $A2$  最右的第 2 列值加 1,置换原最右的第 3 列:  $A2(i,3) = A1(i,2) + 1$ ;

式中  $i$  为行号,  $1 \leq i \leq 4$ .

Step4: 结束.

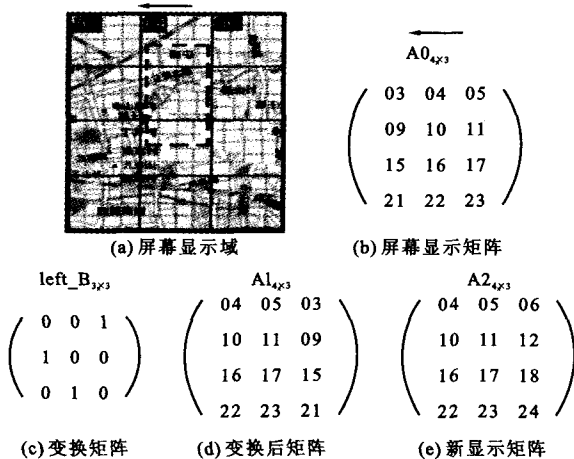


图 4 屏幕显示矩阵的变换

由上述算法可见,任意屏幕显示矩阵  $A_{m \times n}$  的移动变换矩阵与屏幕显示矩阵  $A_{m \times n}$  的行  $m$ 、列  $n$  有关,与移动方向(上、下、左、右)有关. 对任意一个屏幕显示矩阵  $A_{m \times n}$ ,其左、右移动变换矩阵是  $n \times n$  阶的方阵( $B_{n \times n}$ ),其上、下移动变换矩阵是  $m \times m$  阶的方阵( $B_{m \times m}$ ).  $B_{n \times n}$  和  $B_{m \times m}$  可以自动生成.

## 5 工具

基于 GIS 的应用一般涉及空间对象的数字化,

为了方便开发和使用, GIS 引擎提供基本对象的矢量化及空间数据库管理是必要的. 基本对象是指点、线、多边形.

为了规范一个地理管辖区域点阵图格元模型, GIS 引擎提供生成点阵图格元模型的工具.

## 6 结束语

以 GIS 为背景的点阵格元模型表达空间应用对象静态、动态位置是许多移动计算的共同需求. 基于这类 GIS 图形的应用要面临不同管辖区域的要求、不同移动终端的要求,图形移动要求.

点阵图划分为格元矩阵,以格元的行列定位格元所属的地理对象,以格元组织屏幕显示矩阵可以提高运行效率. 格元大小的选择、屏幕显示域的初始化、移动算法、规范的 GIS 点阵图结构是 GIS 应用的基础.

以 GIS 组件为基础,按应用需要,进一步扩展必须应用图层. 不同的图层对应不同的应用. GIS 组件支持地理对象矢量化也是必不可少的功能.

## 参考文献:

- [1] 梁毅,肖迎元,黄静,等. 面向移动设备的基于 Web Service 的图像处理技术[J]. 计算机工程与应用,2009,45(35):62-62.
- [2] 熊庆文,边馥苓. 基于嵌入式数据库系统的移动 GIS 应用体系结构研究[J]. 武汉大学学报,2006,31(1):86-89.
- [3] 杨海东,向华. 栅格地图技术与 VML 融合的 WebGIS 研究[J]. 计算机工程与应用,2009,45(13):68-71.
- [4] 王开义. GIS 领域最短路径搜索问题的一种高效实现[J]. 中国图象图形学报,2003(8):951-956.

## 作者简介:

姜代红 女,(1969-),硕士,副教授. 研究方向为数据库技术、移动计算.

(上接第 244 页)

- [9] Lei Yu, Huan Liu. Efficient feature selection via analysis of relevance and redundancy [J]. Journal of Machine Learning Research, 2004(5):1205-1224.
- [10] Knowledge discovery in databases DARPA archive. Task description [EB/OL]. [2009-12-11]. <http://www.kdd.ics.uci.edu/databases/kddcup99/task.html>.
- [11] Weka machine learning project [EB/OL]. [2009-12-20]. <http://www.cs.waikato.ac.nz/~ml/>
- [12] Joachims T. Support vector machine, SVMlight [EB/OL]. [2009-12-24]. [http://www-ai.cs.uni-dort-](http://www-ai.cs.uni-dortmund.de/SOFTWARE/SVM-LIGHT/svm-light_eng.html)

[mund. de/SOFTWARE/SVM-LIGHT/svm-light. eng. html](http://www-ai.cs.uni-dortmund.de/SOFTWARE/SVM-LIGHT/svm-light_eng.html).

## 作者简介:

陈鑫 男,(1968-),讲师. 研究方向为多媒体应用、图形图像处理、计算机网络.

梁海洁 男,(1974-),讲师. 研究方向为多媒体应用、网站制作.

廖腾峰 男,(1980-),助教. 研究方向为图形图像处理、网页设计制作.