文章编号:1671-251X(2013)02-0062-04
池静,杨振宇,张婷. 基于贝叶斯和决策树的入侵检测方法[J]. 工矿自动化,2013,39(2):62-65.

基于贝叶斯和决策树的入侵检测方法

池静, 杨振宇, 张婷

(河北工程大学 信息与电气工程学院, 河北 邯郸 056038)

摘要:针对目前基于贝叶斯或决策树的入侵检测方法存在检测率低、误检率高的问题,提出了一种基于贝叶斯和决策树的入侵检测方法。该检测方法首先采用基于特征相似度的朴素贝叶斯方法对训练集中的样本进行分类,更新每个样本的类值;然后对训练集中的样本再次使用朴素贝叶斯方法进行分类,对存在误分类样本的类采用决策树的信息增益来确定属性划分子类,再对子类进行分类和划分操作;最后建立贝叶斯和决策树的混合模型进行入侵检测。实验结果表明,与单独使用贝叶斯或者决策树的检测方法相比,该检测方法具有较高的检测率。

关键词:入侵检测; 朴素贝叶斯; 决策树; 相似度; 信息增益
中图分类号:TD67 文献标志码:A 网络出版时间:
网络出版地址:

Intrusion detection method based on Bayesian and decision tree

CHI Jing, YANG Zhen-yu, ZHANG Ting
(School of Information and Electronic Engineering, Hebei University of Engineering, Handan 056038, China)

Abstract: In view of problems of low detection rate and high false detection rate in intrusion detection method based on Bayesian or decision tree, the paper proposed an intrusion detection method based on Bayesian and decision tree. Firstly, Naive Bayesian method based on feature similarity is used to classify samples of training set and class value of each sample is updated. Then the samples are classified with Naive Bayesian method again. Those classes which contain misclassified samples are divided into some subclasses by using the attributes which were determined by information gain of decision tree, and the subclasses are operated on classification and partition. Finally, the model of combination of Bayesian and decision tree is used for intrusion detection. The experiment result showed that the detection method has higher detection rate than the method using either Bayesian or decision tree.

Key words: intrusion detection; Naive Bayesian; decision tree; similarity; information gain

0 引言

随着计算机网络的快速发展,尤其是开放性的计算机网络,越来越多的计算机系统遭受着入侵威胁。因此,网络安全系统日益受到人们的重视和研究[1]。入侵是指那些试图绕开计算机系统安全机制的活动[2]。目前,主要的入侵形式有攻击者通过互联网或者内部访问系统、得到官方授权的用户尝试获得和滥用非授权的权限。因此,入侵可以被认为是威胁网络资源安全性、可用性和机密性的一系列活动。

入侵检测(Intrusion Detection)系统是防火墙之后的第二道安全防线,它能够在发现可疑传输时发出警报或者采取主动反应措施,全方位检测网络攻击,是一种主动防御技术[3]。与防火墙相比,它是一种积极主动的安全防护技术。入侵检测技术主要

收稿日期:2012-11-21。
作者简介:池静(1973-),女,河北邯郸人,副教授,主要研究方向为数据挖掘,E-mail:firewater9@163.com。

有异常入侵检测和误用入侵检测 2 种类型。异常入侵检测是指使用计算机资源判断异常行为进行检测入侵,它尝试用定量的方式描述那些可以接受的行为特征,以区分正常的和非正常的入侵行为。误用入侵检测通过预先定义的入侵行为监视系统的运行,从中找出符合预先定义规则的入侵行为。

入侵检测使用数据挖掘方法,主要有决策树(DT)、朴素贝叶斯(NB)[4]、神经网络(NN)、支持向量机(SVM)[5]、k 临近(KNN)、模糊逻辑模型[6]。目前的入侵检测系统存在许多问题,比如低检测率,不同类型攻击的检测率差别较大,误检率较高。入侵检测系统的另一个困难是在实时高速网络中进行检测,这是因为高速网络要求入侵检测系统在非常短的时间内处理海量数据。另外,由于有不相关和冗余属性的存在,有必要选择有效的属性。因为不相关和冗余属性会使检测过程复杂化,同时也会降低检测率。一些额外属性也可能增加运算时间,影响检测率。

本文提出一种基于贝叶斯和决策树的入侵检测方法。该方法可以处理连续属性和缺失的属性值,降低训练数据中的噪声,与单独使用朴素贝叶斯或决策树的入侵检测系统相比,具有较高的检测率。

1　朴素贝叶斯方法

贝叶斯理论就是依据新的信息从先验概率求得后验概率的一种方法,其关键是利用概率表示各种形式的不确定性[7]。朴素贝叶斯分类器是一种经常用到的分类器。

假设数据样本是一个有 n 维特征的向量 $\boldsymbol{X}=\{x_1,x_2,\cdots,x_n\}$,有 n 个属性分别为 $A_1,A_2,\cdots,A_n$,属性值分别为 $b_1,b_2,\cdots,b_n$,m 个类分别为 $C_1,C_2,\cdots,C_m$。若有一个未知样本 B,朴素贝叶斯分类将预测 B 属于最高后验概率的那个类。

根据贝叶斯公式:$P(C_j\mid B)=\dfrac{P(B\mid C_j)P(C_j)}{P(B)}$,先假设 $P(C_1)=P(C_2)=\cdots=P(C_m)$,再假设各个属性间相互独立,根据式(1)得出预测样本 B 所属类别为

$$C_{nb}=\arg\max_{C\in C}\ P(C_i\mid b_1,b_2,\cdots,b_n)=\arg\max_{C\in C}\ P(C_i)\prod_{i=1}^{n}P(b_i\mid C_i)\qquad(1)$$

式中:$P(C_i)$ 为类 C_i 的先验概率;$P(b_i\mid C_i)$ 为条件概率。

通常情况,使用单纯的贝叶斯方法没有考虑属性对检测结果的影响,因此,会在一定程度上出现分类误差[8]。为了解决这一问题,对训练集在训练之前先进行特征选择。常用的特征选择方法有信息增益[9]、卡方检验等,然而这些方法并不能全面衡量各个特征属性对分类意义的影响大小。本文使用基于特征相似度的特征提取方法来降低数据维数,提高分类效率。

2　决策树方法

决策树[10]是数据挖掘中的一种分类方法,它是一个预测模型,代表的是对象属性与对象值之间的一种映射关系。使用决策树进行决策的过程就是从根节点开始,测试待分类项中相应的特征属性,并按照其值选择输出分支,直到到达叶子节点,将叶子节点存放的类别作为决策结果。决策树的核心是在各级节点上选择属性时,通过计算信息增益来选择属性。具体步骤如下。

假设有一个集合 D,m 个不同类别 C_j,$\mathrm{freq}(C_j,D)$ 为类 C_j 在 D 中出现的次数,则对一个给定样本集合 F 的信息熵的期望为

$$\mathrm{Info}(D)=-\sum_{j=1}^{m}\frac{\mathrm{freq}(C_j,D)}{|D|}\log_2\left(\frac{\mathrm{freq}(C_j,D)}{|D|}\right)\qquad(2)$$

令属性 A 将决策树分为 n 个子集 $\{T_1,T_2,\cdots,T_n\}$,对于一个子集 T_i,其样本集的期望是 $\mathrm{Info}(T_i)$。则属性 A 划分当前样本集合期望的信息熵为

$$\mathrm{Info}(T)=\sum_{i=1}^{n}\frac{|T_i|}{|T|}\mathrm{Info}(T_i)\qquad(3)$$

式中:$|T_i|$ 为 T_i 的样本个数;$|T|$ 为划分之前的样本总数。

由式(2)和式(3)得出信息增益值为

$$\mathrm{InformationGain}(A_i)=\mathrm{Info}(D)-\mathrm{Info}(T)\qquad(4)$$

ID3 算法作为决策树分类最经典的算法,其最大的不足是分裂属性选取时的多值偏向问题,而现实中属性值数量与属性重要性之间并无必然联系,因此,通常会出现归纳出不正确知识的问题[8]。

3　基于贝叶斯和决策树的入侵检测方法

3.1　数据预处理

基于贝叶斯和决策树的入侵检测方法首先遍历整个训练集,找出是否有多个拷贝的样本,如果发现,只保留 1 个(假设如果 2 个样本的所有属性值相等,那么这 2 个样本相等)。然后,将训练集中连续

的属性离散化。

3.2 基于相似度的特征提取

相似度是指 2 个对象的相似程度，它是数学的一个概念。概率统计中的相关系数经常用来刻画 2 个随机变量的分布相似程度。

记 $\rho_{xy}=\frac{\mathrm{Cov}(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}}$，为随机变量 X 与 Y 的线性相关系数，简称相关系数。其中 $D(X)$、$D(Y)$ 为 X、Y 的方差，$\mathrm{Cov}(X,Y)$ 为 X 和 Y 的协方差。ρ_{xy} 越大，表明变量的相似度越高。本文使用 ρ_{xy} 来计算每个攻击类型中属性之间的相似度。删除其他与该属性相似度高的属性，得到一个新的属性集合 $B=\{b_1,b_2,\cdots,b_k\}$。具体步骤：计算与属性 A_i 之间的相似度，选择低于阈值 0.95 的属性对，然后计算两属性的方差，选择方差小的属性；如果两者方差相同，任取其一。这样可以得到每个类型数据的新属性集，用以计算每个类别中属性的条件概率。然后再使用 $P(C_i\mid e)=\arg\max P(C_i)\prod_{j=1}^{n}\rho_j P(x_j\mid C_i)$ 进行分类操作。上式中：e 为样本；x_j 为属性 b_j 的值。

3.3 分类器训练

对于一个训练集 D，包含属性 $\{A_1,A_2,\cdots,A_n\}$。每一个属性 A_i 包括以下属性值 $\{A_{i1},A_{i2},\cdots,A_{im}\}$。这些属性值可以是离散的或者连续的。训练集包含一些类集 C，$C=\{C_1,C_2,\cdots,C_m\}$，训练集中的每个样本有一个专属的类 C_j。对经过数据预处理的训练集，提取每一类的特征，计算先验概率 $P(C_j)$ 和条件概率 $P(A_{ij}\mid C_j)$。然后，假设样本 e_i 的各个属性相互独立，使用朴素贝叶斯公式 $P(e_i\mid C_j)=\prod_{k=1\to p}P(A_{ij}\mid C_j)$ 对训练集中所有的例子进行分类。估计样本 e_i 在每个类的概率，取具有最大的后验概率的类为该样本的归属类。对所有的样本分类完毕后，更新每个样本的所属类值。使用更新后的类值重新计算先验概率和条件概率，再次对训练集的每个样本分类。如果存在误分类的样本，则计算样本集的每个属性的信息增益。选择训练集中信息增益最大的属性 A_i，依据 A_i 的属性值将训练集分成子数据集。计算每一个子数据集先验概率和条件概率，并且使用它们各自的概率对该子数据集中的样本进行分类。如果有任何样本或者任何子集被错误划分，那么计算子集属性的信息增益，在子集中选择有着最大信息增益的属性，使用该属性将子集再划分成子子集。然后为每一个子子集重新计算先验概率和条件概率，并且使用它们的概率分类每一个子子集。重复这个过程直到所有的样本都被正确划分。检测方法终止后，每一个数据集的先验概率和条件概率被保存起来，用来对测试样本进行分类。基于贝叶斯和决策树的入侵检测方法流程如图 1 所示。

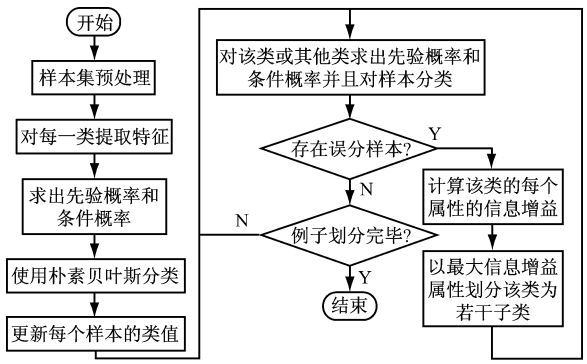

图 1 基于贝叶斯和决策树的入侵检测方法流程

4 实验结果及分析

使用如下环境验证本文提出方法的检测效果：系统 Windows7，CPU(Core i5 2.3 GHz)，内存为 2 GB，编程工具为 Matlab R2010a，入侵检测数据集使用 KDD99 数据集。KDD99 数据集主要有 5 大类：1 个正常类(Normal)和 4 个主要攻击类，即 DOS、U2R、R2L、Probe。其中 DOS 选取 ping-of-death，syn _ flood，smurf；U2R 选取 buffer _ overflow，xterm；R2L 选取 guess _ password，imap，snmpguess；Probe 选取 port-scan，ping-sweep，saint，mscan，ntifoscan。实验结果如表 1 所示。

表 1 各检测方法的检测率比较 %

检测方法	Normal	DOS	U2R	R2L	Probe
朴素贝叶斯	90.27	90.10	85.26	87.50	89.99
ID3 决策树	92.66	89.63	77.19	81.31	88.56
本文方法	97.85	96.80	88.52	91.64	93.82

本文提出的入侵检测方法将朴素贝叶斯和决策树结合起来，对数据样本首先进行相似度特征提取，在不影响样本分类的情况下，保留数据样本中的关键属性，消除数据冗余。从表 1 可看出，与传统的朴素贝叶斯和决策树单独使用时比较，本文提出的入侵检测方法的检测率有明显提升。

5 结语

基于贝叶斯和决策树的入侵检测方法很好地降

低了数据冗余和维数,有效减少了数据的计算量。同时使用 2 次朴素贝叶斯分类来确定存在二义性的分类,然后使用决策树的信息增益来确定属性划分子类,从而建立了基于朴素贝叶斯和决策树的分类模型。实验结果验证了该方法的可行性,具有很好的检测效果。

参考文献:

[1] 马晓春. 数据挖掘技术在网络入侵检测系统中的研究和应用[D]. 西安:西北工业大学,2005.

[2] LIPPMANN R, HAINES J W, FRIED D J, et al. 1999 DARPA off-line intrusion detection evaluation [J]. Computer Networks, 2000, 34(4): 579-595.

[3] BARBARÁ D, COUTO J, JAJODIA S. ADAM: A method for exploring the use of data mining in intrusion detection [J]. SIGMOD, 2001, 30(4): 15-24.

[4] AMOR, BENFERHAT S, ELOUEDI Z, et al. Decision trees in intrusion detection systems [C]// Proceeding of 2004 ACM Symposium on Applied Computing, 2004.

[5] MUKKAMALA S, JANOSKI G, SUNG A H. Intrusion detection using neural networks and support vector machines [C]// Proceeding of the IEEE International Joint Conference on Neural Networks, 2002.

[6] LUO J, BRIDGES S M. Mining fuzzy association rules and fuzzy frequency episodes for intrusion detection [J] International Journal of Intelligent Systems, 2000, 15(8): 687-703.

[7] MICHALSKI R S, CARBONELL J, MITCHE L T. Machine Learning: An artificial intelligence approach [M]. San Mateo: Morgan Kaufmann, 1983: 463-482.

[8] 黄宇达,王迤冉. 基于朴素贝叶斯与 ID3 算法的决策树分类[J]. 计算机工程,2012,38(14):41-43,47.

[9] 何慧,苏一丹,覃华. 基于信息增益的贝叶斯入侵检测模型优化的研究[J]. 计算机工程与科学,2006,28(6):38-40.

[10] QUINLAN J R. Induction of decision trees [J]. Machine Learning, 1986, 1(1): 81-106.