改进的决策树算法在入侵检测中的应用

马志远,曹宝香
(曲阜师范大学 计算机科学学院,山东 日照 276826)

摘 要:为了提高入侵检测系统对入侵行为的速度和检测率,需要引入更好的算法或者对现有的算法进行改进。入侵检测要求能够快速准确地检测出各种入侵行为,因此对算法的执行效率问题要求较高。文中介绍了决策树中的两个经典算法:ID3 算法和 C4.5 算法,分析了它们存在的问题以及寻找如何将改进的决策树算法应用在入侵检测中,并把它们进行了适当的改进以得到更好的效果。通过实验仿真验证了改进的这两种算法在入侵检测系统中对于发现入侵行为能够达到预期的结果。

关键词:入侵检测;决策树算法;入侵行为
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2014)01-0151-04
doi:10.3969/j.issn.1673-629X.2014.01.039

Application of Improved Decision Tree Algorithm in Intrusion Detection System

MA Zhi-yuan, CAO Bao-xiang
(College of Computer Science, Qufu Normal University, Rizhao 276826, China)

Abstract: In order to improve the speed and detection rate of the intrusion detection system for detecting the intrusion, need to introduce better algorithms or improve the existing algorithms. Intrusion detection requires the ability to quickly and accurately detect a variety of intrusion, so the efficiency of the algorithm requires the higher. It describes two classical algorithms of the decision-making tree: ID3 algorithm and C4.5 algorithm, and analyzes their problems and ways to apply them to intrusion detection. Make some appropriate improvements to them in order to get better results. The experimental simulation verifies that these two improved algorithms can achieve the expected results in discovering the intrusion in the intrusion detection system.

Key words: intrusion detection; decision tree algorithm; intrusion behaviors

0 引 言

在信息化建设日益加深的今天,人们在享受网络带来的快捷便利的同时,也必须正视网络的安全问题,它已经成为当今时代亟待解决的首要问题之一。网络已经渗透到人类生活的方方面面,从传统的 E-mail、即时通讯到现在流行的微博、电子银行等,对网络的了解程度几乎成为衡量一个人是否与社会脱节的标准。因此,部署网络安全设施,增加威胁检测防御手段,保障网络的正常运行已经迫在眉睫。

入侵检测系统[1](Intrusion Detection System)作为继网络防火墙等安全设施之后出现的重要的主动防护措施之一,已经成为网络信息安全研究领域的热点问题。如何提高入侵检测系统中网络事件检测的准确率成了制约入侵检测系统发展的瓶颈。为了提高检测率,文中结合决策树算法和已知的攻击行为建立一种能够快速判断网络数据流是否为入侵行为的方法。

1 决策树算法介绍

决策树分类算法起源于概念学习系统(Concept Learning System, CLS)[2],是以实例为基础的学习算法。1986 年由 Quinlan 提出的 ID3 算法[3]是决策树算法中的典型代表,具有描述简单、分类速度快的优点,

收稿日期:2013-04-04 **修回日期**:2013-07-10 **网络出版时间**:2013-11-12
基金项目:国家"973"重点基础研究发展计划项目(2007CB311203);国家自然科学基金资助项目(60803157,90812001);山东省自然科学基金项目(ZR2009GM009)
作者简介:马志远(1986-),男,硕士研究生,研究方向为企业信息化与系统集成、信息安全;曹宝香,教授,研究方向为计算几何与图形学、计算机辅助设计、数据库技术与系统集成等。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20131112.1634.018.html

适用于大规模数据的分析处理,目前流行的决策树算法大多数都是在它的基础上加以改进实现的较好算法。决策树的构造过程不依赖相关的领域知识,它使用属性选择度量来选择将元组最好划分成不同类的属性。

1.1 ID3 算法介绍

ID3 算法[4]是一个从上到下、分而治之的归纳过程。算法的核心:在决策树各级节点上选择属性时,通过计算信息增益来选择属性,以使得在每一个非叶子节点进行测试时,能够获得被测试属性记录最大的类别信息。ID3 算法以信息论为基础,以信息熵和信息增益为衡量标准实现对数据的归纳分类[5]。

设 X 为训练实例集,A 为全体属性集,其中 $A=\{A_1,A_2,\cdots,A_m\}$,通过学习归纳将训练集样本实例分为 n 个类别,记为 $X=\{X_1,X_2,\cdots,X_n\}$。设第 i 类的训练实例个数为 $|X_i|$,X 中总的训练样本实例个数为 $|X|$。对任一个实例属于第 i 类的概率为:

$$p(X_i)=|X_i|/|X|$$

那么对于 X 的不确定程度(即信息熵)为:

$$H(X)=\sum_{i=1}^{m}(-p(X_i)\log_2 p(X_i))$$

决策树学习的过程就是使得决策树对划分的不确定程度逐渐减小的过程。若选择属性 A 进行测试,设属性 A 具有性质 $a_1,a_2,\cdots,a_t$,在 $A=a_j(j=1,2,\cdots,l)$ 的情况下属于第 i 类的实例个数为 C_{ij} 个,Y_j 为 $A=a_j$ 时实例总个数,对于 $A=a_j$ 的先验概率为:

$$p(A=a_j)=|Y_j|/|X|$$

后验概率 $p(X_i\mid A=a_j)$ 为:

$$p(X_i\mid A=a_j)=C_{ij}/|Y_j|$$

即测试属性 A 在取值为 a_j 时属于第 i 类决策的概率。在 $A=a_j$ 的情况下,关于 X 的平均不确定性,即后验熵为:

$$H(X,A=a_j)=-\sum_j(p(X_i\mid A=a_j)\log_2 p(X_i\mid A=a_j))$$

当选择属性 A 后延伸出的每个 $A=a_j$ 叶节点 Y_j 对于分类的条件熵:

$$H(X,A)=\sum p(A=a_j)H(X,A=a_j)$$

属性 A 对于分类提供的信息量,即属性 A 的信息增益为:

$$I(X,A)=H(X)-H(X,A) \quad (1)$$

通过对所有属性的信息增益比较,选择最大值对应的属性为依据,构建决策树的一级节点。同样的方法构造二级及多级决策树节点,直到达到分类目的。

ID3 算法学习能力较强,构建决策树的平均深度较小,分类速度快,适合处理大规模的学习问题。但是也存在一些问题需要进一步的优化,例如在属性选择时趋向于取值较多的属性,在有些情况下这类属性可能不会提供太多有价值的信息。

1.2 C4.5 算法

对于 ID3 暴露出的一些问题,Quinlan 对 ID3 算法进行改进,提出了 C4.5 算法[6-7]。C4.5 算法现已成为最经典的决策树构造算法,排名数据挖掘十大经典算法之首。C4.5 算法继承了 ID3 算法的诸多优点,并对 ID3 算法进行了改进。C4.5 算法使用信息增益率来选择属性作为分类属性,解决了多取值属性对决策树构造过程中产生的影响。信息增益率计算公式为:

$$GR(S,A)=I(S,A)/split(S,A)$$

其中,$split(S,A)=-\sum_{i=1}^{l}(|S_i|/|S|)\log_2(|S_i|/|S|)$,$S_i(i=1,2,\cdots,l)$ 表示属性 A 的 l 个不同值分割样本集 S 形成的 l 个样本子集。

C4.5 算法也解决了 ID3 算法中只能处理离散属性的问题,也就是说 C4.5 算法能够处理连续属性值。C4.5 算法中对离散属性值的处理参照 1.1 节介绍,这里重点分析 C4.5 算法中对连续属性的处理。在训练样本集中,按照连续属性 A' 的具体数值进行排序(可以从小到大,也可以从大到小),得到属性 A' 的取值序列 $\{A'_1,A'_2,\cdots,A'_{total}\}$,在序列中生成 total-1 个分割点。第 i 个分割点的取值设置为 $v_i=(A_i+A_{i+1})/2$,$(i=1,2,\cdots,total-1)$,这 total-1 个分割点把数据集划分为 total 个子集。从第一个分割点开始,分割 A' 并计算两个集合的期望信息,具有最小期望信息的点称为属性 A' 的最佳分裂点,其信息期望作为此属性的信息期望。

1.3 算法存在的问题

在决策树构造的过程中可能会出现这样的情况:所有属性都作为分裂属性用完了,还存在某个(或者某些)节点集中的元素不属于同一类别。在这种情况下,由于没有更多的信息可以使用,一般对这个(或这些)子集进行“多数表决”,即使用此子集中出现次数最多的类别作为此节点类别,并将此节点作为叶子节点。其次,ID3 算法对噪声数据较为敏感,由于训练集中的正反比例很难控制,如何降低或者消除噪声数据对构建决策树的影响也成为研究的重点。

2 入侵检测系统

2.1 入侵检测系统框架

入侵检测交换格式(IDEF:Intrusion Detection Exchange Format)是由互联网工作小组(IETF:Internet Engineering Task Force)的入侵检测工作组(IDWG:Intrusion Detection exchange format Working Group)在 1999 年 6 月开发的安全事件报警的标准格式。IDEF

主要规范了部分术语的使用。安全检测框架模型如图 1 所示。

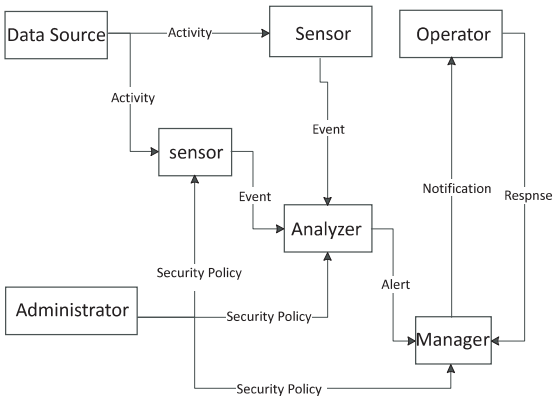

图 1　IETF/IDWG 安全检测框架模型

IETF/IDWG 安全检测框架模型[6]中反映了一个完整的入侵检测系统需要的组成部分。数据源(Data Source)是入侵检测系统的数据来源。探测器(Sensor)主要负责数据的收集与预处理工作,使用数据包嗅探器截获并阅读网络上各个协议层次上的数据包,并把截获的数据经过预处理之后传送到分析器(Analyzer)。分析器也可称为检测引擎(Detection Engine),主要的工作是接收从一个或者多个探测器传来的信息,并通过分析来确定数据是否发生了非法的入侵活动,分析器组件的输出为标识入侵行为是否发生的指示信号,例如警告信号或者 DDos 攻击信号。指示信号中还能包含相关的入侵证据信息。另外,分析器组件还提供了有关可能的防范措施的相关信息。将标识出的信息反映给管理器(Manager),管理器把消息发送通知告诉操作者(Operator),操作者做出反应然后返回给管理器。管理者(Administrator)通过安全策略来影响探测器、分析器和管理器的行为活动。

从模型中可以看出,分析器是 IDS 的核心部分。能否设计出好的分析算法用于判断数据行为的属性,关系到能否检测出是否存在入侵行为的关键。检测率和误检率是入侵检测系统设计人员关注的焦点。他们在努力提高检测的准确率和降低检测的误报率。根据 IETF/IDWG 安全检测框架模型重点设计模型分析器中的检测算法,以便能够达到较好的检测效果。

2.2　数据分析算法改进

决策树算法在分类决策中能够完成较好的分类效果。为了能够把决策树算法应用到入侵检测系统,根据入侵检测系统本身的要求把决策树算法进行适当的改进,以得到较好的检测效率。

网络中哪些是入侵的数据,哪些是正常请求的数据,对入侵检测系统来说都是未知的。发起入侵行为都是不确定的,入侵行为的种类和特点又是多种多样的。对于任何网络安全系统而言,都不能预先判定网络数据的属性。只有在结果出来之后,根据各种日志文件才能判定哪些行为属于入侵行为,以及属于哪类入侵行为。由于属于哪类入侵行为需要多方面的知识和人的参与分析才能最终确定,改进算法主要针对网络数据能够发现入侵行为,不在于能够把入侵行为进行详细的分类。

网络中入侵行为的不可预知性和时间上的不确定性,无法判断入侵行为在何时发生,来源于什么地方。根据这些特性,使用一些先验数据得到的数据作为建立入侵检测系统分析入侵行为的先验知识,建立决策树。ID3 算法中的剪枝条件为:对于某一个分支,当出现下列条件时算法结束:所有的叶子节点同属于一个类别或者为空。修改决策树算法中的剪枝条件如下:当决策树 T 中的所有叶子节点 (X',Q') 满足 $|X_i|/|X'| \geqslant p(X_i \mid A=a_j)^{\alpha}, (0<\alpha\leqslant 1)$ 或者 Q' 为空,即 X' 中包含有的 $X_i(i=1,2,\cdots,n)$ 类的数据个数占总数据个数的比达到 $|X_i|/|X'| \geqslant p(X_i \mid A=a_j)^{\alpha}$,其中 α 是系统设计者预先给出或者管理员通过配置文件给出。通过这样的修改剪枝条件,能够有效地降低噪声数据对决策树生成的影响。根据上面给出的剪枝条件加快了剪枝速度,在建树的过程中能够快速生成决策树。

2.3　算法的主要步骤

在 1.1 和 1.2 部分介绍了决策树算法中的两个经典算法,它们在构造决策树,并用于决策分析中都取得了较好的效果,该小节主要根据前面介绍的算法并结合入侵检测中存在的实际问题,详细描述决策树算法在入侵检测中的步骤。根据 ID3 算法思想,改进的 ID3 算法用于入侵检测中的算法步骤如下:

Step1:初始化决策树 T 为只包含一个树根 $T(X,Q)$,其中 X 是全体样本集,Q 为全体属性集。

Step2:如果 T 中的所有叶子节点 $\mathrm{Leaf}(X',Q')$ 满足 $|X_i|/|X'| \geqslant p(X_i \mid A=a_j)^{\alpha}, (0<\alpha\leqslant 1)$ 或者 Q' 为空,则算法结束。

Step3:否则,对于任何一个不具有 Step2 中的描述的叶子节点 $\mathrm{Leaf}(X',Q')$,执行下面的操作。

Step4:对 Q' 中的属性 A,根据公式(1)计算信息增益 $I(X',A)$。

Step5:选择具有最高信息增益的属性 B 作为节点 $\mathrm{Leaf}(X',Q')$ 的测试属性。

Step6:对于属性 B 的取值 b_i,从节点 $\mathrm{Leaf}(X',Q')$ 伸出分支,代表测试输出 $B=b_i$;求得 X' 中 $B=b_i$ 的子集 X_i,生成相应的叶子节点 $\mathrm{Leaf}(X_i',Q'-\{B\})$。

Step7:返回 Step2。

决策树算法最终目标是生成一棵输出决策树,而

对应的输入则是大量的类别实例。在 IDS 中,决策树中的每一个从树根到叶子节点的一个通路就对应于一条检测规则,这些规则的集合能够作为入侵检测的规则集用在分析器中。

在 C4.5 算法中,使用的剪枝条件和 ID3 算法相同,只是在构建决策树的过程中使用少量的连续属性,以增加检测的准确率。

3 实验验证和结果分析

通过使用 KDD Cup 1999 数据集作为实验数据来验证上面介绍的方法是否能够取得较为满意的效果。

3.1 实验数据来源

文中采用的实验数据为:KDD Cup 1999 数据集[8]。该数据集是 1998 年由麻省理工学院林肯实验室为入侵检测模型评估而建立的测试数据集。该数据集一共 490 万条记录,每条记录共有 41 个数据属性和一个标识属性。41 个数据属性描述数据的特点,除了一些基本属性外,还利用领域知识扩展了一些属性(如登录失败的次数、文件生成操作的数目等),一个标识属性描述数据是否为入侵行为。

KDD99 数据集中共包含 24 种类型的攻击,根据攻击方法和目的可以分为四大类:PROBE、DOS、U2R 和 R2L。PROBE 为进行信息收集的攻击类型;DOS 为拒绝合法用户请求的攻击;U2R 为远程主机非法获取本地主机权限的攻击;R2L 为本地非超级用户获取超级用户权限的攻击。在实验阶段,选取 KDD99 数据集中 10% 的数据集。选取一定量的数据作为训练数据,这些测试数据主要用来验证使用聚类分析算法和决策树算法生成入侵检测的规则能否检测出入侵行为并发现新的入侵行为,并且给出相关的检测率和误检率。其中,检测率和误检率定义如下:

检测率:$D_r = n_i/N_i$,表示入侵行为的检测比例。其中,n_i 为检测出的入侵实例数目;N_i 为数据集中入侵实例的总数。

误检率:$F_r = (N_i - n_i)/N_n$,表示误将入侵数据行为判断为正常数据行为的比例。其中,N_n 表示数据集中的正常实例数目。

3.2 实验结果分析

实验环境主要是基于 Java 的开发应用环境,并且借用 Weka 中的相关源码进行了算法改造。

为了验证文中提出的算法的有效性,通过参考文献中相关算法给出的 D_r 和 F_r 进行了对比。具体对比结果如表 1 所示。通过表 1 中的对比,可以看出文中提出的算法检测率均高于其他算法给出的检测结果,但是在误检率方面存在较大的差距,需要进行相关改进。文中的算法的误检率虽然低于传统的 K-means 算法,但是相比较其他算法如改进的 K-means 算法[9]、聚类分析和关联规则的算法[10]以及 DCA 和 NSA 结合使用[11],具有较高的误检率。

表 1 相关算法检测率和误检率比较 %

	D_r	F_r
文中	90.68	2.53
K-means 算法	87.6	5.5
文献[8]算法	77.4	0.88
文献[9]算法	87.67	0.19
文献[10]算法	78.54	0.43

4 结束语

文中从入侵检测系统的检测率入手,使用决策树分类算法作为检测手段,讨论了入侵检测系统。使用改进决策树算法验证了能够提高入侵检测的准确率和降低入侵检测的误报率,达到了较好的效果。文中提出的改进算法可以作为入侵检测分析器中的一种判断网络数据行为的分类依据。改进算法通过验证虽然能够达到较好的效果,但要想真正应用在入侵检测系统中还有一些问题需要完善和补充。

参考文献:

[1] 陈小辉. 基于数据挖掘算法的入侵检测方法[J]. 计算机工程,2010,36(17):72-73.

[2] Michalski R S, Mozetic I, Hong J, et al. The multi-purpose incremental learning system AQ15 and its testing application to three medical domains[C]//Proc of AAAI. [s. l.]:[s. n.], 1986.

[3] Quinlan J R. Induction of decision trees[J]. Machine learning,1986(1):81-106.

[4] Cios K J, Sztandera L M. Continuous ID3 algorithm with fuzzy entropy measures[C]//Proc of IEEE international conference on fuzzy systems. [s. l.]:[s. n.],1992.

[5] 张燕平,张 玲. 机器学习理论与算法[M]. 北京:科学出版社,2012:227-241.

[6] Quinlan J R. C4.5: Program for machine learning[M]. Los Altos: Morgan Kaufmann Publishers, 1993.

[7] Koshal J, Bag M. Cascading of C4.5 decision tree and support vector machine for rule based intrusion detection system[J]. International journal of computer network and information security, 2012, 4(8):8-20.

[8] KDDCup 1999 Data[EB/OL]. 1999. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

[9] 薛京花. K-means 聚类算法在网络入侵检测中的应用研究[D]. 长沙:中南林业科技大学,2010.

[10] 李 建. 基于聚类分析和关联规则的网络入侵检测研究[D]. 长沙:中南大学,2011.

[11] 张春香. DCA 算法和 NSA 算法结合的入侵检测模型研究[D]. 武汉:武汉科技大学,2011.

改进的决策树算法在入侵检测中的应用

作者： 马志远， 曹宝香， MA Zhi-yuan， CAO Bao-xiang
作者单位： 曲阜师范大学 计算机科学学院,山东 日照,276826
刊名： 计算机技术与发展
ISTIC
英文刊名： Computer Technology and Development
年，卷(期)： 2014(1)