Journal of Computer Applications
计算机应用,2015,35(10):2883-2885
ISSN 1001-9081
CODEN JYIIDU
2015-04-10
http://www.joca.cn

文章编号:1001-9081(2015)10-2883-03
doi:10.11772/j.issn.1001-9081.2015.10.2883

基于决策树与朴素贝叶斯分类的入侵检测模型

姚 潍,王 娟*,张胜利
(深圳大学 信息工程学院,广东 深圳 518000)
(*通信作者电子邮箱 juanwang@szu.edu.cn)

摘 要:入侵检测要求系统能够快速准确地找出网络中的入侵行为,因此对检测算法的效率有较高的要求。针对入侵检测系统效率和准确率偏低,系统的误报率和漏报率偏高的问题,在充分分析C4.5算法和朴素贝叶斯(NB)算法后,提出一种二者相结合的H-C4.5-NB入侵检测模型。该模型以概率的形式来描述决策类别的分布,并由C4.5和NB概率加权和的形式给出最终的决策结果,最后使用KDD 99数据集测试模型性能。实验结果表明,与传统的C4.5、NB和NBTree方法相比,在H-C4.5-NB中对拒绝服务(DoS)攻击的分类准确率提高了约9%,对U2R和R2L攻击的准确率提高约20%~30%。

关键词:入侵检测;决策树;朴素贝叶斯;概率加权和
中图分类号:TP393.08 **文献标志码**:A

Intrusion detection model based on decision tree and Naive-Bayes classification

YAO Wei, WANG Juan*, ZHANG Shengli
(*College of Information Engineering, Shenzhen University, Shenzhen Guangdong 518000, China*)

Abstract: Intrusion detection requires the system to identify network intrusions quickly and accurately, so it also requires high efficiency of the detection algorithm. In order to improve the efficiency and accuracy of intrusion detection system, and reduce the rate of false positives and false negatives, a H-C4.5-NB intrusion detection model combined C4.5 with Naive Bayes (NB) was proposed after fully analyzing the C4.5 and NB algorithm. The distribution of decision category was described in the form of probability in this model, and the final decision results were given in the form of C4.5 and NB probability weighted sum. Finally the performance of the model was tested by KDD 99 data set. The experimental results showed that the accuracy of Denial of Service (DoS) was improved about 9% and the accuracy of U2R and R2L was improved about 20% - 30% in H-C4.5-NB compared to the traditional methods such as C4.5, NB and NBTree.

Key words: intrusion detection; Decision Tree (DT); Naive Bayes (NB); probability weighted sum

0 引言

随着互联网时代的到来,当今社会的网络环境越来越复杂化,网络中的安全问题也层出不穷。因此提高网络的安全防御能力,增加安全检测的机制,保证网络的正常运行,是当前急需解决的问题之一。入侵检测系统(Intrusion Detection System, IDS)[1]是继防火墙之后的又一道网络安全屏障,对网络中出现的入侵行为进行主动防御,也是当前研究的最多、应用最广的安全手段之一。

IDS核心的部分就是检测算法,良好的算法可以全方面地提高网络事件检测的准确率。近些年来的研究表明,将机器学习[2-4]的方法应用到入侵检测领域,可以极大地提高IDS的效率和准确率。然而,在入侵检测中很难再去设计出一个单一的分类器,使其性能比现存的更好,相反混合分类方法近些年来却从边缘化走向了主流[5]。在现有的入侵检测技术中,决策树(Decision Tree, DT)[6-7]和朴素贝叶斯(Naive-Bayes,NB)[8-9]分类算法的模型最为简单、结果易于理解而且分类精度较高。早期,Kohavi提出了一种朴素贝叶斯树(Naive-Bayes Tree,NBTree)算法[10],通过决策树来划分属性,在叶子节点处构建局部朴素贝叶斯分类器,这一方法需要反复地在每个叶子处验证整体分类性能;Jiang等[11]提出一种C4.5-NB算法,利用二者的优势共同来分类新实例,然而C4.5-NB算法并未考虑树的过度拟合问题以及NB的独立性假设条件。

因此,本文以混合混分类为前提,充分考虑和分析了决策树和朴素贝叶斯算法的原理后,结合这二者的特点在C4.5-NB基础上提出一种改进模型(Hybrid C4.5 and Naive-Bayes, H-C4.5-NB),并利用KDD 99数据对H-C4.5-NB性能进行了测试。实验结果表明这种方法能够在一定程度上提升分类的准确率,并且降低了误报率。

1 H-C4.5-NB算法

1.1 算法的思路

入侵检测是要识别网络中的连接,并判断其属于正常访问还是入侵行为,因此可以归结为一个多分类问题。从理论上来说,NB算法由于其坚实的数学理论基础,错误率应当是最低的,然而它最主要的问题就在于其独立性假设条件在实际应用中并不能完全满足;DT算法由于在训练的过程中需要递归地创建测试节点来对训练集进行划分,每次都需要对整个训练集合进行多次顺序扫描,而且当训练集中存在噪声数据时或者训练数据取样不均匀时,会导致树的过度拟合问题。

本文在结合NB和经典决策树算法C4.5[12]的基础上,提

收稿日期:2015-04-02;**修回日期**:2015-07-27。 **基金项目**:国家自然科学基金资助项目(61372078)。
作者简介:姚潍(1990-),男,湖北黄冈人,硕士研究生,主要研究方向:无线通信网; 王娟(1979-),女,广东深圳人,副教授,博士,主要研究方向:无线通信网; 张胜利(1978-),男,广东深圳人,副教授,博士,主要研究方向:物理层网络编码。

出一种新的方法。一方面,在数据子集上训练朴素贝叶斯分类器在某种程度上降低属性之间的相关性,尽量满足条件独立性的假设;另一方面,为了避免决策树的过度拟合,需要控制决策树的规模,在整个数据被完全划分之前就停止决策树的生长。因此,提出 H-C4.5-NB 这样一个混合的分类器,针对整个训练数据集构建两个局部的弱分类器,最后将它们的结果融合起来构成最终的决策规则。

1.2 算法的系统原型和流程

基于 H-C4.5-NB 算法的系统检测过程如图 1 ~2 所示。

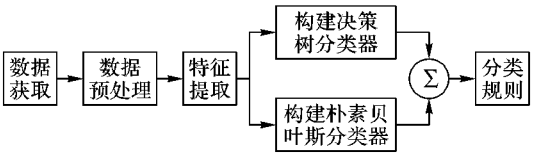

图 1 基于 DT 和 NB 的分类器训练过程

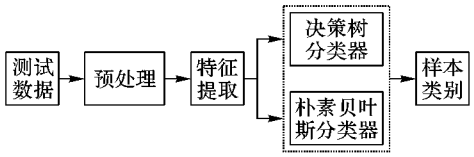

图 2 基于 DT 和 NB 的分类器测试过程

数据获取一般是指使用任何可能的方法从主机或者网络中提取原始数据,包括分类器学习用的训练数据和测试分类器性能的验证数据以及使用分类器所用的测试数据。通常在获得这些原始的数据之后并不能立刻应用到实际当中,不同的场合可能会需要不同的数据格式,因此还需要将这些数据进行预处理,主要的工作包括一些前期的数值化、标准化等。前期工作完成之后,可能还需要对数据进行一些特征提取,以便提高分类器的识别率和效率。最后利用这些处理完毕的数据进行分类器的训练,得到训练好的分类器之后就可以用于新数据的测试。

算法的详细描述如下。

1)训练阶段。

输入 原始数据集合。

输出 C4.5 和 NB 分类器;在训练数据上分类器的准确度$A_{C4.5}$和A_{NB}。

1)对训练数据集合进行预处理,规范数据格式;

2)通过特征提取,将处理好的数据划分为连续属性和离散属性两个子集;

3)针对离散集合,随机抽取 80% 数据用作训练集构建 NB 分类器,20% 用作验证集评估分类器性能,重复执行 10 次得到最终的分类器 NBC 以及其性能A_{NB};

4)针对连续集合,随机抽取 80% 数据用作训练集构建决策树分类器,20% 用作验证集评估分类器性能,重复执行 10 次得到最终的分类器 C4.5 以及其性能$A_{C4.5}$。

2)测试阶段。

输入 测试数据;C4.5 和 NB 分类器;$A_{C4.5}$和A_{NB}。

输出 测试数据类别 C。

1)对测试数据进行与训练数据一致的预处理和特征提取,将其分为离散和连续两个属性子集;

2)对测试数据中每个样本 x:分别利用 NB 分类器评估 x 属于每个类别的后验概率 $P(C_i \mid x)_{NB}$ 以及利用 C4.5 分类器评估 x 属于每个类别的后验概率 $P(C_i,x)_{C4.5}$;

3)根据式(1)评估 x 属于每个类别的概率,最终取 $P(C_i \mid x)$ 最大值对应的类别 C_i 作为 x 所属的类别。

$$P(C_i \mid x) = \frac{A_{C4.5} \times P(C_i \mid x)_{C4.5} + A_{NB} \times P(C_i \mid x)_{NB}}{A_{C4.5} + A_{NB}} \tag{1}$$

需要说明的一点,就是为了缩短决策树的构建时间,降低树的复杂度,这里的 C4.5 算法采用预剪枝方法提早停止树的生长。因此在叶子节点处训练样本并不是纯的,而是以概率的形式来描述:

$$P(C_i \mid x)_{C4.5} = \frac{1}{n}\sum_{i=1}^{n}\delta(C_i,C)$$

其中:n 表示的是叶子节点处包含的样本数。叶子节点含有的样本数对决策树性能的影响如图 3 所示。

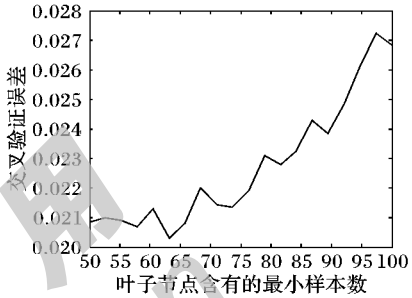

图 3 叶子节点含有的样本数对决策树性能的影响

2 实验结果及分析

本次实验所使用的数据来源于标准的入侵检测数据集 KDD 数据[13],包含 25 133 条训练数据和 22 544 条测试数据,其样本的分布如图 4 所示。

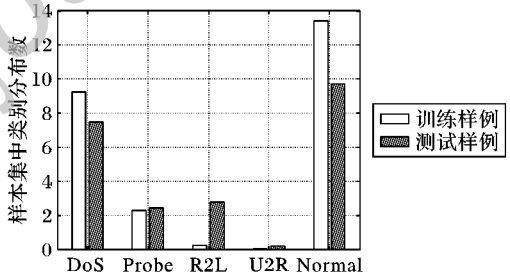

图 4 训练数据和测试数据样本分布

实验之前先要对数据进行一些预处理,包括数值化和标准化。部分属性数值化的关系如表 1 所示。

表 1 部分离散属性数值化关系对照表

属性类型	字符串—数字值转换
Protocol_type	icmp—1, tcp—2, udp—3, others—4
service	domain_u—1, ecr_i—2, eco_i—3, finger—4, ftp_data—5, ftp—6, http—7, hostnames—8, imap4—9, login—10, mtp—1, netstat—12, other—13, private—14, smtp—15, systat—16, telnet—17, time—18, uucp—19, 其他服务—20
flag	REJ—1, RSTO—2, RSTR—3, S0—4, S3—5, SF—6, SH—7, OTHERS—8

决策类别的数值化关系如下:Normal—0、DoS—1、Probe—2、R2L—3、U2R—4。

通常分类器的一个主要性能评估指标就是准确率,本次实验就主要从算法检测到的类别数目、算法检测率、算法误检率以及分类器整体准确率四个方面对几种不同的分类算法作详细的对比。

表 2　不同算法检测到的类别数对比

类别	C4.5	NB	NBTree	C4.5-NB	H-C4.5-NB
DoS	5898	5804	5689	6195	6895
Probe	1262	1297	1556	1392	1302
R2L	240	256	272	277	1328
U2R	15	24	12	24	75
Normal	9128	8329	9135	9043	9415

表 3　不同算法的检测率(TP-rate)对比　%

类别	C4.5	NB	NBTree	C4.5-NB	H-C4.5-NB
DoS	69.76	68.6	76.30	73.27	81.55
Probe	88.56	91.00	64.30	97.61	91.36
R2L	8.70	9.20	9.90	10.00	48.20
U2R	7.50	12.00	6.00	12.00	37.50
Normal	93.90	85.76	94.1	93.10	96.90

表 4　不同算法的误检率(FP-rate)对比　%

类别	C4.5	NB	NBTree	C4.5-NB	H-C4.5-NB
DoS	1.52	3.96	5.6	2.57	2.21
Probe	4.82	6.65	2.40	4.45	3.06
R2L	0.54	0.56	0.10	0.41	0.28
U2R	0.25	5.94	0(约)	0(约)	0(约)
Normal	35.89	26.80	35.30	30.10	19.54

表 2~4 从 KDD 数据五大类别的检测数目、检测率(TP-rate)以及误检率(FP-rate)三个方面对几种不同算法的性能作了直观的对比。分类器最终的结果是要使得其检测率最高而误检率最低。可以看到,H-C4.5-NB 算法相对于其他几种方法,其性能都有一定程度上的提升。虽然在检测 Probe 攻击的过程中,H-C4.5-NB 算法性能有小幅度的下降,但是鉴于数据的复杂性以及其本身已经达到的 90% 多的准确度,这个小幅度的下降是在可以接受的范围内。对于 DoS 攻击、Probe 攻击以及正常访问数据 Normal,H-C4.5-NB 算法都有不错的性能;而对于 R2L 和 U2R 两类攻击,H-C4.5-NB 算法的检测率看上去并不是很高。这个问题有两方面的原因:其一,相比较随机猜测的几率,从五大类别中猜中某一类别的概率仅为 20%,而 H-C4.5-NB 的检测率均要高于 20%,并且优于其他几种方法;其二,通过图 4 也可以看到,用于训练的原始样本中包含的 U2R 和 R2L 类别本来就少,而测试数据的数量又大于训练数据,这也恰好从一定程度上反映了数据样本质量的好坏与分类器性能之间的关系。

评价一个分类器最为直观的指标就是分类准确度。实验结果表明,NB 分类器性能最差,其准确度只有 69.68%,主要就是因为原始数据中包含过多的相关属性,并且连续属性偏多,限制了朴素贝叶斯的条件。C4.5-NB 相对于 C4.5 和 NB 虽然有小幅度的提升,准确度达到了 74.99%,但是其本质并没有改善决策树结构和朴素贝叶斯限制性。NBTree 方法通过树划分属性后构建局部贝叶斯分类器,理论上削弱了部分依赖关系,因此其性能要优于 C4.5 和 NB,准确度接近 73.92%,但是从结果可以看出,其在 DoS 和 Normal 上的误检率却远高于其他几种方法。因为在训练的过程中,H-C4.5-NB 方法只针对数据集的部分属性分别构建决策树和朴素贝叶斯分类器,因此其准确度最高达到了 84.34%。单纯从训练时间上看,H-C4.5-NB 是最优的;在数据样本的存储中,H-C4.5-NB 与 C4.5、NB 是一致的,而 C4.5-NB 方法的存储空间占用率最高。

3　结语

针对朴素贝叶斯的强制条件独立性假设和决策树中过度拟合的问题,本文提出了一种基于这二者相结合的 H-C4.5-NB 算法。实验结果表明,在属性子集上去构建局部的朴素贝叶斯分类器,在一定程度上削弱了条件独立性的限制,同时子集上构建的决策树相比整个数据集而言,复杂度要大大降低,加上剪枝规则的使用在一定程度上也避免了树的过度拟合问题,因此提高了分类的准确率和效率。但是该算法也还不够完善,可以考虑更为有效的属性子集提取方法以及更加精确的结果融合方法,进一步来提升分类的性能。

参考文献:

[1] AXELSSON S. Intrusion detection systems: a survey and taxonomy [J]. Computers and Security, 2000, 20(1):676 - 683.

[2] SOMMER R, PAXSON V. Outside the closed world: on using machine learning for network intrusion detection[C]// Proceedings of the 2010 IEEE Symposium on Security and Privacy. Washington, DC: IEEE Computer Society, 2010:305 - 316.

[3] SHAMSHIRBAND S, ANUAR N B, KIAH M L M, *et al.* An appraisal and design of a multi-Agent system based cooperative wireless intrusion detection computational intelligence technique[J]. Engineering Applications of Artificial Intelligence, 2013, 26(9): 2105 - 2127.

[4] SINGH J, NENE M J. A survey on machine learning techniques for intrusion detection systems[J]. International Journal of Advanced Research in Computer and Communication Engineering, 2013, 12(1): 4349 - 4355.

[5] TSAI C F, HSU Y F, LIN C Y, *et al.* Intrusion detection by machine learning: a review[J]. Expert Systems with Applications, 2009, 36(10): 11994 - 12000.

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

[7] KUMAR M, HANUMANTHAPPA M, KUMAR T V S. Intrusion detection system using decision tree algorithm [C]// Proceedings of the 2012 IEEE 14th International Conference on Communication Technology. Piscataway: IEEE Press, 2012:629 - 634.

[8] JIANG L, CAI Z, ZHANG H, *et al.* Naive-Bayes text classifiers: a locally weighted learning approach[J]. Journal of Experimental and Theoretical Artificial Intelligence, 2013, 25(2):273 - 286.

[9] DESHMUKH D H, GHORPADE T, PADIYA P. Intrusion detection system by improved preprocessing methods and Naive Bayes classifier using NSL-KDD 99 Dataset[C]// Proceedings of the 2014 International Conference on Electronics and Communication Systems. Piscataway: IEEE Press, 2014:1 - 7.

[10] KOHAVI R. Scaling up the accuracy of Naive-Bayes classifiers: a decision-tree hybrid[EB/OL]. [2015-01-10]. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.462.9093&rep=rep1&type=pdf.

[11] JIANG L, LI C, WU J, *et al.* A combined classification algorithm based on C4.5 and NB[C]// ISICA 2008: Proceedings of the Third International Symposium on Advances in Computation and Intelligence, LNCV 5370. Berlin: Springer-Verlag, 2008: 350 - 359.

[12] GONDY L A, THOMAS C R B, BAYES N. Programs for machine learning[EB/OL]. [2014-10-10]. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.81.5894.

[13] SABHNANI M, SERPEN G. Application of machine learning algorithms to KDD intrusion detection dataset within misuse detection context[EB/OL]. [2014-10-10]. http://neuro.bstu.by/ai/To-dom/My_research/Papers-0/For-research/D-mining/Anomaly-D/KDD-cup-99/mlmta03.pdf.