基于信息密度的优化贝叶斯算法

北京交通大学 计算机与信息技术学院

**摘要：**为更好实现对云平台故障检测数据的分类处理，提出了一种基于信息密度的优化贝叶斯算法，采用信息熵描述数据特征的不确定性，信息密度作为数据特征分类的权重，并优化贝叶斯算法使其能够处理数据特征值没有变化的情况。本文对算法进行了细致描述，分析了算法的时间复杂度。实验结果表明，该方法可有效减少数据信息损失，抽取数据特征的内部关系，能够准确将云平台故障检测数据分类。

**关键词：**云平台故障检测 数据特征 信息密度 优化贝叶斯算法

The optimal Bayes algorithm based on information density

DU Ye, ZHANG Tian-tian,

( School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044,China)

**Abstract** For better implementation of the classification of the cloud platform fault detection data processing, this paper proposes an optimal Bayes algorithm based on information density, using information entropy to describe the data characteristics of uncertainty, information density as the weight of the characteristics of data classification. Optimize the Bayesian algorithm to process the data characteristic values which are not changed . This paper describes the algorithm deeply, and analyzes the time complexity of the algorithm. The experimental results show that the method can effectively reduce the data loss, analyze the data characteristics. This method can classify detection data of cloud platform fault accurately.

**Key Words** Cloud platform fault detection; Data characteristic; Information density; Optimal Bayes algorithm

1. 引言

随着互联网信息技术的发展，云数据平台已经影响了人们的生活和工作方式。一方面它为人们的生活带来无穷便利，但另一方面，云数据平台安全还存在着诸多待解决的问题。针对云数据平台的网络攻击也层出不穷，攻击手段方法不断增加，已呈智能化与协同性发展。根根据国家互联网应急中心（CNCERT）发布的《2014年我国互联网网络安全态势报告》[1]数据，中国网络安全形势不容乐观，2014年CNCERT通报的漏洞事件达9068起，较2013年增长3倍。截至2014年12月底，中国网站总量规模为364.7万个，网民规模达6.49亿，手机网民规模5.57亿，互联网普及率达到47.9%。随着互联网的迅速发展，相伴产生的新安全问题也层出不穷，基础网络和新型网络产品带来的漏洞风险日益上升。云数据平台网络攻击事件的频繁发生，对广大网民网络利益产生影响，也深深社会经济发展、国家信息安全造成威胁与挑战。为了保护云数据平台系统资源，作为主动的安全防护技术，云平台故障检测的研究与发展显得重要。

云数据平台网络故障包括入侵系统安全网络，故意地、非授权地试图访问信息、删除信息、使数据系统平台不可靠或不可用。由于云平台的系统资源有共享的要求，识别和追踪非授权用户需要相关安全的技术要求。网络攻击者可以利用大量被劫持的网络主机来占用云平台网络的关键系统网络资源，这样会使得云平台网络服务质量明显降低，即发生云平台故障。云平台故障检测的作用就在于及时地发现各种入侵云平台攻击以及攻击企图，并做出及时有效的反应。云平台故障检测技术认为网络入侵活动是未知的，网络异常活动的是随时存在的。当网络活动行为偏离达到一定程度时，都认为是云平台故障事件发生。

由于云平台故障检测需要对大量的网络访问请求数据进行分析和处理，例如基于网络访问数据的检测需要对连续时间，协议类型，传送的字节数等参数进行分析，且通常处理数据是算法是复杂的，这增加云平台故障检测系统分析的复杂度、时效性。因此，在进行云平台故障检测有必要对算法进行优化处理。优化算法可以适当降低问题规模，减小问题复杂度，去除不相关和冗余的数据，从而增加检测的准确度。

本文提出了一种基于信息密度优化的贝叶斯算法，采用信息密度对数据特征的不确定性进行密度划分，通过优化的贝叶斯算法实现模型学习，提高模型训练精度。文章的结构如下，第1节简单介绍了该领域的相关研究背景；第2节介绍云平台故障检测数据集及数据的编码处理；第3节提出了基于信息密度的优化贝叶斯检测算法，并对算法进行了详细描述；第4节对模型的可用性，模型的收敛性及算法时间复杂度和空间复杂度要进一步分析；第5节给出实验结果与问题；最后对全文的工作进行了总结。

1.相关工作

云计算平台面临诸多安全隐患，平台构建存在漏洞，可用性和完整性差；云计算平台的网络访问存在大量非安全访问，由于互联网的资源共享原则攻击者可以利用大量网络主机来占用网络的关键资源使得网络服务质量显著降低，这种行为将严重影响云计算平台的存储服务和计算能力。

焦从信[2]提出了一种基于有向完全图的贝叶斯分类器，将属性之间的关系加入到分类器的构造中，降低了朴素贝叶斯分类器的强独立性假设,并将其应用于入侵检测中。章倩[3]使用主成分分析法提取网络数据包关键属性、消除冗余属性、降低维数，再用贝叶斯分类器进行分类。文桥[4] 提出了一个改进的贝叶斯模型,对朴素贝叶斯算法进行了改进,降低了朴素贝叶斯算法的强独立性假设,提高了入侵检测的分类精度,并通过试验对算法进行了验证和性能分析，挖掘数据之间的内部关系，得到数据模型，从而对数据进行分类。潘志松[5]使用基于神经网络对入侵检测研究，对数据属性进行分析，用样本数据对神经网络训练，然后使用测试数据来检验样本的准确度。饶鲜[6]提出入侵检测系统中应用支持向量机算法，使得入侵检测系统在小样本(先验知识少)的条件下仍然具有良好的推广能力。首先介绍入侵检测研究的发展概况和支持向量机的分类算法，接着提出了基于支持向量机的入侵检测模型。

由于云计算平台故障检测需要对大量的系统实时数据进行分析和处理，例如基于网络数据的检测需要对链接连续时间，网络协议类型，传送的字节数、访问频率等参数进行分析，且通常这些数据是复杂的且难以处理，因此，要通过优化算法的方式来适当降低问题规模，减小问题复杂度，提高算法的计算精度，从而增加检测的准确度。

本文研究云平台故障检测方法，重点研究了贝叶斯算法处理海量网络访问数据的方法，并在贝叶斯算法的基础上，提出了基于信息密度的优化贝叶斯算法，分析模型的可用性，模型的收敛性分析，算法时间复杂度和空间复杂度。

2.基于信息密度优化贝叶斯算法

贝叶斯分类算法[7]是一种简单有效的分类算法，贝叶斯定理是贝叶斯理论中最重要的一个公式它将事件的先验概率与后验概率巧妙地联系起来利用先验信息和样本数据信息确定事件的后验概率。贝叶斯的思想基础是这样的：对于给出的待分类项，求解在此项出现的条件下各个类别出现的概率，概率最大的类别，就认为此待分类项属于的类别。贝叶斯分类算法假设属性之间相互独立，这个假设在实际数据中往往是不成立的；贝叶斯分类算法没考虑数据特征的内部联系和数据特征值的不确定性，而且忽视了数据特征值的分布关系。以上问题使得贝叶斯分类算法的适用性减小，为了解决贝叶斯分类算法的问题，本文提出了基于信息密度的优化贝叶斯算法。提出新的信息密度的概念，新的信息密度的定义是表示信息的分布密度即信息的不确定度。计算数据特征的信息熵，进而计算出信息密度。

基于信息密度的优化贝叶斯算法的过程：

第一步：将数据矩阵按类别分块，不同类别的数据被分开，得到类别数据矩阵。

第二步：类别数据矩阵中有不同的数据特征，对所有数据特征计算出数据特征值概率取值表格。

第三步：根据特征值的概率，求解除所有数据特征的信息熵。

第四步：对所有数据特征的信息熵[8]进行计算，求出数据特征的信息密度值。

第五步：引入信息密度修正值，用梯度下降法求解出最合适的信息密度修正值。

第六步：信息密度值与密度修正值之和为更新后的信息密度值。最终得到的信息密度值为该数据特征的概率权重值。

第七步：将某条测试数据记录的每个数据特征值在数据特征值概率取值表格进行匹配，得到所有数据特征的概率取值。

第八步：将得到所有数据特征的概率取值与对应数据特征的信息密度值相乘再求和，得到当前测试数据记录的分类结果，算法结束。



图2-1基于信息密度的优化贝叶斯算法流程图

2.1优化信息密度值

开始，这些密度修正值初始化为随机值，梯度下降策略执行贪心的爬山法，因为在每一次迭代或每一步，结果都向此时最优解得方向移动，每次迭代后都要更新密度修正值。最终，它收敛域一个局部最优解。

训练数据被带入各类型数据模型中，与每一个类型的数据模型进行匹配。得到测试数据的每个特征值在各类型的数据模型中出现的概率值。在将这些概率值与信息密度值相乘，得到测试数据经过数据模型处理后的结果。

 (2-1)

其中 表示结果， 表示数据类型， 表示特征出现的概率矩阵， 表示特征 在类型的数据模型中出现的概率值，表示特征 在类型的数据模型中信息密度值。最终测试数据会得到多个结果，比较多个结果并取最大的。最大的所对应的类型，就是测试数据的分类结果。

目标函数：

 (2-2)

信息密度值优化公式：

 (2-3)

其中 为第 次信息密度值矩阵， 为第 次的信息密度矩阵， 为第 次密度修正值。

 (2-4)

第一步：通过对信息熵进行负指数变换可以得到信息密度值。

第二步：按照当前信息密度值，计算分类结果 。

第三步：根据计算出的分类结果，按照实际类型，来判断分类是否正确。

第四步：若正确或当前信息密度值已经为1，则密度修正值为0；若错误，则密度修正值为随机值（不高于当前信息密度值的10%）。

第五步：根据新的新的信息密度值，计算分类结果。

第六步：使用梯度下降法[9]对密度值进行计算，求出最优的信息密度值。根据计算出的分类结果，按照实际类型，来判断分类是否正确。

第七步：重复四到六步，直到密度修正值均为0或者目标函数 的达到稳定状态或者达到迭代次数。

第八步：数据模型训练完毕。

2.2时间复杂度分析

假设 代表数据集中数据的条数， 代表每条数据中数据特征的个数。算法将数据矩阵按类别分块，不同类别的数据被分开，得到类别数据矩阵，时间复杂度为 。类别数据矩阵中有不同的数据特征，对所有数据特征计算出数据特征值概率取值表格，时间复杂度为。根据特征值的概率，求解除所有数据特征的信息熵，时间复杂度为。对所有数据特征的信息熵[8]进行计算，求出数据特征的信息密度值，时间复杂度为。引入信息密度修正值，用梯度下降法求解出最合适的信息密度修正值，时间复杂度为。信息密度值与密度修正值之和为更新后的信息密度值。最终得到的信息密度值为该数据特征的概率权重值，时间复杂度为。将某条测试数据记录的每个数据特征值在数据特征值概率取值表格进行匹配，得到所有数据特征的概率取值，时间复杂度为。将得到所有数据特征的概率取值与对应数据特征的信息密度值相乘再求和，得到当前测试数据记录的分类结果，算法结束。

综合所述，算法的时间复杂度为。

3.实验

3.1实验环境

实验在使用操作系统是windows7 64位系统，CPU的型号是Core i5，内存空间为4GB，硬盘空间为1T，使用的软件环境为MATLAB R2012b。在MATLAB的工作环境中编写程序对数据进行处理，计算数据矩阵，调整数据特征格式，使数据符合数据模型的要求；统计数据特征出现的概率，并求解相关的参数，实验平台会整合自动化应用程序提交和交叉程序数据兼容性。

表3-1实验平台环境表

|  |  |  |
| --- | --- | --- |
| 实验平台 | 型号 | |
| 操作系统 | windows7 64位 | |
| CPU | Core i5 | |
| 内存 | 4GB |
| 硬盘 | 1TB | |
| 软件环境 | MATLAB R2012b | |

3.2云平台故障检测数据集

美国国防部高级规划署(DARRA)的入侵检测评估项目,是由麻省理工学院林肯实验室主持,并由DARPA机构和美国空军研究实验室资助的。该项目的目的是对入侵检测技术实现全面的技术评估。该项目开始于1998年,在1999年和2000年也提供了评估数据。它是利用与空军基地相似的背景通信量来评估各种入侵检测的检测率和误报警率。KDDCUP99数据集中有38种攻击被分为4类,Dos类、Porbing类、R2L类和U2L类。而R2L类和U2L类攻击所占比例很小。训练数据中包含有22种攻击,而测试数据中包含有17种未见攻击。攻击分类见下表。

表3-2数据集类型表

|  |  |  |
| --- | --- | --- |
| 数据集  类别 | 训练集中出现的攻击 | 测试数据中出现的新攻击 |
| DOS | Back,land,neptune,pod,smurf,  teardrop | Apache2,mailbomb  Processtable,udpstorm |
| Probe | Ipsweep,nmap,portsweep,satan | mscan,saint |
| R2L | ftp\_write,guess\_passwd,imap,  mulitihop,phf,spy,warezclient,  warezmaster | named,sendmail,snmpgetattack,  snmpguess,sqlattack,xlock,xsnoop |
| U2R | buffer\_overflow,loadmodule,  perl,rootkit | httptunnel,ps,worm,xterm |

为了从包含大量冗余信息的数据中提取出尽可能多的安全信息,抽象出有利于进行判断和比较的特征集合,Wenke Lee在数据特征抽取方面已做了很多的工作,从DARPA1998[10]数据中抽取41维特征, 分为基本特征(basic feature)、内容特征(content feature)、两秒钟内的流量特征(traffic feature computed using a two-second time window)、主机流量特征(host-based traffic features)。

3.3数据预处理

对收集的数据进行预处理。某条数据特征值出现空白记录，则用相邻两个相邻的数据的平均值进行填充，如果当前特征值有明显的变化趋势，则使用回归拟合的方法拟合符合规律的数值，并进行填充。对数据进行观察，按照不同的类型将数据划分成数据矩阵，并计算出每种类型的数据数量。由于原始KDDCUP数据集[11]中的数据类型众多，为了使实验结果具有代表性，将四大类型的攻击与数据集中具体的攻击对应起来。

表3-3数据类型和攻击类型表

|  |  |  |
| --- | --- | --- |
| 类型 | 攻击类型 | 数量 |
| DOS | teardrop | 829 |
| Probe | nmap | 816 |
| R2L | spy | 813 |
| U2R | perl | 52 |
| NORMAL | normal | 4761 |
| 共计 |  | 7271 |

为了方便算法处理，对数据特征进行编码，将不具有数值意义的数据和具有数值意义的数据统一编码。这样使无数值意义的字符项变成具有数值意义，方便算法的处理，部分编号省略。

表3-4数据字符编码表

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 协议类型 | 编号 | 连接状态 | 编码 | 类型 | 编码 | 服务类型 | 编号 |
| tcp | 1 | OTH | 201 | DOS | 401 | auth | 101 |
| udp | 2 | REJ | 202 | Probe | 402 | bgp | 102 |
| icmp | 3 | RSTO | 203 | R2L | 403 | courier | 103 |
|  |  | RSTOS0 | 204 | U2R | 404 | csnet\_ns | 104 |
|  |  | RSTR | 205 | NORMAL | 405 | ctf | 105 |
|  |  | S0 | 206 |  |  | daytime | 106 |
|  |  | S1 | 207 |  |  | discard | 107 |
|  |  | S2 | 208 |  |  | domain | 108 |
|  |  | S3 | 209 |  |  | domain\_u | 109 |
|  |  | SF | 210 |  |  | echo | 110 |
|  |  | SH | 211 |  |  | eco\_i | 111 |
|  |  |  |  |  |  | ecr\_i | 112 |

3.4计算数据特征的概率值

数据特征包含着许多关于数据项的信息。数据类型的不同，则数据特征也表现出不同。对已经分类的数据矩阵进行概率统计，计算不同的数据特征的数值分布结果。以DOS类型数据为例，DOS类型数据的第1列的数据全为0，共829次，数值0出现的概率为1；第27列的数值类型共6种，出现次数分别为749,34,26,12,6,2共829次；出现概率为0.9035,0.0410,0.0314,0.0145,0.0072,0.0024概率之和为1。

表3-5 DOS类型数据特征值概率统计表

|  |  |  |
| --- | --- | --- |
| 数值类型 | 出现次数 | 出现概率 |
| 0 | 749 | 0.9035 |
| 0.01 | 34 | 0.0410 |
| 0.02 | 26 | 0.0314 |
| 0.03 | 12 | 0.0145 |
| 0.04 | 6 | 0.0072 |
| 0.05 | 2 | 0.0024 |
| 共计 | 829 |  |

3.4计算出数据特征的信息熵和信息密度

根据得到的数据特征的概率，可以求出数据的信息熵和信息密度。信息熵的计算公式为

 (3-1)

以DOS类型数据为例，DOS类型数据的第1列的数据全为0，数值0出现的概率为1，信息熵为0，信息熵之和为0；则第1个数据特征的信息熵为0。第27列的数值类型共6种，出现次数分别为749,34,26,12,6,2；出现概率为0.9035,0.0410,0.0314,0.0145,0.0072,0.0024；对应的信息熵为0.1323,0.1860,0.1567,0.0884,0.00515,0.0210，信息熵之和为0.6388；则第27个数据特征的信息熵为0.6388。

表3-6 DOS类型第27列数据特征值信息熵统计表

|  |  |  |  |
| --- | --- | --- | --- |
| 数值类型 | 出现概率 | 信息熵 | 第27特征的信息熵 |
| 0 | 0.9035 | 0.1323 | 0.6388 |
| 0.01 | 0.0410 | 0.1890 |
| 0.02 | 0.0314 | 0.1567 |
| 0.03 | 0.0145 | 0.0884 |
| 0.04 | 0.0072 | 0.0515 |
| 0.05 | 0.0024 | 0.0210 |

本文提出新的信息密度的概念，新的信息密度的定义是表示信息的分布密度即信息的不确定度。计算特征的信息密度需要先特征的信息熵，计算出信息密度的计算公式为

 (3-2)

表3-6 DOS类型第27列数据特征值信息密度统计表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 数值类型 | 出现概率 | 信息熵 | 第27特征的信息熵 | 第27特征的信息密度 |
| 0 | 0.9035 | 0.1323 | 0.6388 |  |
| 0.01 | 0.0410 | 0.1890 |  |
| 0.02 | 0.0314 | 0.1567 | 0.5279 |
| 0.03 | 0.0145 | 0.0884 |  |
| 0.04 | 0.0072 | 0.0515 |  |
| 0.05 | 0.0024 | 0.0210 |  |

表3-7全体数据特征值信息密度统计表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 信息密度  特征 | 401信息密度 | 402信息密度 | 403信息密度 | 404信息密度 | 405信息密度 |
| 1 | 1 | 0.8956 | 0.1111 | 0.0194 | 0.0734 |
| 2 | 1 | 0.4717 | 1 | 1 | 0.6302 |
| 3 | 1 | 0.332 | 0.3854 | 0.3472 | 0.6276 |
| 4 | 1 | 0.2506 | 0.9468 | 0.7904 | 0.9263 |
| 5 | 1 | 0.4296 | 0.018 | 0.0293 | 0 |
| 6 | 0.9758 | 0.8461 | 0.1982 | 0.0098 | 0.0036 |
| 7 | 1 | 1 | 1 | 1 | 1 |
| 8 | 0.9321 | 1 | 1 | 1 | 1 |
| 9 | 1 | 1 | 1 | 1 | 1 |
| 10 | 1 | 0.9865 | 0.2495 | 0.1552 | 0.9971 |
| 11 | 1 | 1 | 1 | 1 | 1 |
| 12 | 1 | 0.9656 | 1 | 1 | 0.6191 |
| 13 | 1 | 1 | 1 | 0.1305 | 0.9947 |
| 14 | 1 | 1 | 1 | 0.3695 | 1 |
| 15 | 1 | 1 | 1 | 1 | 0.9971 |
| 16 | 1 | 1 | 1 | 0.6261 | 0.9804 |
| 17 | 1 | 1 | 1 | 0.2737 | 0.9026 |
| 18 | 1 | 1 | 1 | 1 | 1 |
| 19 | 1 | 1 | 1 | 1 | 0.9141 |
| 20 | 1 | 1 | 1 | 1 | 1 |
| 21 | 1 | 1 | 1 | 1 | 1 |
| 22 | 1 | 1 | 0.4054 | 1 | 1 |
| 23 | 0.0008 | 0.0112 | 0.4126 | 0.2351 | 0.5541 |
| 24 | 0.0013 | 0.2311 | 0.3888 | 0.4094 | 0.1845 |
| 25 | 0.1983 | 0.0099 | 0.9622 | 0.6762 | 0.9542 |
| 26 | 1 | 0.4959 | 0.9492 | 1 | 0.9505 |
| 27 | 0.5279 | 0.0095 | 0.9622 | 0.6261 | 0.9697 |
| 28 | 1 | 0.3768 | 0.9754 | 0.7904 | 0.9538 |
| 29 | 0.1333 | 0.0977 | 0.9263 | 0.5369 | 0.7322 |
| 30 | 0.2234 | 0.1543 | 0.9263 | 0.4972 | 0.7363 |
| 31 | 1 | 1 | 0.8577 | 1 | 0.3346 |
| 32 | 0.2067 | 0.7516 | 0.0097 | 0.075 | 0.0007 |
| 33 | 0.0008 | 0.1993 | 0.0012 | 0.041 | 0.0005 |
| 34 | 0.0032 | 0.215 | 0.0316 | 1 | 0.0021 |
| 35 | 0.0179 | 0.0611 | 0.1478 | 1 | 0.0216 |
| 36 | 0.0032 | 0.1598 | 0.1744 | 0.1353 | 0.089 |
| 37 | 1 | 0.9865 | 0.0255 | 0.3033 | 0.1658 |
| 38 | 0.1794 | 0.0088 | 0.4659 | 1 | 0.6057 |
| 39 | 1 | 0.4894 | 0.4575 | 1 | 0.7836 |
| 40 | 0.0272 | 0.0034 | 0.4012 | 0.3158 | 0.4763 |
| 41 | 1 | 0.3768 | 0.8849 | 0.3158 | 0.5492 |

3.5计算信息密度修正算法

直接求解出来的信息密度值可能存在误差，通过对信息熵进行负指数变换可以得到信息密度值，但是数据特征值具有不确定性，可能导致信息熵存在细小的误差，进而使得信息密度值也有误差。那么为了解决这个问题，就引入了密度修正值，可以使信息密度值更为准确，具有客观性。使用梯度下降策略对信息密度值进行优化，在每次迭代后，这些信息密度都会被修正，并最终收敛到一个局部最优解。假定密度修正值的值可能设置是等可能的，梯度下降策略用于搜索更好的数据建模的密度修正值。这种策略是迭代的，它沿着准则函数的梯度的负方向（陡峭下降的方向）搜索解。要找到最大化的密度修正值的集合。

3.6测试数据验证数据模型

使用测试数据的数据特征，与每一个类型的数据模型进行匹配。得到测试数据的每个特征值在各类型的数据模型中出现的概率值。在将这些概率值与信息密度值相乘，得到测试数据经过数据模型处理后的结果。

 (3-3)

其中 表示结果， 表示数据类型， 表示特征出现的概率矩阵， 表示特征 在类型的数据模型中出现的概率值，表示特征 在类型的数据模型中信息密度值。最终测试数据会得到多个结果，比较多个结果并取最大的 。最大的所对应的类型，就是测试数据的分类结果。

3.7实验结果

使用ROC曲线[12]反映数据模型分类器的敏感性和特异性，揭示敏感性和特异性的相互关系，通过将连续变量设定出多个不同的临界值，从而计算出一系列敏感性和特异性，再以敏感性[13]为纵坐标、特异性[14]为横坐标绘制成曲线。混淆矩阵的用二元分类器为研究对象，下面的混淆矩阵显示了一个分类器可能遇到的所有情况描述混淆矩阵的每一列代表了预测类别，每一列的总数表示预测为该类别的数据的数目；每一行代表了数据的真实归属类别，每一行的数据总数表示该类别的数据实例的数目；表格中的百分比例表示测试数据的分类结果的准确率，误检率等。

表3-8混淆矩阵数据意义表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | 实际 | |  |
|  |  | 1 | 0 | 合计 |
| 预测 | 1 | True Positive（TP） | False Negative（FN） | P’=Actual Positive(TP+FN) |
| 0 | False Positive（FP) | True Negative(TN) | N’=Actual Negative(FP+TN) |
| 合计 |  | P=Predicted Positive(TP+FP) | N=Predicted Negative(FN+TN) | Total=TP+FP+FN+TN |

真正类率(true positive rate ,TPR), 刻画的是分类器所识别出的正实例占所有正实例的比例。计算公式为：

 (3-4)

负正类率(false positive rate, FPR),计算的是分类器错认为正类的负实例占所有负实例的比例。计算公式为

 (3-5)

真负类率（True Negative Rate，TNR），计算公式为

 (3-6)

混淆矩阵还有其他评价比率：

 其中  (3-7)

 (3-8)

 其中 (3-9)

 (3-10)

 (3-11)

 (3-12)

实验结果的ROC曲线和混淆矩阵：



DOS数据类型的分类结果ROC曲线图和混淆矩阵

6

Probe数据类型的分类结果ROC曲线和混淆矩阵



R2L数据类型的分类结果ROC曲线和混淆矩阵



U2L数据类型的分类结果ROC曲线和混淆矩阵



NORMAL数据类型的分类结果ROC曲线和混淆矩阵

根据数据模型的ROC曲线的特点，曲线下的面积大，分析出数据模型分类的准确度较高。在ROC曲线上，最靠近坐标图左上方的点为敏感性和特异性均较高的临界值。数据模型能够准确的检测不同的数据类型，真正类率、负正类率、分类准确率等评价指标都达到较好水平。DOS类数据的分类准确率为99.7%，DOS类数据被误判为非DOS类数据的比率为0.3%，非DOS类数据被误判为DOS类数据的比率为0%；Probe类数据的分类准确率为98.7%，Probe类数据被误判为非Probe类数据的比率为1.4%，非Probe类数据被误判为Probe类数据的比率为0%；R2L类数据的分类准确率为100%，R2L类数据被误判为非R2L类数据的比率为0%，非R2L类数据被误判为R2L类数据的比率为0%；U2L类数据的分类准确率为100%，U2L类数据被误判为非U2L类数据的比率为0%，非U2L类数据被误判为U2L类数据的比率为0%；NORMAL类数据的分类准确率为99%，NORMAL类数据被误判为非NORMAL类数据的比率为1.0%，非NORMAL类数据被误判为NORMAL类数据的比率为0%。

分别用本文提出基于信息密度的优化贝叶斯算法，神经网络算法，决策树算法，K-邻近算法处理数据，将四种算法的分类准确率进行对别得到

表3-9四种算法准确率结果表

|  |  |
| --- | --- |
| 分类算法 | 分类准确度 |
| 基于信息密度的优化贝叶斯算法 | 97.33% |
| 神经网络算法[18] | 97.00% |
| 决策树算法[19] | 98.64% |
| K-邻近算法[20] | 94.93% |

由上述表格可以得到，基于信息密度的优化贝叶斯算法对数据的分类准确度可以接受，能够较准确的实现对数据分类。

4.结论

云平台故障检测样本数据规模的大小，影响到算法的性能和检测率。为了更好地对样本数据进行分类，提出了一种基于信息密度的优化贝叶斯算法，采用信息熵描述数据特征的不确定性，信息密度作为数据特征分类的权重，并优化贝叶斯算法使其能够处理数据特征值没有变化的情况。通过对网络链接持续时间、网络协议类型，网络服务类型等云平台网络故障主要性能指标进行分析，实验表明，与改进前及其他相关算法相比，本文算法不仅有效减小信息损失而且抽取数据特征的内部关系，能够准确将云平台故障检测数据分类，可为后续的故障恢复工作提供了更好的保障，有利于提高故障恢复工作的效率。

参考文献

[1]岳逍远. 国家互联网应急中心发布“2011年中国互联网网络安全态势报告”[J].

信息网络安全, 2012, 04期:93-93.

[2]焦从信, 王崇骏, 陈世福. 基于完全无向图的贝叶斯分类器在入侵检测中的应用[J].

计算机科学, 2008, 09期:83-86. DOI:doi:10.3969/j.issn.1002-137X.2008.09.021.

[3]章倩, 周健. 基于贝叶斯的入侵检测模型与仿真研究[J]. 微电子学与计算机,

2013, 11期.

[4]文桥, 王卫平. 基于改进贝叶斯算法的入侵检测方法[J]. 计算机工程, 2006,

12期:160-162.DOI:doi:10.3969/j.issn.1000-3428.2006.12.061.

[5]潘志松. 基于神经网络的入侵检测研究[D]. 南京航空航天大学,

2003. DOI:doi:10.7666/d.y579036.

[6]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,

04期:798-803.

[7]石洪波, 王志海, 黄厚宽,等. 一种限定性的双层贝叶斯分类模型[J]. 软件学报,

2004, 第2期:193-199.

[8]周薇, 李筱菁. 基于信息熵理论的综合评价方法[J]. 科学技术与工程, 2010, 23

期:5839-5843. DOI:doi:10.3969/j.issn.1671-1815.2010.23.058.

[9]李太福, 熊隽迪. 基于梯度下降法的自适应模糊控制系统研究[J]. 系统仿真学报,

2007, 第6期:1265-1268. DOI:doi:10.3969/j.issn.1004-731X.2007.06.022.

[10]林龙涛. 高速网络环境下入侵检测系统研究[D]. 青岛大学, 2007.

DOI:doi:10.7666/d.y1096759.

[11]张新有, 曾华燊, 贾磊. 入侵检测数据集KDD CUP99研究[J]. 计算机工程与设计,

2010, 第22期:4809-4812.

[12]陈卫中, 倪宗瓒, 潘晓平,等. 用ROC曲线确定最佳临界点和可疑值范围[J]. 现代预

防医学, 2005, 07期:729-731. DOI:doi:10.3969/j.issn.1003-8507.2005.07.010.

[13]邹洪侠, 秦锋, 程泽凯,等. 二类分类器的ROC曲线生成算法[J]. 计算机技术与发展,

2009, 06期:109-112.DOI:doi:doi:10.3969/j.issn.1673-629X.2009.06.030.

[14]董元方. 机器学习中的模型选择问题研究[D]. 吉林大学, 2011.

[15]戴红，常子冠，于宁　主编. 数据挖掘导论[M] 清华大学出版社, 2014.

[16] 张良均, 杨坦, 肖刚 MATLAB数据分析与挖掘实战 机械工业出版社 2015

[17] 田景文. 人工神经网络算法研究及应用[M]// 北京理工大学出版社, 2006.

[18] 唐华松, 姚耀文. 数据挖掘中决策树算法的探讨[J]. 计算机应用研究, 2001, 08

期:18-19. DOI:doi:10.3969/j.issn.1001-3695.2001.08.006.

[19] 朱美玲, 陈勇明, 罗廷友. 基于k-最邻近分类法的灰色评估方法的改进[J]. 统计与决

策,2011, 19期:4-8.