基于最大信息熵模型的异常流量分类方法 *

钱亚冠[1], 关晓惠[2], 王 滨[3]
(1. 浙江科技学院 理学院, 杭州 310023; 2. 浙江水利水电高等专科学校 计算机工程系, 杭州 310018; 3. 浙江大学 计算机科学与技术学院, 杭州 310027)

摘 要: 最大信息熵原理已被成功地应用于各种自然语言处理领域,如机器翻译、语音识别和文本自动分类等,提出了将其应用于互联网异常流量的分类。由于最大信息熵模型利用二值特征函数来表达和处理符号特征,而 KDD99 数据集中存在多种连续型特征,因此采用基于信息熵的离散化方法对数据集进行预处理,并利用 CFS 算法选择合适的特征子集,形成训练数据集合。最后利用 BLVM 算法进行参数估计,得到满足最大熵约束的指数形式的概率模型。通过实验,比较了最大信息熵模型和 Naive Bayes、Bayes Net、SVM 与 C4.5 决策树方法之间的精度、召回率、F-Measure,发现最大信息熵模型具有良好的综合性能,尤其在训练数据集样本数量有限的情况下仍然能保持较高的分类精度,在实际应用中具有广阔的前景。

关键词: 最大信息熵模型; 异常流量; 离散化; 特征选择; 参数估计
中图分类号: TP393.08 **文献标志码**: A **文章编号**: 1001-3695(2012)03-1019-05
doi:10.3969/j.issn.1001-3695.2012.03.060

Anomalous traffic classification based on maximum entropy model

QIAN Ya-guan[1], GUAN Xiao-hui[2], WANG Bin[3]
(1. *College of Science, Zhejiang University of Science & Technology, Hangzhou* 310023, *China*; 2. *Dept. of Computer Engineering, Zhejiang Water Conservancy & Hydropower College, Hangzhou* 310018, *China*; 3. *College of Computer Science, Zhejiang University, Hangzhou* 310027, *China*)

Abstract: The machine learning model based on maximum entropy principles has been applied successfully in natural language processing, such as machine translation, text auto-classification and speech recognition. This model was first used in network anomalous traffic classification with our exploration. As the maximum entropy model used binary feature function, which was fit for processing nominal feature, it adopted the discrete method based on entropy to preprocessing the training data set. It generated the final feature set by extracting features from KDD99 dataset with CFS algorithm. Finally, employed the BLVM algorithm to evaluate the parameters and got an exponential model subjected to maximum entropy constrain. The model was compared with Naive Bayes, Bayes Net, SVM and C4.5 by precision, callback and F-Measure. The results of experiment show that the maximum entropy model has better classification efficiency, especially under small size of training data set.
Key words: maximum entropy model; anomalous traffic; discretezation; feature selection; parameter evaluation

0 引言

互联网应用已经渗透到各个领域,成为一个十分重要的基础设施,因此网络的安全性越来越受到各方面的重视。攻击者已经不局限于将主机作为攻击目标,开始将整个网络作为实施攻击的对象。尤其像 DDoS、网络蠕虫这样的攻击,它的危害已经不仅仅是终端主机的正常运行,而且会使网络大面积的瘫痪。作为网络运营商,及时地检测和发现异常网络流量,快速地采取相应的措施显得非常必要。

模式识别和机器学习是通过对数据的统计分析,选择合适的模型,对目标对象进行分类的学科。建立的模型称为分类器,每种分类器有相应的学习算法。利用实例数据不断地训练机器,让模型的各种参数趋于优化,从而使得模型具有更好的预测性能。利用模式识别和机器学习的方法,可以事先采集大量的流量数据,利用这些流量数据训练所选择的模型,从而使它具有预测各种流量类型的能力。本文的创新之处就是成功地将自然语言处理领域中应用广泛的最大信息熵模型应用到网络安全领域,并与多种成熟的机器学习方法进行了比较,包括分类精度、对训练集合大小的稳定性等。就目前所知,由于最大信息熵模型善于处理离散的符号特征,还没有论文利用最大信息熵模型来处理大量连续特征的网络流量的分类问题。

1 相关研究

Pietra 等人[1]在 1992 年将最大信息熵模型应用于自然语言处理之后,大量的研究围绕最大信息熵模型在该领域展开,并取得了令人瞩目的成果。最大信息熵模型在自然语言的应用领域方面包括机器翻译[2]、分本分类[3]、语音识别[4]等,但是没有看到其他领域的更多应用,尤其在网络领域的应用。最

收稿日期: 2011-07-27; **修回日期**: 2011-08-30 **基金项目**: 国家"973"计划基金资助项目(2007CB307102);国家科技支撑计划基金资助项目(2008BAH37B02)

作者简介:钱亚冠(1976-),男,浙江嵊州人,讲师,博士研究生,主要研究方向为互联网流量分析与建模(qianyg@yeah.net);关晓惠(1977-),女,讲师,硕士,主要研究方向为互联网流量测量;王滨(1978-),男,讲师,博士,主要研究方向为互联网安全.

大信息熵方法的两个基本任务是特征选择和模型选择。Berg 等人[2]提出用贪心策略,每次选择对数似然增益最大的特征来获得相应的特征集合。这种方法的缺点是可能无法找到全局最优解,而只能找到局部最优解;同时计算量太大。后来在此基础上又提出近似算法,但是计算量仍然很大。频率阈值法(cut-off)是一种简单而直观的方法,它将训练集合中出现次数很少的特征过滤掉。但这种方法的缺点是最终选择的特征之间存在很大的冗余,造成模型过于复杂,使得下一阶段的模型训练时间太长。

关于模型选择,即估计模型的参数 λ,Darroch 等人[5]最早提出通用迭代算法 GIS(general iterative scaling)来估计特征的权重。Berg 等人[2]又提出改进算法 IIS(improved iterative scaling),并从理论上证明比 GIS 算法速度快。但 IIS 算法的实际实现反而比 GIS 复杂,且占用的空间更多。Benson 等人[6]提出的 LMVM 方法是一类高效的拟牛顿算法。由于它不需要求 Hessian 矩阵,使得存储容量减少,搜索效率大大加快。

在网络流量异常检测方面,早期的方法非常简单、直观,利用流量的大小变化来发现异常[7],但这种方法对于流量变化不明显的异常很难检测得到。目前大量工作围绕机器学习方法展开。机器学习方法主要分两大类,即有监督和无监督学习。在网络流量异常检测中常用的基于监督分类方法有朴素 Bayes 方法[8]、Bayes 网络[9]、支持向量机[10,11]、决策树[12]等。无监督学习方法中通常使用聚类算法,如 K-means[13]、增量聚类[14],包括最近出现的基于信息熵方法[15]。除了机器学习方法外,也有相当多的工作围绕信号处理方法展开,如小波分析[16]。

2 最大信息熵模型

2.1 基本原理

在概率论领域存在两大学派,即客观概率和主观概率学派。客观概率学派强调系统的固有客观属性,认为概率是相同条件下重复实验的频率极限。而主管概率学派认为概率是观测者在已知有限知识的前提下对系统状态分布的信任程度。最大信息熵模型正是体现了主观概率的这种思想,它强调在有限知识下预测系统未知部分的概率分布时,应该选取符合这些已知知识且最随机的概率分布。而最随机的概率分布,即意味着信息熵值最大,因此将建立的随机模型称为最大信息熵模型。随着对系统的进一步认识,即获得的知识增多,相应的概率模型也作出调整,但始终保持对未知部分最随机分布这一假设。

本文将最大信息熵模型应用于网络异常流量的分类。网络流量类型用分类集合 Ω 表示,所有的 flow 构成的集合用 V 表示。每个 flow 由特征向量 $v=(\tau_1,\tau_2,\cdots,\tau_k)$ 表示,$v\in V$。训练样本实例由特征向量 $v\in V$ 和分类标记 $\omega\in\Omega$ 组成的序偶表示 $\langle v,\omega\rangle$,整个训练样本集合可表示为 $T=(\langle v_1,\omega_1\rangle,\langle v_2,\omega_2\rangle,\cdots,\langle v_i,\omega_i\rangle,\cdots,\langle v_N,\omega_N\rangle)$,这里的下标 i 仅用于区分不同的训练样本,有可能 $\langle v_i,\omega_i\rangle=\langle v_j,\omega_j\rangle(i\neq j)$。$V\times\Omega$ 构成一个事件空间,在这个事件空间上建立随机模型,来解决对未知 flow v' 的归类问题,即给出条件概率 $p(\omega|v')(\omega\in\Omega)$。

即使有一个很大的训练样本集合,也往往无法提供所有可能的样本 $\langle v,\omega\rangle$ 的信息,但能给出部分样本的经验概率值。在符合已知信息的情况下,去除人为的主观假设,满足无偏性的要求,使未知事件的概率分布尽可能均匀,即信息熵(条件信息熵)最大。

为了表示已知信息,引入特征函数的概念:$f:\langle v,\omega\rangle\to\{0,1\}$,在最大信息熵模型中也简称特征。给定训练集合 T,可得到实例 $\langle v,\omega\rangle$ 的经验概率分布 $\tilde{p}(v,\omega)=\dfrac{\mathrm{freq}(v,\omega)}{N}$。特征函数 k 在经验分布 T 下的经验期望就是模型要重现的统计量:$E_{\tilde{p}}f=\sum_{v,\omega}\tilde{p}(v,\omega)f(v,\omega)$。特征函数的模型期望:$E_p f=\sum_{v,\omega}p(v,\omega)f(v,\omega)$。通常可以定义一组特征函数 $F=\{f_i\}$。为了使模型与训练样本一致,引入一组矩约束(由于特征函数是两值函数,该矩约束实际上表达的是该特征的概率要在经验模型上和模型保持一致):

$$E_p f_i=E_{\tilde{p}}f_i \tag{1}$$

满足上述约束的概率分布不止一个,而最优解就是分布最均匀的那个,因此可将信息熵函数作为目标函数:

$$H(p)=-\sum_{v,\omega}\tilde{p}(v)p(\omega|v)\log p(\omega|v) \tag{2}$$

$$p^*=\underset{p\in P}{\operatorname{argmax}}\,H(p) \tag{3}$$

使式(2)信息熵最大的解即为最优解。应用拉格朗日乘数法求解,可得指数形式的解为

$$p^*(\omega|v)=\frac{1}{Z(v)}\exp\Big(\sum_i\lambda_i f_i(v,\omega)\Big) \tag{4}$$

其中:$Z(v)=\sum_{\omega}\exp\Big(\sum_i\lambda_i f_i(v,\omega)\Big)$ 为归一化因子。λ_i 是需要确定的参数,将在 2.4 节说明如何估计 λ_i。

2.2 数据预处理

由于 KDD99 数据集中存在大量连续型属性,如发送的分组数量。最大信息熵模型适合处理离散的符号型特征,因此需要对这些连续型属性离散化。本文采用文献[17]中提出的基于信息熵的特征离散化方法。该方法在对特征值离散的过程中,充分考虑了分类标签对离散区间的影响,因此也被称为有监督的离散化方法。通过实验比较多种离散化方法后,证明该方法目前是最有效的。

定义 1 设训练样本集合 S 内有 k 个分类标签 $C_1,\cdots,C_k$,$P(C_i,S)$ 表示 C_i 在集合 S 内的经验概率,那么集合 S 的分类熵(class entropy)可表示为

$$H(S)=-\sum_{i=1}^{k}P(C_i,S)\log_2(P(C_i,S)) \tag{5}$$

设 T 是分割点,它将已按升序排列的特征 A 分割为两个区间,对应的训练样本集合 S 也被分为两个子集 S_1、S_2。

定义 2 由 T 对训练样本集合 S 的分割而引起的分类信息熵(class information entropy)表示为

$$H(A,T;S)=\frac{|S_1|}{|S|}H(S_1)+\frac{|S_2|}{|S|}H(S_2) \tag{6}$$

因此选择的分割点就是使 $H(A,T;S)$ 最小的 T:$T_{\min}=\underset{T}{\operatorname{argmin}}\,H(A,T;S)$。文献[18]已经证明满足上述条件的分割点必定落在两个不同分类之间,这样可使分割点数量大大减少。对子集 S_1、S_2 可用上述方法继续递归分割,最终形成不同的分割区间集合。根据 MDL(minimal description length)原则[19],递归结束的条件为

$$\text{gain}(A,T;S) < \frac{\log_2(N-1)}{N} + \frac{\Delta(A,T;S)}{N} \tag{7}$$

其中:N 是集合 S 中的样本数;

$$\text{gain}(A,T;S) = H(S) - H(A,T;S) \tag{8}$$

$$\Delta(A,T;S) = \log_2(3^k - 2) - [k \times H(S) - k_1 H(S_1) - k_2 H(S_2)] \tag{9}$$

其中:k_i 是 S_i 中的分类标签数。

2.3　特征选择

特征选择是从特征空间中选择一个合适的特征空间子集,去除不相关和冗余特征。本文采用基于相关的过滤器(correlation-based filter,CFS)[20]进行特征选择。CFS 与其他特征选择方法相比具有更好的分类正确率和计算效率。条件信息熵可以用来衡量特征与分类及特征与特征之间的相关程度。

设 $H(X)$ 是特征 X 的信息熵,$H(X|Y)$ 表示其他特征或类 Y 出现时的条件信息熵,信息增益可用来定义 X 与 Y 之间的相关程度为

$$C(X|Y) = \frac{H(X) - H(X|Y)}{H(X) + H(Y)} \tag{10}$$

一个特征子集的合理性可通过如下指标来衡量:

$$G_{\text{subnet}} = \frac{k\bar{r}_{ci}}{\sqrt{k + k(k-1)\bar{r}_{ii}}} \tag{11}$$

其中:k 是特征子集的特征数量,

$$\bar{r}_{ci} = \sum_{X \in \text{subnet}} P(X|Y) \cdot C(X|Y) \tag{12}$$

$$\bar{r}_{ii} = \sum_{X,Y \in \text{subnet}} P(X|Y) \cdot C(X|Y) \tag{13}$$

最后以 G_{subnet} 的值为依据,采用最佳优先(best first)搜索方法得到需要的特征子集。

2.4　参数估计

为了确定最大信息熵模型,必须通过训练样本估计出模型参数 λ_i。用 $\varphi(\lambda)$ 表示拉格朗日函数的最大值,λ 是一个参数向量,$\lambda = (\lambda_1, \lambda_2, \cdots, \lambda_k)$。

$$\varphi(\lambda) = -\sum_v \tilde{p}(v)\log Z(v) + \sum_i \lambda_i E_{\tilde{p}}(f_i) \tag{14}$$

满足 $\lambda^* = \arg\max_{\lambda} \varphi(\lambda)$ 的 λ^* 就是估计出的参数。本文采用 LMVM(limited memory variable metric)算法来求解这个无约束优化问题。它的关键步骤就是用一个近似矩阵来代替 Hessian 矩阵,同时采用 BFGS 修正式,使搜索方向近似牛顿方向。

2.5　模型算法描述

假设已经获得了预处理后的训练数据,可通过下述的算法框架获得最大信息熵模型。

输入:流量训练数据 $T = (\langle v_1, \omega_1\rangle, \langle v_2, \omega_2\rangle, \cdots, \langle v_i, \omega_i\rangle, \cdots, \langle v_N, \omega_N\rangle)$

输出:条件概率模型 p^* 和参数 λ_i

1)初始化数据

从训练数据中获得经验概率分布 $\tilde{p}(v,\omega) = \frac{\text{freq}(v,\omega)}{N}$。

2)特征选择

根据 CFS 算法和最佳优先搜索方法获得特征函数集合 $F = \{f_1, f_2, \cdots, f_K\}$。

3)参数估计

a)对每个特征 f_i;

b)线性搜索算法确定步长 α^k,$\lambda_i^{k+1} = \lambda_i^k + \alpha^k d^k$;

c)计算目标函数的梯度 $G(\lambda_i^{k+1})$ 及其投影 $T_\Omega G(\lambda_i^{k+1})$;

d)若满足终止条件,获得参数 λ_i,$i \leftarrow i+1$,转 a);

e)调用 BGFS 算法[6],转 b)。

最后获得概率模型 $p^*(\omega|v) = \frac{1}{Z(v)}\exp(\sum_i \lambda_i f_i(v,\omega))$。

3　实验数据

KDD99 源于 MIT 林肯实验室收集的网络流量数据,对正常连接和各种异常连接都进行了标记,目前被广泛用于入侵检测方面的研究实验。连接类型共有五类,分别为正常(normal)、DoS、U2R(user to root attacks)、R2L(remote to user attacks)、扫描(probe)。KDD99 数据中的各种连接类型的分布如表 1 所示。

表 1　KDD99 10% 训练数据集连接类型的分布

class	number of records	% of occurrence
normal	97 278	19.69
DoS	391 458	79.24
U2R	52	0.01
R2L	1 126	0.23
probe	4 107	0.83
total	494 021	100.00

由于 KDD99 10% 的数据量过于巨大,本文在这个基础上又进行了一次随机抽取,获得用于本文实验的训练数据集,如表 2 所示。

表 2　本文实验训练数据集连接类型的分布

class	number of records	% of occurrence
normal	9 841	19.84
DoS	39 092	78.82
U2R	13	0.03
R2L	213	0.43
probe	437	0.88
total	49 596	100.00

4　实验分析

用 Java 语言实现了最大信息熵模型,朴素 Bayes、Bayes 网络、SVM 和 C4.5 等算法则利用开源的机器学习软件 Weka(Java)实现。实验用的环境是安装有 Linux 的双核 Intel Core2 PC 机,主频 1.66 GHz,配有内存 2 GB。

4.1　评价指标

1)精度(precision)　某类被正确归类的实例数占分类器分类到该类的所有实例的比例。

2)召回率(recall)　某个类中被正确归类的实例数占该类应该被正确归类的总数的比例。

3)F-Measure　采用调和平均的方法将 precision 和 recall 的同时考虑进去的一个综合指标:2 × precision × recall/(precision + recall)。

4.2　不同的机器学习方法的比较

4.2.1　精度、召回率和 F-Measure

为了确定最大信息熵模型的分类能力,分别选择朴素 Bayes 方法、Bayes 网络、SVM 和决策数 C4.5 算法进行比较分析。将 KDD99 1% 的数据集随机抽取 10%、30%、50%、70%、

90% 用于训练,从 KDD99 提供的测试集中随机抽取 15 000 条记录用于测试。

在 normal 分类上,如图 1 所示,五种方法在精度、召回率和 F-Measure 三个指标上非常接近,而且均在 95% 以上。在 DoS 分类上,如图 2 所示,五种方法在三个指标上均在 98% 以上。在这两个分类上很难区分五种方法的优劣,最大信息熵模型几乎接近 100%。在 U2R 分类上,如图 3 所示,五种方法却表现出显著的差异,在精度上,SVM 最好,当训练数据超过 30% 后精度接近 100%,其次为最大信息熵模型,在 50% 以上;而在召回率上,最大信息熵模型最好,当训练数据超过 30% 后接近 100%,其次为 Bayes 网络,仅在 70% 的训练数据上时召回率为 50%,其他大小的数据集上与最大信息熵模型相当。在 F-Measure 这个综合指标上,最大信息熵模型得分最高。分析在 U2R 上五种方法差异明显的原因,从训练数据集的角度看,由于 U2R 的记录数非常少,仅占 0.03%。在如此少的数据量时,最大信息熵模型仍然保持良好的分类性能,这点是值得肯定的。在 R2L 和 Probe 这两个类上,如图 3、4 所示,最大信息熵模型、SVM 和 C4.5 在精度上比较相似,均可保持 95% 以上的分类精度。在召回率上,随着训练数据集的增大,五种方法不同程度地提高了召回率,当训练数据集超过 70% 时,五种方法都在 90% 以上。但最大信息熵模型始终处于前列,并保持良好的稳定性。所以在 F-Measure 指标上,最大信息熵模型略优于 SVM 和 C4.5。

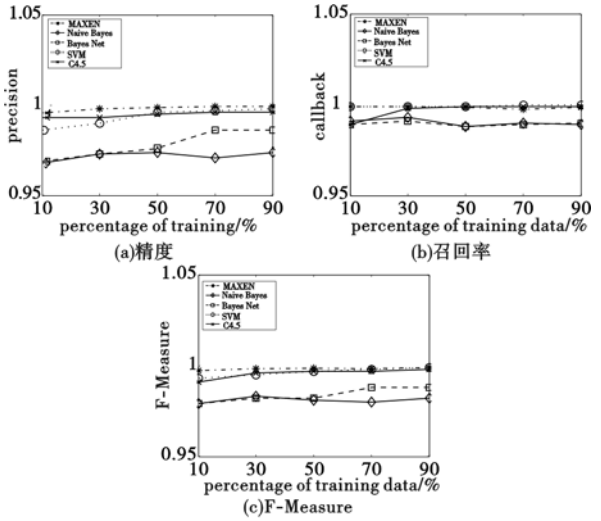


图1　不同大小数据集上，五种分类方法在normal上的精度、召回率和F-Measure

4.2.2　不同大小的训练数据集合影响

训练数据集的大小对于模型训练的时间有直接关系,一般来说训练数据越多,训练的时间就越长。人们总是希望在小的数据集上就能训练得到性能较好的模型。如图 1、2 所示,五种方法在 Normal 和 DoS 类上,随着训练集合的增大,分类性能并没有明显的变化。在 U2R 上,如图 3 所示,朴素 Bayes 和 Bayes Net 反而出现随着训练集合的增大,分类精度下降,但召回率在超过 30% 的数据集时趋于稳定;最大信息熵模型的精度在 50% 的数据集上比 30% 时要小,这可能与 U2R 的记录数太少有关系。在 R2L 上,如图 4 所示,最大信息熵模型、SVM 和 C4.5 的精度都很稳定,召回率上最大信息熵模型和 C4.5 比 SVM 更稳定。在 probe 上,如图 5 所示,最大信息熵模型和 SVM 的精度比其他方法更稳定。在召回率上,最大信息熵模型、朴素 Bayes 和 Bayes Net 比 SVM 和 C4.5 更稳定。综合在各个分类上的表现,最大信息熵模型比其他方法的稳定性更好。

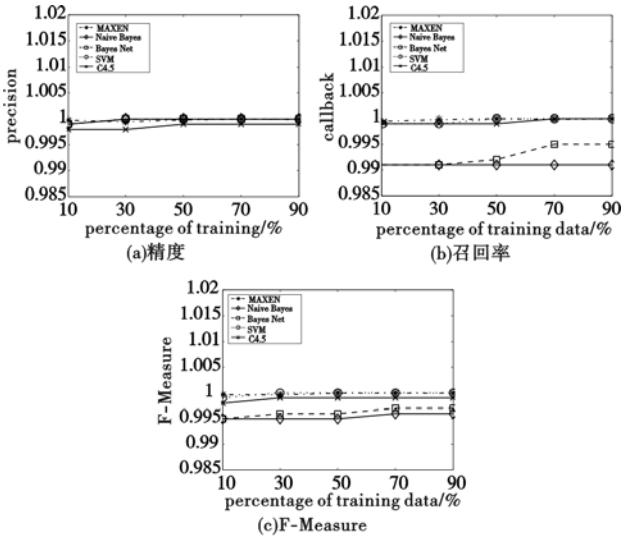


图2　不同大小数据集上，五种分类方法在DoS上的精度、召回率和F-Measure

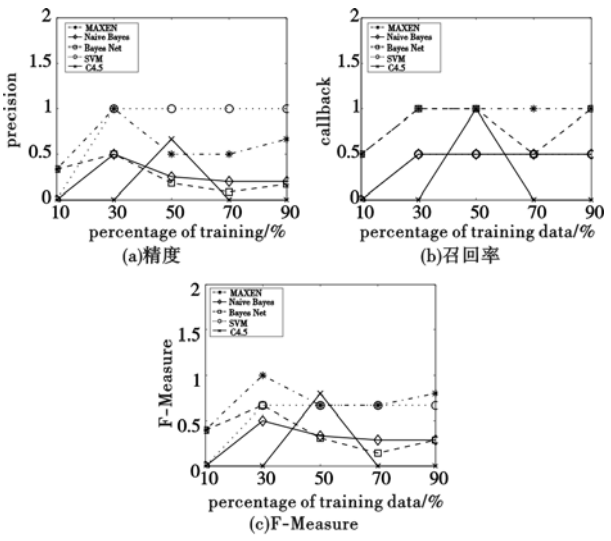


图3　不同大小数据集上，五种分类方法在U2R上的精度、召回率和F-Measure

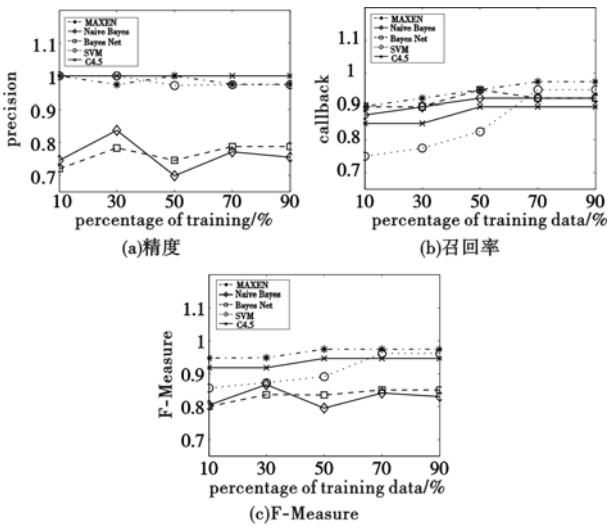
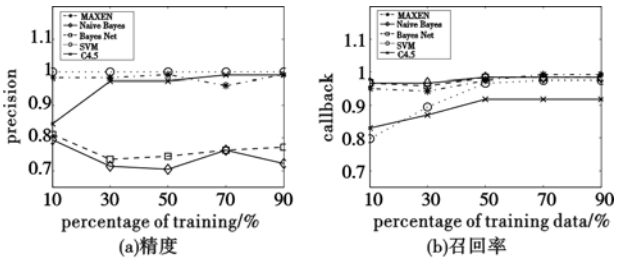


图4　不同大小数据集上，五种分类方法在R2L上的精度、召回率和F-Measure

4.2.3　模型训练时间的比较

如图 6 所示,随着训练数据集合的增大,模型的训练时间都会增加。朴素 Bayes 和 Bayes 网络的训练时间是最少的。SVM

的训练时间是最长的,90% 的数据集需要的训练时间大概为 43 s。最大信息熵模型的训练时间介于 SVM 和 C4.5 之间,90% 的数据集大概需要 5.4 s 的时间,总体上是可以接受的。



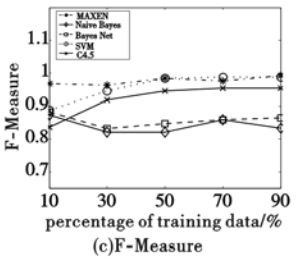


图5　不同大小数据集上，五种分类方法在Probe上的精度、召回率和F-Measure

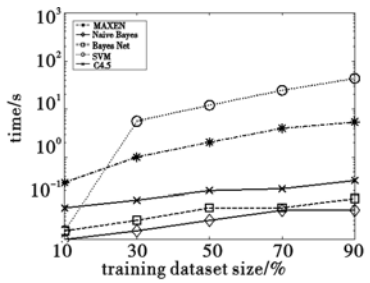


图6　最大信息熵模型与其他方法在训练时间上的比较

4.2.4　实验分析小结

最大信息熵模型的分类性能、稳定性略好于 SVM,比 C4.5 要好,远远高出朴素 Bayes、Bayes 网络。在模型训练时间上,比 SVM 少 10 倍的时间,但比 C4.5 要多。综合各方面的因素,最大信息熵模型是一个非常不错的分类器。最大信息熵模型的训练时间主要消耗在参数估计上,将来继续要改进的地方就是缩短参数估计的运行时间。

5　结束语

利用机器学习的方法对异常流量进行分类是一种很有价值的方法。最大信息熵模型建立在主观概率理论的基础上,其核心的观点就是对事件的未知因素不作任何人为的主观假设,从而避免模型的偏倚性。本文给出了利用 KDD99 流量数据构建最大信息熵模型的完整过程。通过实验,比较了它与其他几种主流的机器学习算法,可以发现最大信息熵模型具有良好的分类性能,具有很好的实用意义。

参考文献:

[1] PIETRA S D, PIETRA V D, MERCER R L, *et al*. Adaptive language modeling using minimum descriminant estimation[C]//Proc of IEEE International Conference on Acoustic, Speech and Signal Processing. 1992:633-636.

[2] BERGE A L, PIETRA S D, PIETRA V D. A maximum entropy approach to natural language processing[J]. Computational Linguistic, 1996, 22(1):39-71.

[3] NIGAM K, LAFFERTY J, MCCALLUM A. Using maximum entropy for text classification[C]//Proc of IJCAI'99 Workshop on Machine Learning for Information Filtering. 1999:61-67.

[4] KHUDANPUR S, WU J. A maximum entropy language model integrating n-grams and topic dependencies for conversational speech recognition[C]//Proc of the IEEE International Conference on Acoustics, Speech and Signal Processing. 1999:553-556.

[5] DARROCH J N, RATCLIFF D. Generalized iterative scaling for log-linear models[J]. Annals of Mathematical Statistics, 1972, 43(5):1470-1480.

[6] BENSON S J, MORE J J. A limited-memory variable-metric method for bound-constrained minimization[R]. [S. l.]: Mathematics and Computer Science Division, Argonne National Laboratory, 2001.

[7] BARFORD P, KLINE J, PLONKA D, *et al*. A signal analysis of network traffic anomalies[C]//Proc of ACM SIGCOMM Internet Measurement Workshop. 2002: 71-82.

[8] PEDRO D, PAZZANI M. On the optimality of the simple Bayesian classifier under zero-one loss[J]. Machine Learning, 1997, 29(2-3):103-137.

[9] ZHANG N L, POOLE D. A simple approach to Bayesian network computations[C]//Proc of the 10th Biennial Canadian Artificial Intelligence Conference. 1994:171-178.

[10] GARDNER A B, KRIEGER A, VACHTSEVANOS G, *et al*. One-class novelty detection for seizure analysis from intracranial EEG[J]. Journal of Machine Learning Research, 2006, 7:1025-1044.

[11] AIZERMAN M, BRAVERMAN E, ROZONOER L. Theoretical foundations of the potential function method in pattern recognition learning[J]. Automation and Remote Control, 1964, 25:821-837.

[12] QUINLAN J R. Improved use of continuous attributes in C4.5[J]. Journal of Artificial Intelligence Research, 1996, 4:77-90.

[13] MÜNZ G, LI S, CARLE G. Traffic anomaly detection using K-means clustering[C]//Proc of Leistungs-, Zuverlässigkeits- und Verlässlichkeitsbewertung von Kommunikationsnetzen und Verteilten Systemenv. 2007.

[14] CAN F. Incremental clustering for dynamic information processing[J]. ACM Trans on Information System, 1993, 11(2):143-164.

[15] LAKHINA A, CROVELLA M, DIOT C. Diagnosing network-wide traffic anomalies[C]//Proc of SIGCOMM. 2004.

[16] HUANG C T, THAREJA S, SHIN Y J. Wavelet-based real time detection of network traffic anomalies[C]//Proc of Workshop on Enterprise Network Security and the 2nd International Conference on Security and Privacy in Communication Networks. 2006:1-7.

[17] FAYYAD U M. On the induction of decision trees for multiple concept learning[D]. Ann Arbor: University of Michigan, 1991.

[18] FAYYAD U M. On the induction of decision trees for multiple concept learning[D]. [S. l.]: EECS Dept, University of Michigan, 1991.

[19] QUINLAN J R, RIVEST R L. Inferring decision trees using the minimum description length principle[J]. Information and Computation, 1989, 80(3):227-248.

[20] HAMILTON N Z. HALL M. Correlation-based feature selection for machine learning[D]. [S. l.]: Department of Computer Science, Waikato University, 1998.

[21] KDD cup 1999 data[EB/OL]. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.