基于主题模型的网络异常行为分类学习方法研究

马钲然 张博锋 王勇军

(国防科学技术大学计算机学院 长沙 410072)

摘 要 提出了一种新的用于学习和分辨网络异常行为的方法。与之前的工作相比,将采用主题模型对网络异常行为进行建模并构建分类器。根据连接的分类标签,在训练模型之前将数据集分成两部分,即正常的部分和异常的部分。通过分析模型参数对结果的影响可以发现α(主题的狄利克雷参数)和主题数量对于预测结果具有正相关性,而β(特征号的狄利克雷参数)对于预测结果具有负相关性。通过KDDCUP'99数据集对该模型进行评估,结果显示预测的准确度达到91.69%,比SVM等算法在正常和异常行为分类上的表现更好。

关键词 主题模型,异常行为,分类器

中图法分类号 TP309 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.9.010

Research on Studying Method of Network Anomalous Behaviors Classification Based on Topic Model

MA Zheng-ran ZHANG Bo-feng WANG Yong-jun

(School of Computer, National University of Defense Technology, Changsha 410072, China)

Abstract A novel approach to learn and identify the anomalous behaviors in network was proposed. Unlike previous work, the intrusion detection problem is mapped into the topic model and a classifier is built. Two kinds of connections, namely normal and anomalous ones, are separated before training the model according to the labels of the connections. By analyzing the effect of the parameters, it shows that α (Dirichlet parameter of topics) and the number of topics have positive correlation with the results of prediction, while β (Dirichlet parameter of feature numbers) has negative correlation with the results of prediction. This model was evaluated using KDDCUP'99 dataset. The result suggests that the prediction accuracy is up to 91.69% which outperforms SVM algorithm in normal and anomalous behaviors classification.

Keywords Topic model, Anomalous behavior, Classifier

1 介绍

如今,大部分入侵检测系统都以不同的程度发展[1],然而主要方向是通过基于特征检测的手段提高对已知威胁的检测。对于未知威胁,基于异常行为的检测方法则并没有广泛使用在产业界。

就我们所知,所有的威胁行为都是由一系列的网络连接组成。除此以外,我们可以非常容易地获取这些连接并得到这些连接的一些具有重要含义的特征。例如,可以利用Bro等IDS对网络中的流量进行特征提取,通过系统监视器对主机进行特征提取。因此,分辨网络流量的异常与否,对于发现网络中的异常行为是非常有帮助的。

本文提出了一种新颖的学习和辨别网络连接类型的方法。该方法将每一个连接通过一个特征号表示,每一个特征号都是一个主题的一部分,且这个主题会帮助我们构建一个分类器来识别一个网络连接是正常还是异常。将该模型命名为基于网络的LDA,即NLDA。其主要优势在于主题可以反映一个连接异常与否的趋势或倾向,每个主题中特征号的概率帮助我们对新的连接进行分类。

本文的主要工作包括以下3个方面:

(1)将LDA模型应用于有标签的数据集上,并将其变量含义映射到NLDA模型中。

(2)基于LDA模型,提供了一个可以对没有标签的数据集进行结果预测的算法。

(3)通过KDDCUP'99数据集对提出的模型和算法进行评估,实验结果显示所提模型比SVM等算法在对网络连接的正常和异常行为分类的预测结果的准确率上要高。

本文第2节对相关工作进行简明的介绍;第3节对模型及相关的算法进行详细阐释;第4节通过实验对模型进行评估;最后总结全文。

2 相关工作

在过去的十年里,网络世界出现了各种各样的网络威胁行为。更糟糕的是,它们之中很大一部分都没有被及时地发现。一方面,未知类型的威胁行为很难被发现,因为对于新型威胁行为来说没有现成的检测规则或特征;另一方面,异常检

到稿日期:2015-08-20 返修日期:2015-10-28 本文受国家自然科学基金项目(61472439,61303264,61271252)资助。

马钲然(1991-),男,硕士,主要研究方向为网络安全、数据挖掘,E-mail:zhengranma@163.com;**张博锋**(1978-),男,博士,副研究员,主要研究方向为网络安全、数据挖掘;**王勇军**(1971-),男,博士,研究员,主要研究方向为网络安全、信息安全。

测在产业界并不成熟，并且大部分算法还没有应用到更广泛的场景。

LDA(Latent Dirichlet Allocation)[2,3]被 D. Blei 在 2003 年首次提出，并被用来寻找一系列文档中的主题或者观点信息。该方法的主要特征在于研究人员不需要对文档中相关领域的背景知识的了解。不仅仅是文本建模，LDA 还可以应用到其他的相关领域，如图像识别处理等[4]。此外，一些作者也将 LDA 应用到网络安全建模和异常网络行为检测。

Cramer 等人首次提出将 LDA 模型应用到网络领域[5]。Benjamin D. Newton 等人则将 LDA 应用到网络异常行为检测中[6]。Huang 等人[7]认为超级计算机系统中的检测设备和系统自身产生的日志可以用于 LDA 做知识发现等应用。他们同时提出了文本建模和入侵行为检测问题的映射关系。但遗憾的是，他们并没有将其文中所提的方法应用到实际中。Bhavesh Kasliwal 等人提出了一种使用 LDA 和遗传算法来检测异常网络行为的方法 G-LDA，但是其准确率不高，因为该方法忽略了很多网络连接中相关的特征[8]。

LDA 对于网络流分类的主要优势在于其变量适用于 LDA 模型，尽管 LDA 是一个关注于文本建模研究的模型。下面将给出 LDA 模型和网络连接中的变量之间的映射关系。

3 模型

对网络中的每一条连接进行建模，将其视作一系列特征的集合。每一个特征都由一个特征号表示，所有的特征号组成一个巨大的“词汇表”，称为“特征号库”。我们的目标是将网络中的正常和异常行为进行分类。

图 1 是对所提算法的总体概括，包括学习和识别过程。该算法基于 LDA 模型，主要的不同在于根据训练数据集的标签进行训练，并且利用生成的主题对没有标签的测试数据集进行分类。称该模型为 NLDA，其含义为用于网络安全的 LDA 模型。

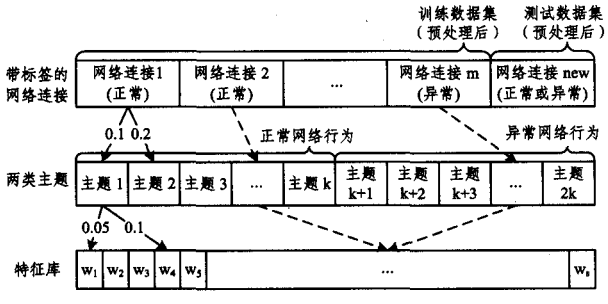


图 1 算法框架

3.1 基本模型结构

NLDA 模型是基于 LDA 模型的，它们之间唯一的不同在于其采用的变量所代表的物理意义之间的差异。为了更好地解释该模型，以一个连接的产生作为例子。对于一个异常连接行为，在生成特征库中的特征号时，抽取一个能决定哪一个主题被选中的概率向量。对于网络连接中的每一个特征号，首先选取一个可能的主题，然后抽取一个特征库中的特征号。重复上述抽取主题和其对应的特征号的过程，最后得到一个特征号的“袋子”，从而组成一个网络连接。图 2 展示了基本模型的概率图形式。

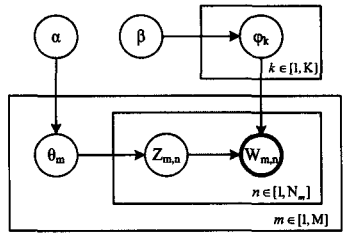


图 2 网络连接分类的主题模型

3.1.1 表示过程

图 2 展示了主题模型的基本结构，下面说明该主题模型的概念和定义。由于模型主要被用于文本建模研究，因此首要任务就是对其相关的概念与网络异常行为检测找出映射关系。在所提模型中，一个最基本的单位即一个网络连接，是由一个 N 维的向量组成，向量中的元素即是网络连接的一系列基本特征。

• 一个特征号 w 是一个网络连接的基本元素，其数据类型是离散型，所有的特征号可以组成一个“特征库”。模型中的特征号就如同 LDA 中的词语。

• 一个网络连接是一个 N 维特征号组成的序列，可以表示成 $\boldsymbol{w}=(w_1,w_2,\cdots,w_N)$，其中，$w_n$ 是该网络连接中的第 n 个特征号。模型中的网络连接如同 LDA 中的文档。

• 一个特征库 $\boldsymbol{W}=\{w_1,w_2,\cdots,w_S\}$ 是所有的网络连接中的特征号组成的集合。模型中的特征库如 LDA 中的语料库。

根据上面定义的概念以及同 LDA 的映射关系，可以描述模型中生成一个网络连接的详细过程。将这个过程分成以下两个阶段。

$\boldsymbol{\alpha}\rightarrow\boldsymbol{\theta}\rightarrow z$：该过程代表当第 $\boldsymbol{m}$ 个网络连接产生后，首先抽取一个能够决定主题分布的参数，即对每一个网络连接抽取一个 $\boldsymbol{\theta}\sim p(\boldsymbol{\theta}|\boldsymbol{\alpha})$。$\boldsymbol{\theta}$ 是选择主题的多项式分布参数。$\boldsymbol{\alpha}$ 是 K 维的狄利克雷分布参数。K 是主题的总数。之后可以得到第 m 个网络连接中第 n 个特征号的主题号 $z_{m,n}$。

$\boldsymbol{\beta}\rightarrow\boldsymbol{\varphi}\rightarrow w|z$：该过程描述了网络连接中的特征号是如何产生的。同第一个过程一样，首先选择能够决定特征号分布的参数 $\boldsymbol{\varphi}\sim p(\boldsymbol{\varphi}|\boldsymbol{\beta})$。$\boldsymbol{\varphi}$ 是多项式分布的参数，而 $\boldsymbol{\beta}$ 作为狄利克雷分布参数，是一个 V 维向量。V 是特征库中特征的总数。之后，对于网络连接中的每一个特征号 w，有以下两个步骤。

选择一个主题 $z_{m,n}\sim Mult(\boldsymbol{\theta})$，其中 $z_{m,n}$ 是一个单位向量，$z_{m,n}=1$ 代表这个主题被抽中，假设它的主题号为 k。选择一个特征号 $w_{m,n}\sim p(w_{m,n}|z_{m,n},\boldsymbol{\varphi})$。$w_{m,n}$ 是第 m 个网络连接中的第 n 个特征号，其与 $z_{m,n}$ 是相对应的。

3.1.2 推理过程

根据表示过程，可以将模型的完全生成等式描述如下，这同时也是变量 $\boldsymbol{w},\boldsymbol{z},\boldsymbol{\theta}$ 和 $\boldsymbol{\varphi}$ 的联合概率分布的贝叶斯展开。

$$p(\boldsymbol{w}_m,\boldsymbol{z}_m,\boldsymbol{\theta}_m,\boldsymbol{\Phi}|\boldsymbol{\alpha},\boldsymbol{\beta})=\prod_{n=1}^{N_m}p(w_{m,n}|\boldsymbol{\varphi}_{z_{m,n}})p(z_{m,n}|\boldsymbol{\theta}_m)\cdot p(\boldsymbol{\theta}_m|\boldsymbol{\alpha})\cdot p(\boldsymbol{\Phi}|\boldsymbol{\beta}) \quad (1)$$

接着，可以得到一个初始特征号的概率表达。

$$p(\boldsymbol{w}_m=t|\boldsymbol{\theta}_m,\boldsymbol{\Phi})=\sum_{k=1}^{K}p(w_{m,n}=t|\boldsymbol{\varphi}_k)p(z_{m,n}=k|\boldsymbol{\theta}_m) \quad (2)$$

整个网络连接集合的似然函数可以表示为如下形式。

$$p(\boldsymbol{W}|\boldsymbol{\Theta},\boldsymbol{\Phi})=\prod_{m=1}^{M}p(\boldsymbol{w}_m=t|\boldsymbol{\theta}_m,\boldsymbol{\Phi})$$
$$=\prod_{m=1}^{M}\prod_{n=1}^{N_m}p(w_{m,n}|\boldsymbol{\theta}_m,\boldsymbol{\Phi}) \quad (3)$$

可将联合概率分布改写成如下形式：

$$p(\boldsymbol{w},\boldsymbol{z}|\boldsymbol{\alpha},\boldsymbol{\beta})=p(\boldsymbol{w}|\boldsymbol{z},\boldsymbol{\beta})p(\boldsymbol{z}|\boldsymbol{\alpha}) \quad (4)$$

由于多项式分布和狄利克雷分布是共轭结构，可以利用这一性质将式(4)改写成以下的形式。这种形式对于接下来的学习过程是必不可少的。

$$p(\boldsymbol{w},\boldsymbol{z}|\boldsymbol{\alpha},\boldsymbol{\beta})=\prod_{z=1}^{K}\frac{\Delta(\boldsymbol{n}_z+\boldsymbol{\beta})}{\Delta(\boldsymbol{\beta})}\cdot\prod_{m=1}^{M}\frac{\Delta(\boldsymbol{n}_m+\boldsymbol{\alpha})}{\Delta(\boldsymbol{\alpha})} \quad (5)$$

3.1.3 学习过程

我们现在可以得到吉布斯采样的规则，即根据其他特征号(除了当前特征号)的主题分布来计算当前特征号的主题概率公式。

$$p(z_i=k|\boldsymbol{z}_{\neg i},\boldsymbol{w})\propto p(z_i=k,w_i=t|\boldsymbol{z}_{\neg i},\boldsymbol{w}_{\neg i})=\theta_{m,k}\cdot\varphi_{k,t} \quad (6)$$

可以得到两个多项式分布的参数，$i=(m,n)$代表一个二维索引，其含义为第 m 个网络连接中的第 n 个特征号的二维索引。$n_{k,\neg i}^{(t)}$代表特征号 t 被分配到除了第 i 个主题 k 的数量。$n_{m,\neg i}^{(k)}$代表主题 k 被分配到除了第 i 个网络连接 m 的数量。

$$\varphi_{k,t}=\frac{n_{k,\neg i}^{(t)}+\beta_t}{\sum_{t=1}^{V}n_{k,\neg i}^{(t)}+\beta_t} \quad (7)$$

$$\theta_{m,k}=\frac{n_{m,\neg i}^{(k)}+\alpha_t}{\sum_{k=1}^{K}n_{m,\neg i}^{(k)}+\alpha_t} \quad (8)$$

可以发现式(6)中的右边部分就是 $p(topic|connection)\cdot p(feature\ number|topic)$，而这个概率即是 $connection\rightarrow topic\rightarrow feature\ number$ 的路径概率。对这 K 条路径进行采样(由于 topic 有 K 个)就是吉布斯采样的物理意义。

3.2 训练和分类方法

在 3.1 节得到了一个 LDA 模型，从而更加接近文中的两个主要的目标：1)对模型中的参数 $\boldsymbol{\varphi}_1,\cdots,\boldsymbol{\varphi}_K$ 和 $\boldsymbol{\theta}_1,\cdots,\boldsymbol{\theta}_M$ 进行估计；2)计算新的网络连接中的主题分布参数 $\boldsymbol{\theta}_{new}$。

为了解决以上两个问题，分别采用训练算法和分类算法。

3.2.1 训练算法

训练过程的目的是通过吉布斯采样得到特征库中$(\boldsymbol{z},\boldsymbol{w})$的样本，并且根据得到的样本估计所有模型中的参数。

算法 1 训练算法

1. foreach connection c_m do
2. foreach feature number $w_{m,n}$ do
3. randomly sample topic index $z_{m,n}$ from k~Mult(1/K)
4. $n_m^{(k)}+=1;n_m+=1;n_k^{(t)}+=1;n_k+=1;$
5. foreach iteration till convergence do
6. foreach connection c_m do
7. foreach feature number $w_{m,n}$ do
8. $n_m^{(k)}-=1;n_m-=1;n_k^{(t)}-=1;n_k-=1;$
9. randomly sample topic index $k\sim p(z_i|z_{\neg i},w)$
10. $n_m^{(k)}+=1;n_m+=1;n_k^{(t)}+=1;n_k+=1;$
11. readout parameter set Φ and Θ according to Eq. 7 and Eq. 8

3.2.2 分类算法

在对模型进行训练后，可以得到一个分类器。当得到一个新的网络连接时，可以像在训练过程中一样计算主题分布。对于每一个新的网络连接，认为吉布斯采用中的 $\varphi_{k,t}$ 部分是固定不变的。换句话说，$\varphi_{k,t}$是由训练模型提供得到的。因此只需要关注对新网络连接中的 θ_{new} 分布进行估计。

算法 2 分类算法

1. foreach feature number $w_{new,n}$ in new connection do
2. randomly sample topic index $z_{new,n}$ from k~Mult(1/K)
3. $n_{new}^{(k)}+=1;n_{new}+=1;n_k^{(t)}+=1;n_k+=1;$
4. foreach iteration till convergence do
5. foreach feature number $w_{new,n}$ in new connection do
6. $n_{new}^{(k)}-=1;n_{new}-=1;n_k^{(t)}-=1;n_k-=1;$
7. randomly sample topic index $k\sim p(z_i|z_{\neg i},w)$
8. $n_{new}^{(k)}+=1;n_{new}+=1;n_k^{(t)}+=1;n_k+=1;$
9. readout parameter set θ_{new} according to Eq. 8

4 实验

采用 KDDCUP'99 数据集[9]来训练提出的模型，并提出了 3 个主要的步骤。首先，对数据集中的数据进行预处理，将其转换为 LDA 模型可用的形式；其次，用准备好的数据对模型进行训练；最后，对网络连接中的正常和异常行为进行分类并评估模型中的参数。图 3 展示了实验的整体框架。

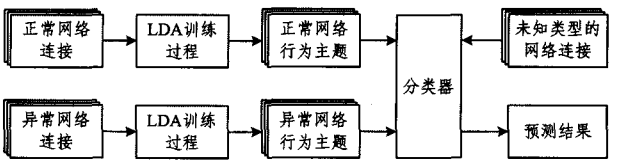


图 3 实验基本框架

4.1 数据预处理

在模型训练之前，需要对数据进行预处理，因为原始数据集并没有标准化。对数据集的标准化处理(包括对训练数据集和对测试数据集的处理)主要采取以下几个步骤。

原始数据集中的某些特征的数据类型是 string 类型，因此需要对它们的值进行编号。例如特征 *protocol_type* 代表协议类型，包括 3 个可能的值 TCP，UDP 和 ICMP，分别将其编号为 0，1 和 2。

尽管特征的数量并没有约减并且部分特征对于网络连接类型的区分并不能起到关键作用，但是我们并不会利用如 PCA 或其他相关技术来对数据集进行降维处理。原因在于网络连接中的每个特征都如同文档中的一个词语，这个词语对于每一个连接或多或少都会有着贡献值。此外，具备更少的特征的网络连接在使用 LDA 模型时，产生的主题有较弱的说服力。

考虑到特征包含不同的数据类型，包括离散和连续类型，我们首先对数据集做离散化处理。离散化的方法是将特征的数值分割到不同的区间并对其编号。例如，代表着文件生成的数量的特征 *num_file_creations* 是一个连续型变量，将其数值分割到 101 个区间然后为其依次编号为 0 到 100。

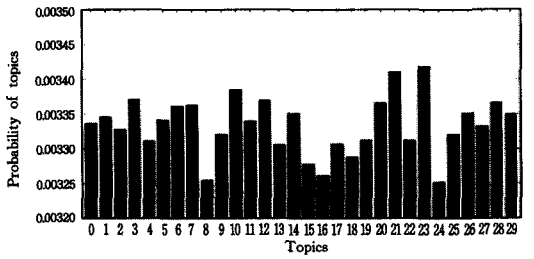
为了构建特征号组成的特征库，必须让每一个特征号都有所区别。因此对每一个特征号提供一个新的独特的数值。例如对第一个特征 *duration* 进行离散化处理后得到 201 个区间并对其编号为 0 到 200。对于第二个特征 *protocol_type*，对其数值编号则为 201，202 和 203。从而可以得到所有的特征号组成的集合，即特征库，共包括 3375 个特征号。

4.2 训练过程

在所提模型中，根据标签把训练集分割成两个部分。一部分是正常网络行为，而另一部分是异常网络行为。之后，将其分别应用于 LDA 模型中，从而得到两组关于不同网络行为

的主题，分别是正常和异常网络行为主题。

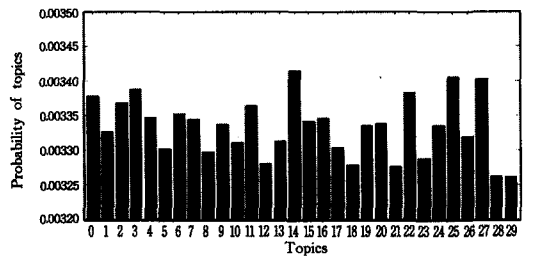
图 4(a)展示了正常网络行为分类中的主题分布。以 21 号和 23 号主题作为例子。图 4(b)分别展示了 21 号和 23 号主题中的前 3 个特征号以及它们的含义。可以发现，主题中的特征号所代表的含义充分说明了正常网络连接的特征。图 4(c)和图 4(d)展示了异常网络行为中的情形。



(a)正常主题分布

0:网络连接的持续时间不超过 1 秒钟	1145:没有访客登录过程发现
921:没有试图使用 root 账户的命令请求	1141:FTP 出站会话的连接数目为 0
689:没有错误的碎片	285:从源地址到目的地址的字节数少于 1024
主题 21	主题 23

(b)前 3 个正常特征号



(c)异常主题分布

1126:Shell 命令被使用	3276:存在“REJ”错误
0:网络连连接的持续时间不超过 1 秒钟	818:存在有危险的条件
922:尝试使用“suroot”命令	2670:在过去的 100 个连接中存在的相同主机不同端口的连接
Tpoic 14	Tpoic 25

(d)前 3 个异常特征号

图 4　正常、异常网络连接主题示例及其特征示例

4.3　测试过程

在训练过程中得到了整个特征号的集合，我们认为这个集合是完整的，因为特征是确定的。训练过程之后，得到的分类器可以用来对测试数据集进行预测。使用 10000 条训练数据在不同的参数下构建不同的分类器。为了评估参数和最终预测结果的准确率，将对以下情况进行实验。

参数：首先固定训练集中的训练数据的数量为 10000，针对 3 个参数做 3 组实验。由于有 3 个参数需要评估，因此在每一组中固定其中一个参数，如表 1 所列。

表 1　参数测试的 3 组实验

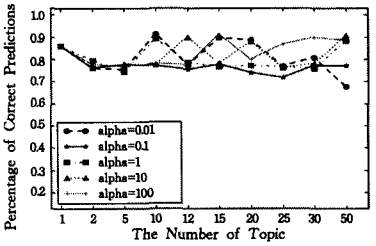
参数	α	β	主题数
第 1 组	0.01,0.1,1,10,100	0.01	[1,50]
第 2 组	[0.01,100]	0.01,0.1,1,10,100	10
第 3 组	10	[0.01,100]	1,5,10,20,50

训练集：训练集中测试数据的数量可能是最影响预测结果的一个因素，因此固定上面的 3 个参数(主题数＝10，α＝10，β＝1)，然后对不同规模的训练集进行测试。

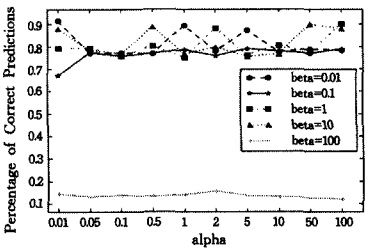
与 SVM 的比较：分别使用 NLDA 和 SVM 对 10000 个训练数据进行建模并构建分类器，之后对 300000 个测试样本进行分类，对最终得到的结果进行比较分析。

4.4　实验结果

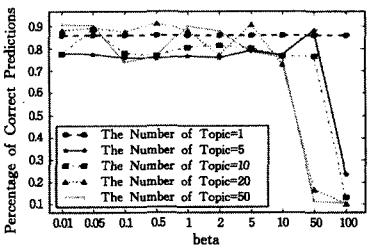
通过 20000 条测试数据分析了模型参数对结果的影响。图 5 示出了实验结果。



(a)当 β=0.01 时，在不同 α 和主题数的取值下，预测准确率的结果比较



(b)当主题数为 10 时，在不同 α 和 β 的取值下，预测准确率的结果比较



(c)当 α=10 时，在不同 β 和主题数的取值下，预测准确率的结果比较

图 5　模型参数对预测准确率的影响

对于 α，从图 5(a)中可以看出，随着主题数量的增加，α 越大，预测结果的准确率越高。但是在主题数量很小的情况下，α 对预测结果的准确率的影响则非常小。

对于 β，从图 5(b)中可以看出，β 越大，尤其在 β 的值为 100 时，预测结果的准确率越低。图 5(c)中剧烈的下滑部分也说明了这一点。

在主题数量方面，从图 5 中可以很明显地看出，主题数量越大，预测结果的准确性也越高，但同时其稳定性越低；主题数量越少，预测结果的准确性也越低，但稳定性越高。因此，并不需要高数值的主题数量，但是一定数量的主题数仍然是必须的。

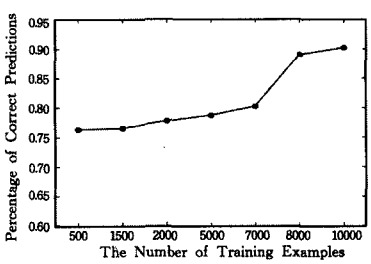


图 6　当主题数为 10，α=10，β=1 时，不同训练样本数对预测准确率的影响

从图 6 可知，训练数据集的数据样本数量对预测结果的

(下转第 81 页)

图的关系网络不仅能帮助我们清晰地理解事件发展趋势，还能根据关系网络有向图的连通性勾画出话题的主题性，运用基于主题的 PAGERANK 计算各节点的相对重要度，从而得到各子主题在整个话题中的重要程度，并在关系元组的重要程度排序的基础上按照时间要素先后进行排序。最后逐步删除排序了的语句集合中对信息贡献最小的子句，直到剩余的句子长度之和达到目标文摘长度。实验证明这种方法能更全面、更简洁地概括出文本的主要内容。

参 考 文 献

[1] Harabagiu S, Hickl A. Relevance modeling for microblog summarization [C]// Fifth International AAAI Conference on Weblogs and Social Media. 2011

[2] Long R, Wang H, Chen Y, et al. Towards effective event detection, tracking and summarization on microblog data[M]// Web-Age Information Management. Springer Berlin Heidelberg, 2011: 652-663

[3] Zhao W X, Jiang J, He J, et al. Topical keyphrase extraction from twitter[C]// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011: 379-388

[4] Zhong Z, Liu Z. Ranking events based on event relation graph for a single document [J]. Information Technology Journal, 2010, 9(1): 174-178

[5] Guo Xi-yue, He Ting-ting, Hu Xiao-hua, et al. Chinese Named Entity Relation Extraction Based on Syntactic and Semantic Features[J]. Jounal of Chinese Informatio Processing, 2014, 28(6): 183-189(in Chinese)
郭喜跃，何婷婷，胡小华，等. 基于句法语义特征的中文实体关系抽取[J]. 中文信息学报，2014，28(6)：183-189

[6] Zhao W X, Jiang J, Weng J, et al. Comparing twitter and traditional media using topic models[M]// Advances in Information Retrieval. Springer Berlin Heidelberg, 2011: 338-349

[7] Chen Dan-qi, Manning C D. A Fast and Accurate Dependency Parser using Neural Networks[C]// Proceedings of EMNLP 2014. 2014

[8] Nenkova A, Maskey S, Liu Y. Automatic summarization[C]// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts of ACL 2011. Association for Computational Linguistics, 2011

[9] Zhao W X, Jiang J, He J, et al. Topical keyphrase extraction from twitter[C]// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011: 379-388

(上接第 60 页)

准确率的影响是显而易见的，数据样本越多，则预测结果的准确率越高。

从与 SVM 算法的比较结果来看，NLDA 的预测准确率基本上维持在 90%左右。最终在 300000 个测试数据上进行测试，得到 91.69%的准确率，比 SVM 的 89.64%有明显提升，结果如图 7 所示。

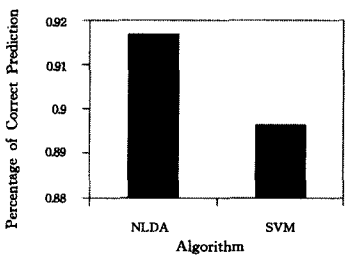


图 7 NLDA 和 SVM 算法之间的比较

结束语 本文提出了一个用于学习和识别网络异常行为分类的主题模型。该模型是对网络安全的一个应用并且在 KDDCUP'99 测试集上的效果显著。本文主要的创新点在于使用主题模型对网络行为进行分类并在参数分析的基础上提高了识别的准确率。

在接下来的工作中，将在单机和分布式环境下做进一步的实验。考虑使用 Spark 和 Petuum 平台作为分布式环境。之后，将对网络流量如何转换为网络特征做进一步研究。

参 考 文 献

[1] Garcia-Teodoro P, Diaz-Verdejo J, Macia-Fernandez G, et al. Anomaly-based network intrusion detection: Techniques, systems and challenges[J]. Computers & Security, 2009, 28(1/2): 18-28

[2] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. J. Mach. Learn. Res., 2003, 3: 993-1022

[3] Blei D M. Probabilistic topic models[J]. Commun. ACM, 2012, 55(4): 77-84

[4] Fei-Fei L, Perona P. A bayesian hierarchical model for learning natural scene categories[C]// IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005(CVPR 2005). IEEE, 2005, 2: 524-531

[5] Cramer, Christopher, Carin L. Bayesian topic models for describing computer network behaviors[C]//2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2011: 1888-1891

[6] Newton B D. Anomaly Detection in Network Traffic Traces Using Latent Dirichlet Allocation[OL]. http://www.cs.unc.edu/~bn/BenNetwonFinalProjectReport.pdf

[7] Huang J, Kalbarczyk Z, Nicol D M. Knowledge Discovery from Big Data for Intrusion Detection Using LDA[C]// 2014 IEEE International Congress on Big Data (BigData Congress). IEEE, 2014: 760-761

[8] Kasliwal B, Bhatia S, Saini S, et al. A hybrid anomaly detection model using G-LDA[C]//2014 IEEE International Advance Computing Conference (IACC). IEEE, 2014: 288-293

[9] http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html