第35卷 第12期 计 算 机 工 程 2009年6月
Vol.35 No.12 Computer Engineering June 2009
·网络与通信· 文章编号：1000—3428(2009)12—0090—02 文献标识码：A 中图分类号：TP393

基于半监督学习的网络流量分类

佘 锋，王小玲
(中南大学信息科学与工程学院，长沙 410083)

摘 要：利用攻击在网络通信中独特的流特征，给出一个可以适应已知和未知攻击的半监督分类方法。在训练分类器中，提出使用加权采样技术得到训练流，同时采用顺序前向选择算法得到最佳的特征子集。使用KDD CUP1999性能评估数据，可以得到较高的流和字节分类准确度。

关键词：网络流量分类；半监督学习；模糊C均值；入侵检测

Network Traffic Classification Based on Semi-supervised Learning

SHE Feng, WANG Xiao-ling
(School of Information Science and Engineering, Central South University, Changsha 410083)

【Abstract】 This paper exploits distinctive flow characteristics of attacks when they communicate on a network, and proposes a semi-supervised classification method that can accommodate both known and unknown attacks. In training the classifier, it employs Sequential Forward Selection(SFS) to get the best feature subset. Meanwhile, it proposes weighted sampling techniques to obtain training flows. Performance evaluation using KDD CUP1999 data shows that high flow and byte classification accuracy can be achieved.

【Key words】 network traffic classification; semi-supervised learning; Fuzzy C-Means(FCM); intrusion detection

1 概述

网络流量分类是众多网络活动的基础，根据具体类型动态识别和分类流量的能力非常关键，如网络运营商可以找出异常网络流量以减轻恶意行为的影响。

最近，Moore[1]等人采用基于Naive Bayes分类的监督学习方法分类网络流量。然而，监督分类面临2个主要挑战。首先，标记的例子稀缺且难以获取，传统监督学习方法用少量已标记的例子产生的分类器，往往不能归纳以前没出现的流类型。其次，并非所有攻击的流量类型是事先所知的，并且随着时间推移新攻击类型的流量可能出现。

本文针对上述问题进行研究，提出一个不同于传统半监督学习应用的结合监督和非监督的方法。

2 半监督网络流量分类原理

设$\boldsymbol{X}=\{X_1,X_2,\cdots,X_N\}$是一个流的集合；$X_i=\{X_{ij}\,|\,1\leqslant j\leqslant m\}$，其中，$m$是属性数量，并且$X_{ij}$是第$i$个流的第$j$个属性值。另外，设$\boldsymbol{Y}=\{Y_1,Y_2,\cdots,Y_q\}$为一个分类集，其中，$q$是分类数目。因此，目标是研究从$m$维向量$\boldsymbol{X}$到$\boldsymbol{Y}$的映射。该映射形成分类模型的基础，也被称为分类器。表1列出了分类网络流量的例子。

表1 根据所属范畴分类网络流量的例子

分类	实例
DoS(Denial of Service)	Mailbomb, Teardrop, Apache2
R2L(Remote to Local)	ftp_write, Xlock
U2R(User to Root)	buffer_overflow
Probing	ipsweep, nmap, satan
Data	secret

半监督分类方法包括2个步骤，首先采用聚类方法处理由少量标记流和大量未标记流构成的训练数据集。然后利用现有的标记流量获得从簇到已知不同的q类映射，这一步也允许一些簇未能映射，因为可能流中没有已知类型的标记。学习结果是一个簇的集合，其中有些映射到不同的流类型。这种分类方法被称为半监督学习[2]。

2.1 聚类方法

首先利用标记和未标记的流进行聚类训练分类器。聚类的好处是有能力确定隐藏类型，如可以通过检查形成的新簇识别新攻击类型以及改变行为的现有攻击。不失一般性，本文采用欧氏距离作为属性向量$\boldsymbol{x}_i$和$\boldsymbol{y}_j$之间的相似性度量：

$$d(\boldsymbol{x}_i,\boldsymbol{x}_j)=\left[\sum_{k=1}^{m}(x_{ik}-x_{jk})^2\right]^{1/2} \tag{1}$$

在机器学习领域有许多不同的聚类算法，早期研究中调查了3种聚类算法：FCM，EM和DBSCAN聚类。其中，FCM算法是非常有效的一种非监督模糊聚类方法。即使对于很难明显分类的变量，FCM算法也能得到较为满意的效果，并且它简单且容易实现，算法复杂度低，故本文采用FCM算法。

2.2 簇到分类的映射

FCM算法输出一个簇集，由聚类中心代表γ_k。给出流属性向量$\boldsymbol{x}$，通过找到距离$\boldsymbol{x}$最近的中心C_k将它分配到其中一个簇，C_k由式(2)得到：

$$C_k=\arg\min_k d(\boldsymbol{x},\gamma_k) \tag{2}$$

其中，$d(\boldsymbol{x},\gamma_k)$由式(1)得到。

使用一个概率分布发现从簇到分类的映射：$P(\boldsymbol{Y}=\boldsymbol{y}_j\,|\,C_k)$，其中，$j=1,2,\cdots,q$（$q$是分类类型数量），$k=1,2,\cdots,c$（$c$是聚类数目）。使用在训练数据中被标记为不同类型的流集合$(\boldsymbol{x}_i,y_i)$估计这些概率，$i=1,2,\cdots,L$，其中，L是被标记为不同的分类类型总数。$P(\boldsymbol{Y}=\boldsymbol{y}_j\,|\,C_k)$由极大似然

基金项目：国家自然科学基金资助项目(60773013)
作者简介：佘 锋(1983－)，男，硕士研究生，主研方向：网络安全，模糊聚类；王小玲，教授
收稿日期：2009-01-12 **E-mail**：symphony.sf@msn.com

估计估算，n_{jk}/n_k，n_{jk}是被分配到簇 k 标签为 j 的流数量，且 n_k 是被分配到簇 k(已标记的)的流总数量。完成映射后，那些没有包含任何标记实例的簇被定义成未知类型，因而允许对以前未识别的类型的表示。分类流 $\boldsymbol{x}$ 的决策函数 y 为

$$y = \arg \max_{y_1,y_2,\cdots,y_q} (P(\boldsymbol{y}_j \mid C_k)) \tag{3}$$

其中，C_k 是到 $\boldsymbol{x}$ 最近的簇，由式(2)可得。根据 $P(\boldsymbol{y}_j \mid C_k)$ 得到流标签的置信度，标签置信度低于一个给定阈值可被认为“未知类型”。

3 半监督分类实现方法

实现方法使用 KDD CUP1999 数据集，采用 SFS 算法从原始数据集中得到一个精简的特征子集用于 FCM 聚类，选定 FCM 算法中聚类数目 c 和模糊指数 m，并使用抽样方法解决分类不平衡问题。

3.1 数据集

实验选用的样本数据是入侵检测领域比较权威的 KDD CUP1999 测试数据[3]，这是一个已审核过的标准数据集。

攻击数据分为 5 大类，包括：拒绝服务(DoS)，远程主机非授权访问(R2L)，普通用户进行本地超级用户的非授权访问(U2R)，侦查和探测(Probing)和数据传输攻击(Data)。每条记录共包括 41 种定性和定量的特征属性。

3.2 特征子集选择

为了从原有庞大数据集中获得一个精简数据集合，并保持原有数据集的完整性，采用 SFS(Sequential Forward Selection)算法搜索特征子集，从中选择一些重要的相关特征。SFS 算法是一个贪婪选择过程，从一个空的特征子集开始，每次添加一个特征，并且加入的特征具有最大度量值。SFS 算法不能保证获得最优解决办法，但它的复杂性为 $O(d^2)$，定义如下：

$$S_w = \sum_{j=1}^{k} \pi_j E\{(\boldsymbol{X}-\boldsymbol{\mu}_j)(\boldsymbol{X}-\boldsymbol{\mu}_j)^{\mathrm{T}} \mid \omega_j\} = \sum_{j=1}^{k} \pi_j \boldsymbol{\Sigma}_j \tag{4}$$

$$S_b = \sum_{j=1}^{k} \pi_j (\boldsymbol{\mu}_j - M_o)(\boldsymbol{\mu}_j - M_o)^{\mathrm{T}} \tag{5}$$

$$M_o = E\{\boldsymbol{X}\} = \sum_{j=1}^{k} \pi_j \boldsymbol{\mu}_j \tag{6}$$

其中，$\boldsymbol{X}$ 是一个表示流的 d 维属性向量；π_j 是一个流属于簇 ω_j 的概率；k 是聚类数；$\boldsymbol{\mu}_j$ 是簇 ω_j 的样本均值向量；M_o 是总样本均值；$\boldsymbol{\Sigma}_j$ 是簇 ω_j 的样本协方差矩阵；$E\{\bullet\}$ 是期望值的算符。由式(4)、式(5)得到一个数据集的类可分性 S，定义：$S = trace(S_w^{-1} S_b)$。较高的可分性度量 S 值，确保类较好地被分离。由于特征子集选择准则涉及到维数，因此本文采用规格化标准值[4]。

3.3 聚类数目 c 和模糊指数 m 选择

模糊指数 m 是一个控制算法的柔性参数，Pal 等人[5]从聚类有效性方面的实验研究中得到 m 最佳取值区间为 $(1.5,2.5)$，在不做特殊要求下可取区间中值 2，这里取 $m=2$。为了避免在聚类数目 c 选取上的主观性，提高 FCM 算法实用性，把聚类有效性函数引入到 FCM 算法中。对于给定的聚类中心数 c，待分析的样本个数 n 和隶属度矩阵 $\boldsymbol{U}$，定义聚类有效性函数 $FP(\boldsymbol{U};c)$ 为

$$FP(\boldsymbol{U};c) = \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} \mu_{ij}^{\ 2} - \frac{1}{c}\sum_{i=1}^{c}(\sum_{j=1}^{n} \mu_{ij}^{\ 2} / \sum_{j=1}^{n} \mu_{ij}) \tag{7}$$

记 Ω_c 为所有划分矩阵 $\boldsymbol{U}$ 的“最优”有限集，对于 $\boldsymbol{U} \in \Omega_c$，若存在 $(U^*;c^*)$ 满足式(8)，则 $(U^*;c^*)$ 为“最优”有效性聚类。

$$FP(U^*;c^*) = \min_{c}\{\min_{\Omega_c} \mathrm{FP}(\boldsymbol{U};c)\} \tag{8}$$

给定 $m=2$ 的情况下，根据聚类有效性函数 $FP(\boldsymbol{U};c)$，可以给出 FCM 算法对聚类数目 c 优选的自适应算法：对不同聚类数目值，在 c 未达到最优时，$FP(\boldsymbol{U};c)$ 随 c 的增加而减小，当 c 到达最优值而继续增大时，$FP(\boldsymbol{U};c)$ 将会由最小值而增大，因此，选 $FP(\boldsymbol{U};c)$ 随 c 的增加而成为最小点的值作为最佳聚类数 c^*。

3.4 长流和短流

T. Mori 和 M. Uchida[6]记录了在网络流量中出现的长流(elephant flows)和短流(mice flows)现象。根据这一现象，大部分网络流量是小规模的短流，只有一小部分是大中型的长流，不过，短流只占一小部分包及网络传送的字节比率。如果在训练数据集中没有对这 2 种流进行适当处理，那么开发的分类器可能会有一些不足，例如，高流量精度低字节精度。这个问题被称为分类的不平衡问题。

针对分类不平衡问题，解决方案采用顺序和随机抽样技术。对于顺序抽样，每个数据集随机选择一个点开始流的顺序选择。至于随机抽样包括简单随机抽样，还考虑加权随机抽样，即根据一个流传输的字节或流的持续时间有重点地选择样本。这种加权方案可以让更多分类更好地代表大的流量。

4 实验结果

为评估基于 FCM 的半监督分类系统性能，并实现最优配置，实验采用的半监督分类系统框架，如图 1 所示。实验在 KDD 99 数据集中随机选取 10 组数据，每组数据各 10 万条记录。每组数据均满足检测算法假设，即正常行为数目远远大于入侵行为数目，每组使用约 10%的标记数据。迭代停止阈值 $\varepsilon = 10^{-5}$。

图 1 基于 FCM 的半监督分类系统框架

实验有 3 个部分：第 1 个实验集目的是采用 SFS 算法搜索特征子集；第 2 个实验集评估聚类数目 c 优选；第 3 个实验集评价采样技术对分类正确率的影响以及分类器的检测率和误警率百分比。

在第 1 个实验集中，发现一些在不同试验中被选择得最多的属性子集，最终选择的 17 个流量属性如下：duration, protocol_type, service, src_bytes, dst_bytes, flag, logged_in, num_access_files, is_hot_login, is_guest_login, count, cerror_rate, rerror_rate, srv_count, srv_serror_rate, srv_rerror_rate, srv_diff_host_rate。实验使用这个属性集作为分类器的基础。

在第 2 个实验集中，采用聚类数目 c 优选的自适应算法，得到了 10 组数据集最佳聚类数目，具体值见表 2。

表 2 各组数据集最佳的聚类数目

组次	最佳聚类数目	组次	最佳聚类数目
1	5	6	8
2	6	7	6
3	6	8	6
4	8	9	6
5	7	10	7

在第 3 个实验集中，从图 2 显示的结果，观察到顺序采样超过 86%的高流量正确率，但相应的字节正确率较低。结果还表明，顺序抽样字节正确率与随机和加权抽样技术相比很差。加权字节抽样技术达到了最佳字节分类正确率，利用

(下转第 94 页)

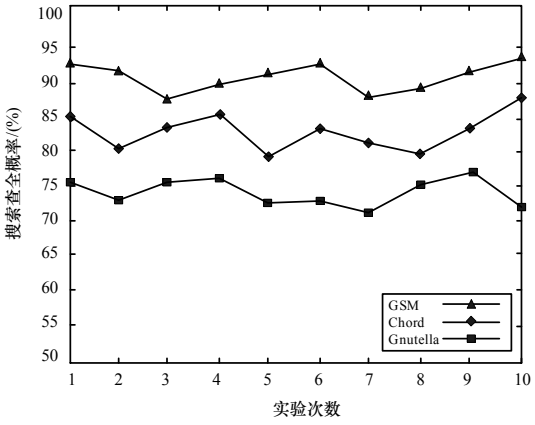

图 4　搜索查全率

本文还针对 GSM, Chord 和 Gnutella 在查询过程中经过的逻辑跳数分布情况进行实验，结果如图 5 所示。

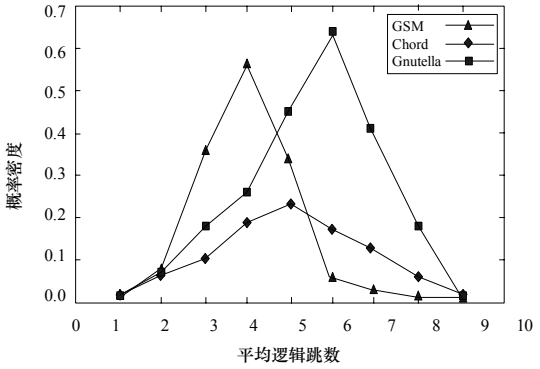

图 5　逻辑跳数分布

可见，GSM 利用余弦相似度分组，搜索在组内获得结果的概率极大，组内不存在时，通过组成员向其余分组洪泛查询的形式保证了查全率，整个查全率在 90%左右，相比 Chord 的 85%左右和 Gnutella 的 75%左右要优越；GSM 查找成功平均经过跳数为 4 跳左右，Chord 的平均跳数为 5 跳左右，而 Gnutella 平均跳数在 6 跳左右，GSM 能更快地查找到目标，具有更好的查找性能。

4　结束语

本文提出一种基于余弦相似度的 P2P 分组搜索机制，根据标引词权重向量的余弦相似度快速建立逻辑分组，并由组员分担组间链路的管理维护，在缓解中心节点负担的同时也增强了网络鲁棒性，而且采用 *Hash* 值与节点 ID 对应的形式，提高了组间链路建立的效率。下一步研究方向将进行 GSM 的实现并加强可用性。

参考文献

[1] Stoica I, Morris R, Karger D, et al. Chord: A Scalable Peer-to-Peer Lookup Service for Internet Applications[C]//Proc. of SIGCOMM'01. New York, USA: ACM Press, 2001.

[2] Hui K Y K, Lui J C S, Yau D K Y. Small World Overlay P2P Networks[C]//Proc. of the 12th Workshop on Quality of Service. Montreal, Canada: [s. n.], 2004.

[3] Triantafillou P. PLANES: The Next Step in Peer-to-Peer Network Architectures[C]//Proc. of FDNA'03. Karlsruhe, Germany: ACM Press, 2003.

[4] Sripanidkulchai K, Maggs B, Zhang Hui. Efficient Content Location Using Interest-based Locality in Peer-to-Peer Systems[C]//Proc. of INFOCOM'03. San Francisco, USA: [s. n.], 2003.

编辑　金胡考

(上接第 91 页)

它改善分类性能，以增加形成更多具有代表偶尔发生的长流的聚类的概率，比如 DoS 攻击，并且可能只损失轻微的流量正确率。通过对形成的分类进行统计，表 3 表明分类器取得了较高的检测率和较低的误警率。

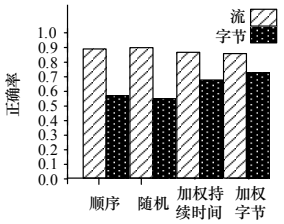

图 2　采样技术对分类正确率的影响

表 3　各组测试数据集结果汇总　(%)

组次	检测率	误警率	组次	检测率	误警率
1	85.33	0.34	6	89.23	0.43
2	88.37	0.53	7	83.17	0.48
3	86.54	0.29	8	90.01	0.28
4	87.58	0.26	9	87.14	0.31
5	86.26	0.39	10	86.25	0.36

5　结束语

本文提出了一种基于 FCM 的半监督网络流量分类算法，以确定不同流量类型。与监督分类算法比较，该算法只利用了少量标记数据，同时，对于基于 FCM 的半监督分类的最优配置，有利于提高分类性能。实验结果表明，该算法在入侵检测中的应用尽量减少了手工和经验的成分，检测效率和可靠性明显得到提高。对分类器进行再训练是延长其使用寿命的关键，如何有效地进行分类器的再训练和采样技术的评估，是进一步研究的重要方向。

参考文献

[1] Moore A W, Zuev D. Internet Traffic Classification Using Bayesian Analysis Techniques[J]. Performance Evaluation, 2005, 33(1): 50-60.

[2] Erman J, Mahanti A, Arlitt M, et al. Offline/Realtime Traffic Classification Using Semi-supervised Learning[J]. Performance Evaluation, 2007, 64(9-12): 1194-1213.

[3] Hettich S, Bay S D. The UCI KDD Archive[EB/OL]. (1999-10-20). http://kdd.ics.uci.edu/databases/kddcup99/kddcup99. html.

[4] Dy J G, Brodley C E. Feature Selection for Unsupervised Learning[J]. The Journal of Machine Learning Research, 2004, 5(1): 845-889.

[5] Pal N R, Bezdek J C. On Clustering for the Fuzzy C-means Model[J]. Proc. of the IEEE, 1995, 31(3): 370-379.

[6] Mori T, Uchida M, Kawahara R, et al. Identifying Elephant Flows Through Periodically Sampled Packets[C]//Proc. of IMC'04. Taormina, Italy: [s. n.], 2004: 115-120.

编辑　索书志