基于 NetFlow 的应用协议半监督识别算法

刘 炯,徐同阁
(北京航空航天大学 计算机学院 网络技术北京市重点实验室,北京 100191)

摘 要:传统协议识别算法无法适应当前多变复杂的网络环境,尤其在当今复杂网络环境中 P2P 应用中广泛的动态端口应用。因此针对传统端口识别方法的局限性,提出了一种基于 NetFlow 的应用协议半监督识别算法。采用 NetFlow 报文数据为基础,通过对 NetFlow 属性维进行子空间聚类,生成每种协议类型的关键维组,采用半监督算法,根据每种协议的关键维特征识别 NetFlow 数据。实验结果表明,文中提出的基于 NetFlow 的应用协议半监督识别算法在测量准确度上优于传统识别算法,有效地解决了当前复杂网络中的协议识别问题。

关键词:协议识别;半监督;子空间聚类
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2010)07-0009-04

A Semi-Supervised Clustering Algorithm for Application Protocol Recognition Based on NetFlow

LIU Jiong, XU Tong-ge
(Beijing Key Laboratory of Network Technology, School of Computer Science and Engineering, Beihang University, Beijing 100191, China)

Abstract: The traditional protocol recognition algorithms cannot adapt to the current complex and changing network environment, especially in the P2P applications, which use a wide range of dynamic port. Therefore, the limitation of the traditional port classification method, presents the semi-supervised recognition algorithm based on NetFlow. With NetFlow technology, through the sub-space clustering, select the main properties set, using semi-supervised algorithm on NetFlow data for identification application protocol. The results show that semi-supervised clustering algorithm for application protocol recognition base on NetFlow, the measurement accuracy is better than the traditional recognition algorithms to effectively solve the problem of today's complex network protocols recognition.

Key words: protocol recognition; semi-supervised; subspace clustering

0 引 言

随着 IP 网络的发展,网络环境日益复杂,对网络进行合理的监控管理,判断出当前流量的协议组成是相当重要的。目前网络应用协议组成分析方法主要有基于 SNMP、基于抓包分析、基于网络探针、DPI 等常规分析方法。

Flow 技术是当前用于采集网络流量信息的常用技术,它以流的概念为出发点,利用流数据测量网络的流量信息。目前主流的 Flow 技术有两种:sFlow 技术和 NetFlow 技术。在 Flow 技术中,"流"被定义为在给定的源端点和目的端点之间持续一段时间的单向数据包/帧序列。在 IP 网络中,可以根据 IP 数据包的源/目的 IP 地址、源/目的端口号以及协议类型等信息构成多元组来定义网络流,即具有相同源/目的 IP 地址、源/目的端口号以及协议类型等信息的 IP 数据包可归于一个流。基于 Flow 数据的端口分析流量应用协议是目前基于 Flow 的常规的流量分析算法。此算法的优点是算法复杂度比较低,处理时间快,对一些常规协议分析比较准确[1]。但是随着网络的逐步发展,越来越多的应用采用了动态端口或者随机的端口,导致基于端口的 Flow 分析算法失效。针对以上问题,文中提出了一种基于 Flow 的半监督学习算法,用来提高对动态端口的应用协议的分析,弥补传统算法[2]的不足。

收稿日期:2009-11-01;**修回日期**:2010-02-28
基金项目:北京市教育共建项目基金(JD100060630)
作者简介:刘 炯(1984-),男,山西人,硕士研究生,研究方向为网络管理;徐同阁,副教授,主要研究方向为网络管理(1997~2001 年在新加坡惠普公司、美国惠普公司任高级软件工程师,从事研发工作)。

1 基于 NetFlow 应用协议半监督识别算法

机器学习及相关技术近年开始运用在应用协议识

别领域[3~5]。文中基于 NetFlow 的提供的属性组和半监督学习算法的特点[6,7]，提出一种基于 NetFlow 的应用协议的半监督识别算法[8]。算法的基本原理为：首先通过子空间聚类，从 NetFlow 的属性集，及基于 NetFlow 属性生成的扩展属性集中选取部分属性作为聚类属性；然后根据选择的子空间，将 NetFlow 数据聚类，并提取每个分类的特征作为每个应用协议的聚类识别特征。

1.1 子空间聚类

1.1.1 NetFlow 属性的高维聚类问题

在对 NetFlow 数据的聚类研究中，要处理的数据经常有几十个属性，将这些数据对象表示成高维属性空间中的点或者向量，就可以把 NetFlow 的对象集用高维空间中点的集合来表示[9]。这些数据与低维数据相比在许多方面表现出不同的特性，如稀疏性、空空间现象以及高维数据处理过程中的“维度效应”现象。在高维数据聚类中这些问题主要表现在两个方面：一方面距离函数难于定义；另一方面在高维空间中最近邻的概念常常会失去其计算意义。

如果依照传统的聚类方法，在处理 NetFlow 数据的过程中会遇到如下问题：

(1)随着需要计算的属性维数增长，聚类的时间和空间复杂度迅速上升而导致算法的性能下降。

(2)相比低维数据，高维数据空间很难确定一个聚类的中心。高维数据集中存在大量无关的属性，并且在这些不相关的维上十分稀疏。

(3)距离函数难于定义，高维情况下会产生维度效应。

1.1.2 子空间选取

针对高维数据的问题，一般有两种解决方法：特征选择[10]和特征变换[11,12]。特征选择就是从原始高维空间中选取若干属性维组成新的坐标系，但保留的维没有旋转，不同的是，特征变换是指构造降维映射将原数据集的维合并至 k 个新维，保持基本特征不变，使得维数减少，从而使算法能在这 k 个新维中进行有效聚类。

文中提出一种 NetFlow 的特征选择算法，提供一些数据的先验知识，通过迭代和枚举选择相关性高的属性维，消去与所需特征无关，或与已有数据特征高度相关的属性维，已达到维度约减的目的。不同于常用的随机搜索、向前搜索、向后搜索、加权方法、基于分型等方法，文中是在数据集的不同子空间中搜索聚类簇的过程，同样的数据集在不同的属性子空间上形成的聚类不同，所以子空间聚类的结果包括搜索到的聚类簇及其所在的子空间。

此方法需要进行两次搜索过程，一个是搜索 NetFlow 数据聚类簇所在的子空间，即相关属性的集合；另一个则是要搜索存在于子空间中的 NetFlow 数据集。通过对选取子空间的 NetFlow 数据集簇中先验数据的判断来确定此子空间是否符合选取要求。

1.1.3 子空间聚类算法

首先将未标记的数据集、已标记的数据集做预处理，将数据库内的数据转化成程序将使用的格式。

然后根据一个初始化种子[13]，从空间中选取若干空间做为子空间。将已标记的数据映射到子空间，成为新的子空间数据，不在子空间的其他维的数据(属性)可以舍弃。为了提高准确性，可以分几组分别获得子空间，再根据相似性进行合并。

接着，将映射过的数据跟未标记的数据，通过聚类算法进行聚类。聚类算法可以选择多种聚类算法。不同的算法，聚类后的结果也不尽相同。算法中一个可以改进和提高的步骤是距离的计算。目前系统是将属性标准化后，采用几何距离计算。

聚类得到若干个数据簇，每个簇理论上应该对应一个实际应用协议。如果出现分簇过细，区分度不高，离散点过多等问题，则说明子空间选取不当。返回重新选择子空间，如果聚类结果符合要求则将此子空间记录为最终使用的子空间。迭代过程中，也可记录一个符合标准的子空间，然后继续计算，当找到多个符合标准的子空间，可以通过对这些子空间的结果指标做对比，择优保存。目前判断标准是同一个簇中元素距离中央的距离方差较小。且已标记的同类型数据均在一个簇中，不同类型的已标记数据不在同一个簇中。

子空间聚类算法的整体算法流程，如图 1 所示。

在选取子空间过程中，除了 NetFlow 固有的属性之外，还提供一些额外的属性维，这些属性维是 NetFlow 原有属性中提取和变化出来的。部分属性如表 1 所示。

表 1 部分自增属性

属性名称	属性说明
包的平均字节数 avgOctetsPerPkt	dOctets / dPkts
流的持续时间 Time	First - Last
平均每秒的包数 avgPkts	dPkts / (First - Last)
平均每秒的字节数 avgOctets	dOctets / (First - Last)
包频率 PktFret	一定时间内的平均包数

1.2 NetFlow 数据聚类

通过子空间选取算法，可以将 NetFlow 多维数据映射到较小的子空间中，降低维度，只保留主要的相关属性做为聚类的属性判断标准。

整个基于 NetFlow 的应用协议的半监督识别算法框架如图 2 所示。

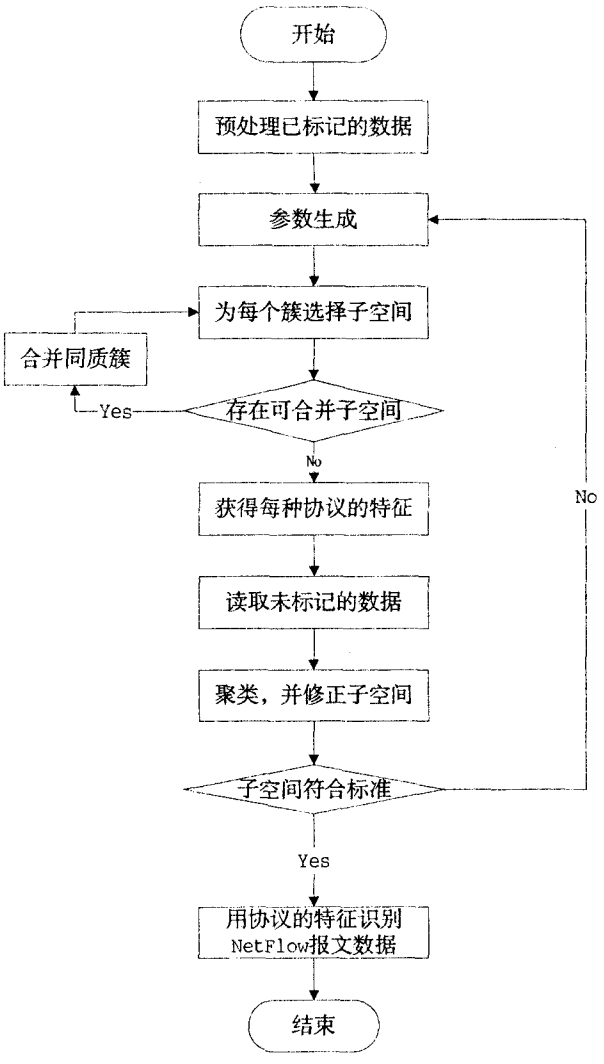

图 1　子空间选取算法

各模块介绍如下:

(1)Flow 生成程序/Flow 收集程序。

Flow 生成程序采集互联网上的原始流量数据,依照一定的采样算法生成 NetFlow 数据,并发送给 Flow 收集程序。Flow 收集程序将获得的 NetFlow 包数据整理后存入数据库中。

(2)离线数据数据库。

用来保存历史 NetFlow 数据的数据库。根据时间段划分表格,以防止数据量过大。目前系统采用 MySQL 数据库。

(3)未标记的数据/已标记的数据。

从数据库中选取一个时间段的数据做为未标记的数据,另外准备部分已标记的数据,两者做为训练集。

(4)半监督聚类学习算法。

通过对未标记和已标记的数据做半监督聚类算法,获得有效子空间和不同协议的特征参数。下文将会详细描述。

(5)已学习的属性组和协议特征类型。

上述的半监督聚类学习算法会算出有效的子空间。这个子空间的维度标示,也就是 NetFlow 协议规定的属性组中的一个子组。系统将通过此子组做为判定依据。此外,系统还会记录每种协议在这个子组中的属性参数值和置信区间。

(6)基于 Flow 的协议分析。

通过已选定的属性组和协议特征参数做为基础,导入历史 NetFlow 数据,判断数据最可能是的协议,将此记录标记为判定的协议,并做标记。

2　实验结果及分析

基于上述的算法,文中以 2008 年 3～5 月,已经标记过(即协议类型已知)的三个月的 Flow 数据为基础,做了几组实验。这些数据取自本单位实验室学生用机的真实流量数据。

首先,将三个月的 NetFlow 数据按月分成三组。选其中一组作为训练集,剩下两组作为判定集。训练

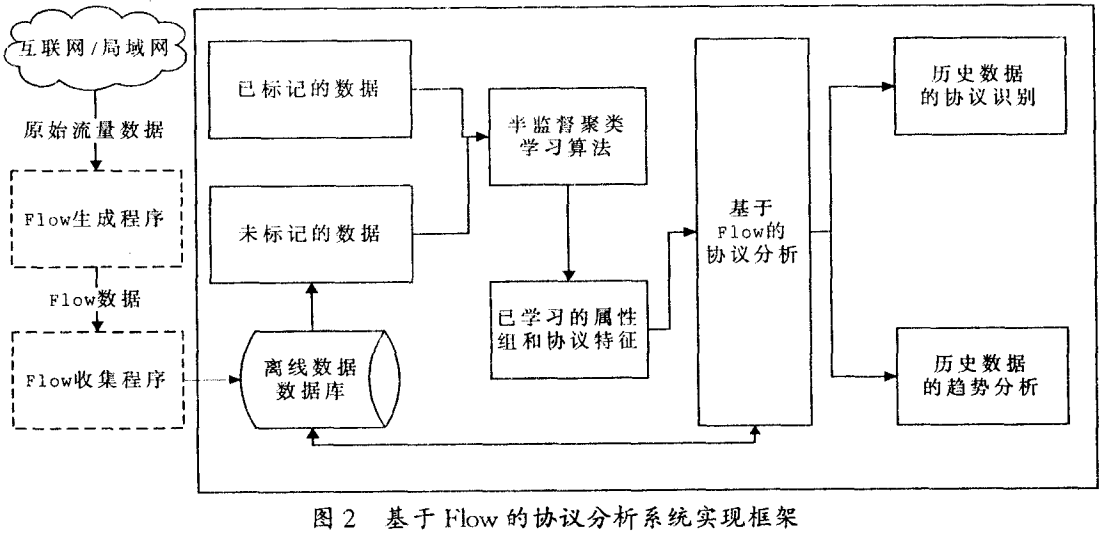

图 2　基于 Flow 的协议分析系统实现框架

集中，只留部分数据的标记符号，其余数据假设不知其协议类型。然后将这组数据做半监督训练，得到子空间和协议特征。

通过子空间聚类选取算法，计算出最可能做为最优子空间的前十项属性维。如表 2 所示。

表 2 选取的子空间属性

选取的属性	属性在子空间中出现的频率
流的持续时间 Time	91%
传输层协议 Prot	83%
包的平均字节数 avgOctetsPerPkt	77%
流内的数据包数量 dPkts	74%
平均每秒的字节数 avgOctets	73%
平均每秒的包数 avgPkts	68%
流内传输层的字节数量 dOctets	65%
源端口 SrcPort	49%
包频率 PktFret	45%
目的端口 DstPort	31%

然后以得到的子空间和协议特征作为判断标准，去分析剩下两组数据。将分类结果与实际结果作对比，得到如下数据，如图 3 和图 4 所示。

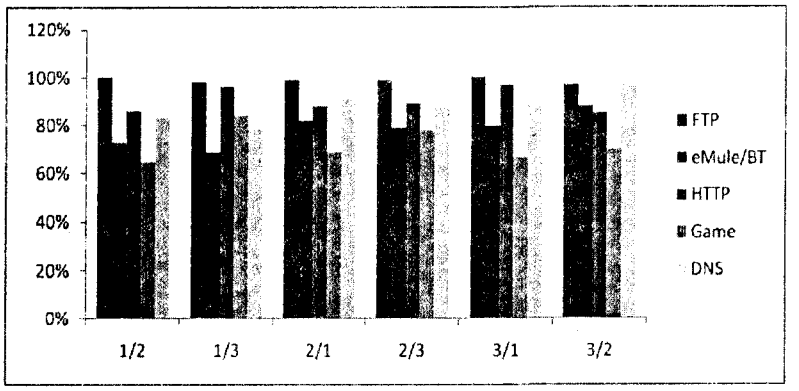

图 3 正确识别率实验结果

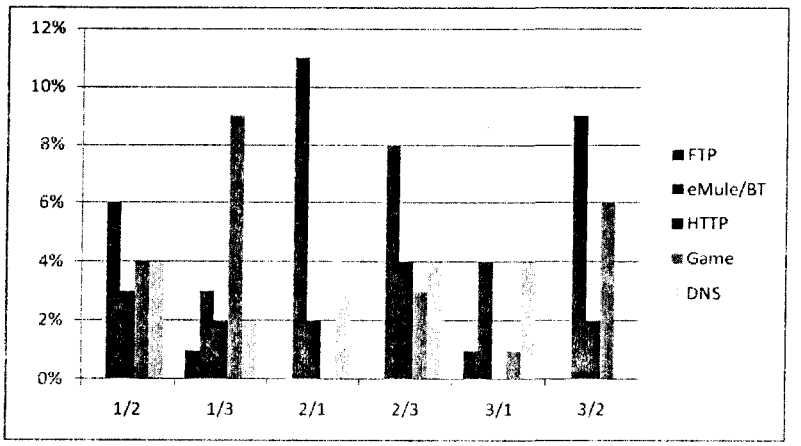

图 4 错误识别率实验结果

图中横轴表示不同的协议，纵轴表示协议的识别的准确率。试验共分 6 组。横轴下方的数字编号表示了每组试验中训练集和测试集的序号，如 1/3 表示：第三组数据做为训练集，用第一组数据做验证，得到的各组协议的准确率。

图 3 是正确识别率，即准确标记为此协议的流的数量除以标记为此协议的总的数量。可以看出对于大部分流量数据均可做出很好的识别，平均识别率约为 85%。实验结果可以看出对于基本的 FTP/HTTP/DNS 等协议可以很好地判断，eMule/BT 下载的流量特征差异性比较大，识别率稍差。

图 4 是错误识别率，即错误标记为此协议的流的数量除以标记为非此协议的流的数量。基本错误率保持在 5%以下，平均错误率为 3%。以上试验验证了文中提出的模型是可行的。

3 结束语

文中将 NetFlow 技术和半监督聚类理论应用到网络应用协议识别分析中，提出了应用半监督聚类算法，分析收集的 NetFlow 数据，筛选子空间，并在选取的子空间的基础上构建半监督聚类算法，将 NetFlow 数据聚类，进而识别协议。本应用协议识别算法及配套的采集和分析系统已在实际中取得了良好的效果。

参考文献：

[1] Zander S, Nguyen T, Armitage G. Automated traffic classification and application identification using machine learning[C]// Proceedings of the IEEE Conference on Local Computer Networks - 30th Anniversary. [s. l.]: IEEE Computer Society, 2005:250 - 257.

[2] Kim M S, Won Y J, Hong J W K. Application level traffic monitoring and analysis on IP networks[J]. ETRI Journal, 2005, 27(1):22 - 41.

[3] Dreger H, Feldmann A, Mai M, et al. Dynamic Application - Layer Protocol Analysis for Network Intrusion Detection[C]// Proceedings of USENIX Security Symposium. Berkeley, CA, USA: USENIX Association, 2006:257 - 272.

[4] 马永立，钱宗珏，寿国础，等. 机器学习用于网络流量识别[J]. 北邮学报，2008，32(1):65 - 68.

[5] McGregor A, Hall M, Lorier P, et al. Flowclustering using machine learning

（下转第 16 页）

验结果见表 1,表中 T 表示算法时间开销,S 表示查出的相似记录数,R 为查出的正确的相似记录数,P 表示准确率,下标 1 和 2 分别表示原算法和改进算法。

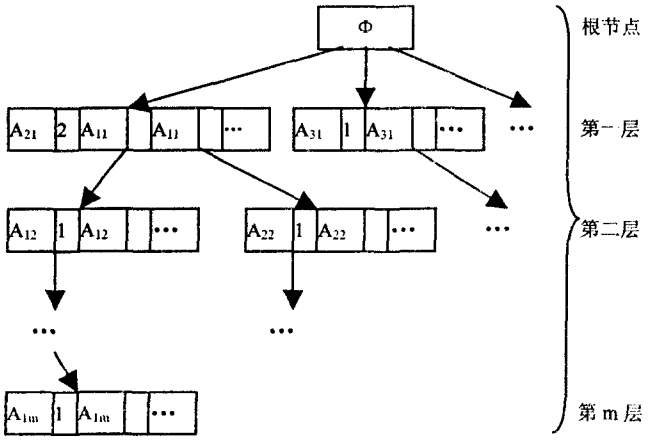

图 3 构造的聚类树[8]

表 1 实验结果表

记录数	T_1(ms)	S_1	R_1	P_1(%)
10000	35281	510	428	83.9
20000	145625	1315	1057	80.4
30000	314219	1574	1316	83.6
记录数	T_2(ms)	S_2	R_2	P_2(%)
10000	1891	524	452	86.1
20000	11172	1247	1014	81.3
30000	13937	1621	1379	85.1

从实验结果可知,与原算法相比,文中算法的总时间开销减少了很多,准确率略有提升。实验表明,本改进算法与基于聚类树的清洗算法相比优点为:避免了因属性排序不当使有些相似记录未被识别的情况,提高了精度;减少了一些不必要的相似度比较过程,在一定程度上提升了效率。但是本算法中优先队列的使用降低了结果的精度,这将是下一步的研究方向。

5 结束语

介绍并分析了数据清洗领域中针对增量式数据库的基于聚类树的相似重复记录检测算法,针对该算存在的不足提出了改进算法,并用实验证明了改进算法的有效性。

参考文献:

[1] 邱越峰,田增平,季文贝斌,等.一种高效的检测相似重复记录的方法[J].计算机学报,2001,24(1):69-77.

[2] Hernandez M, Stolfo S. The Merge/Purge Problem for Large Databases[C]//Proceedings of the ACM SIGMOD International Conference on Management of Data. New York, USA: ACM,1995:127-138.

[3] Hernandez M,Stolfo S. Real-world data is dirty:data cleansing and the merge/purge problem [J]. Data Mining and Knowledge Discovery,1998,2(1):9-37.

[4] Monge A E. An adaptive and efficient algorithm for detecting approximately duplicate database records[EB/OL]. 2003-03. http://citeseer.nj.nec.com/monge00adaptive.html.

[5] 陈 伟,王 昊,朱文明.一种提高相似重复记录检测精度的方法[J].计算机应用与软件,2006,23(10):29-30.

[6] 佘春红.基于优先队列的增量式重复记录识别[J].计算机应用,2003,23(9):61-63.

[7] 许向阳,佘春红.近似重复记录的增量式识别算法[J].计算机工程与应用,2003,39(12):191-193.

[8] 刘 芳,何 飞.一种基于聚类树的增量式数据清洗算法[J].华中科技大学学报,2005,33(3):46-48.

[9] 李星毅,包从剑,施化吉.数据仓库中的相似重复记录检测方法[J].电子科技大学学报,2007,36:1273-1277.

[10] Smith T F, Waterman M S. Identification of common molecular subsequences[J]. Journal of Molecular Biology, 1981, 2(3):195-197.

(上接第 12 页)

techniques[C]//International Workshop on Passive and Active Network Meaurement. Berlin:Springer,2004:22-41.

[6] 孙海波.基于协议行为特征的协议识别方法[C]//全国网络与信息安全技术研讨会论文集(上册).北京:中国通信学会,2007:245-251.

[7] 张春男,龙 翔,高小鹏.面向应用的网络流控系统的设计与实现[J].微计算机信息,2008,24(10-3):88-90.

[8] 谭 炜,吴 健.基于半监督学习的 P2P 协议识别[J].计算机工程与设计,2009,30(2):291-293.

[9] 温 超,郑雪峰,戚 翔,等.基于流量分析的 P2P 协议识别方法的研究[J].微计算机应用,2007,28(7):714-717.

[10] Woo Kyoung-Gu, Lee Jeong-Hoon, Kim Myoung-Ho, et al. FINDIT:a fast and intelligent subspace clustering algorithm using dimension voting[J]. Information and software technology,2004,46:255-271.

[11] Basu S, Bilenko M, Mooney R J. A Probabilistic Framework for Semi-Supervised Clustering[C]//Proceeding of Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004). Seattle, WA, USA: ACM,2004:59-68.

[12] Grira N, Crucianu M, Boujemaa N. Active semi-supervised fuzzy clustering[J]. The journal of the pattern recognition society,2008,41:1834-1844.

[13] Basu S, Banerjee A, Mooney R. Semi-supervised Clustering by Seeding[C]//Proceedings of the 19th International Conference on Machine Learning (ICML-2002). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.,2002.