多决策树算法在P2P网络流量检测中的应用

孙名松[1]，邸明星[1]，王湛昱[2]
(1.哈尔滨理工大学 网络信息中心，黑龙江 哈尔滨 150080；
2.哈尔滨工业大学 机电工程学院 工程训练中心，黑龙江 哈尔滨 150001)

摘 要：P2P技术作为一种全新的网络应用，正主导着互联网的发展方向，P2P的管理问题也成为当前互联网络中最大的难题。通过分析P2P流量特征及控制P2P流量过程中存在的问题，比较目前P2P流量检测的几种技术，提出一种基于属性关键度的多决策树分类方法，设计了一个基于多决策树算法的P2P流量检测模型，阐述了模型的工作原理。从虚警率和漏警率以及检测率三个方面评价了采用多决策树算法进行P2P流量检测的有效性。通过大量实验证明，该方法具有较高的检测率，说明采用多决策树分类算法进行P2P流量检测的有效性。

关键词：对等网络；流量检测；数据挖掘；决策树算法
中图分类号：TP393 **文献标识码**：A **文章编号**：1673-629X(2010)06-0126-04

Application of Decision Tree Algorithm in Traffic Detection of P2P Network

SUN Ming-song[1], DI Ming-xing[1], WANG Zhan-yu[2]
(1. Network Information Center, Harbin University of Science and Technology, Harbin 150080, China;
2. Engineering Training Center, School of Mechano-Electric Engineering, Harbin Institute of Technology, Harbin 150001, China)

Abstract: As a new network application, P2P technology is leading the direction of development of the Internet. The management of P2P network has also become the biggest problem in the Internet. By analyzing the P2P traffic features and the problems in the process of controlling P2P traffic and comparing existing technologies about P2P traffic detection, proposed a decision-tree algorithm based on the degree of important of property, designed a P2P detection model based on multi-decision tree algorithm, and also described the working principle of model. Evaluate the classification in term of false positive, false negative and detection rate. The experiment showed that the method has a high identification rate to the protocol of P2P, which showed that an effectiveness in using decision-tree classification algorithm on the identification of P2P traffic.

Key words: peer to peer; traffic detection; data mining; decision tree algorithm

0 引 言

对等网络(peer to peer，简称P2P)，是指在网络上起相同作用的对等用户之间进行相互通信的技术。目前P2P应用技术主要有P2P文件共享、P2P即时通信、P2P协同计算、P2P流媒体传输技术[1]。在P2P网络环境中，成千上万台彼此连接的计算机都处于对等地位，整个网络一般不依赖于专用的集中服务器。网络中的每一台计算机既能充当网络服务的请求者，又能对其他计算机的请求做出响应，提供资源与服务。

随着P2P技术的迅猛发展和广泛应用，人们在受益于P2P技术所带来的巨大利益的同时，也不得不面对信息安全的严峻挑战。P2P技术的大规模使用产生了网络带宽的巨大消耗，甚至引起网络拥塞，降低其它关键业务的性能[2]。同时，P2P技术往往采用动态随机端口、跳跃端口、冒充HTTP、有效载荷加密等技术逃避检测。传统的检测P2P流量方法是基于固定端口的检测技术[3]，由于P2P应用从最初使用固定端口发展到随机端口进行数据传输，在传输的具体内容方面也从使用明文传输发展到对传输数据进行加密处理，面对这些应用一些基于端口、基于协议[4]以及基于流量特征[5,6]的检测技术就显得有些困难了。实现对P2P流量的有效识别及控制成为了当前亟待解决的问题。

决策树算法(decision tree algorithm)是一种应用较

收稿日期：2009-10-13；**修回日期**：2010-01-15
作者简介：孙名松(1963-)，男，教授，研究方向为网络应用、网络安全。

为广泛的数据挖掘分类算法。决策树算法能够通过训练数据自动生成分类模型——决策树,并且可以利用生成的决策树对未知分类的数据进行预测。如果将 P2P 流量检测的过程看成是对流量是否是 P2P 应用以及哪种 P2P 应用的预测过程,则 P2P 流量检测系统可以利用决策树算法来创建检测模型和利用决策树模型进行 P2P 流量检测。其实质是通过决策树分类算法将所有网络数据包分为 P2P 流和非 P2P 流,然后在决策树分类算法二分类[7]的基础上,进行多值分类,将 P2P 流划归为具体属于哪一类协议的 P2P。因为决策树的预测过程是从决策树根到叶节点一条路径的匹配过程,无需对整个决策树进行遍历,匹配次数降低,减少了运算量,所以将决策树算法运用到 P2P 流量检测中将提高检测效率。

1　决策树算法

决策树方法是以实例为基础的归纳学习算法,它从一个无次序、无规则的实例集合中归纳出一组采用树型结构来表示的分类规则[8,9]。决策树方法在分类、预测、规则提取等领域得到广泛应用。利用决策树处理分类问题通常分为两步:通过训练集的学习,形成决策树分类模型;利用生成的决策树模型对类型未知的样本进行分类。使用决策树模型对类型未知样本进行分类时,从根节点开始逐步对该样本的属性进行测试,并沿着相应的分支向下行走,直至到达某个叶节点,此时叶节点所代表的类型即为该样本的类型。由此可见,利用决策树方法进行分类的关键是根据训练集合构建决策树分类模型[10]。

2　决策树算法在 P2P 流量检测中的应用

2.1　基于决策树算法的 P2P 流量检测模型

基于决策树算法的 P2P 流量检测模型由网络连接信息提取模块、数据预处理模块、多决策树检测模块三部分组成,如图 1 所示。该检测模型从网络连接信息提取模块中得到的网络数据流中分析出网络连接记录,并提取出每条网络连接的特征信息;将网络连接的特征信息由数据预处理模块进行数据预处理,得到决策树的输入形式;多决策树检测模块通过分析来确定是否是 P2P 流量。

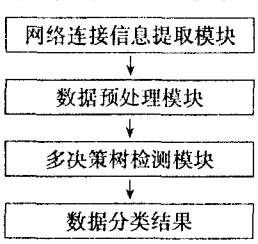

图 1　基于多决策树的 P2P 流量检测模型

模块的作用如下:

1)网络连接信息提取模块。该模块的主要工作是在捕获的网络数据包中提取出网络连接的特征数据信息,即将网络数据信息转换成网络连接记录的形式,每个记录包含了从原始数据中得到的各种特征值。

2)数据预处理模块。数据预处理模块的作用是对当前截获的数据包进行预先处理。决策树模型的生成是以训练数据为基础的。网络连接信息提取模块得到的网络连接记录包括了网络连接记录的特征信息,需要经过协议分析后对原有数据进行处理,最终整理为属性—值形式的数据。

3)多决策树检测模块。该模块的主要工作是对预先选定的训练数据集进行训练,训练数据集中的数据是从数据预处理模块中获得的。因为 P2P 网络流量具有高速、海量性传输的特点,文中使用多决策树分类算法进行 P2P 流量检测。

2.2　算法设计

2.2.1　改进多决策树分类算法

训练数据集经过数据预处理模块生成决策树可以处理的属性—值的二维表形式。表中每一列代表一个属性,每一行代表一个连接信息。设 $PE = \{PE_1, PE_2, \cdots, PE_k\}$ 为训练数据集的全部属性集合,把整个属性集 PE 分成 N 个小的属性集,每个小的属性集互不相交。$V_i = \{v_{i1}, v_{i2}, \cdots, v_{imj}\}$ 为属性 PE_i 所有不同取值的集合,当属性为连续属性时,将其离散化。r_i 为属性 P_i 的关键度,r_i 越大属性关键度越高。$\{TR_1, TR_2, \cdots, TR_N\}$ 为生成的 N 棵决策树,$C = \{c_1, c_2, \cdots, c_N\}$ 为数据分类。U 为数据集合,其中 x_i 对应于集合中的第 i 条记录。训练数据和测试数据是以 (x_i, y_j) 的形式给出的,表示第 i 条记录属于 y_j 类。

一个信息系统的分辨矩阵是一个 $|U| \times |U|$ 的对角矩阵。其中每一项定义为:

$$u_{ij} = \begin{cases} \{a \in A \mid a(x_i) \neq a(x_j)\}, \\ d(x_i) \neq d(x_j), d(x) \in D \\ \phi, d(x_i) = d(x_j), d(x) \in D \end{cases} \quad (1)$$

采用分辨矩阵结合专家经验的方法来确定数据集中每个属性的权值。文中给出了一种计算属性关键度的方法。设 $\rho(a_i)$ 是属性 a_i 的关键度。

1) 初始时对所有 $a_i \in A$,令 $\rho(a_i) = 0$。

2) 对分辨矩阵中每一项 u_{jk} 计算:

$$\rho(a_i) = \rho(a_i) + \frac{|A|}{|u_{jk}|}, a_i \in u_{jk}, 0 < k < j \leqslant |U| \quad (2)$$

其中,$|A|$ 是所有属性的基数,$|u_{jk}|$ 是分辨矩阵中

u_{jk} 的基数。

对于一个数据量非常大的信息系统，可以将整个信息系统划分为 K 个子系统，对于第 $k(k < K)$ 个子系统求出属性 a_i 的关键度 $\rho_k(a_i)$，则 a_i 在整个系统中的关键度为：

$$\rho(a_i) = \frac{\sum_k \rho_k(a_i)}{K} \tag{3}$$

设由专家经验给出的权值为 $\rho'(a_i)$，则属性 a_i 的关键度 $\rho_{ai} = \rho(a_i) + \rho'(a_i)$。

当多决策树建成后，在训练的过程中，用训练数据集中的数据更新每棵子决策树中相应节点中各分类的统计数据。经过训练后的每棵子决策树的每个节点保存落入该节点的不同分类训练数据的数量。属性关键度多决策树生成算法如下：

步骤 1：在全部属性集 PE 中任选 W' 个属性，把整个数据集分成 N 类，并按 V_i 中属性值的多少对该数据集进行递增排序，得到属性集 $v_i = \{v_{i1}, v_{i2}, \cdots, v_{iW'}\}$。

步骤 2：计算每棵子决策树的关键度 $\rho_i = \sum_{v_i} r_i / W'$。

步骤 3：以 v_{i1} 为根节点建立一棵子决策树，树的第 k 层对应的分裂属性为 v_{ik}。

步骤 4：重复步骤 1，直到 N 棵决策树建立完毕。

P2P 检测的本质是通过对大量数据进行训练，在训练的过程中逐步修正错误，形成一个较为精确的预测模型。因此当多决策树建立完成后，开始对数据集进行训练。设 TA 为训练数据集，n[P2P] 为节点中保存经过该节点且属于 P2P 类的训练数据的数量，而 n[Non - P2P] 为节点中保存经过该节点且属于 Non - P2P 类的训练数据的数量，i 为子决策树的个数，整个训练过程如下所示：

步骤 0：开始。

步骤 1：$i = 1$，对训练数据集 TA 中的每一条数据，$(x, y) \in TA$。

步骤 2：repeat

步骤 3：n[P2P] = 0，n[Non - P2P] = 0。

步骤 4：从根节点出发，如果该训练数据最终所对应的叶子节点属于 P2P 类，则从根节点到叶子节点这条路径上所有内部节点中的 P2P 域都加 1，即 n[P2P] = n[P2P] + 1，否则转到步骤 5。

步骤 5：n[Non - P2P] = n[Non - P2P] + 1。

步骤 6：$i = i + 1$。

步骤 7：until $i = N$。

步骤 8：结束。

因为各子决策树中的根节点和内部节点仅存储了经过该节点的数据的数量，因此当训练数据集中增加新的训练数据时，不会改变已生成好的树的结构，用多决策树生成算法生成的决策树具备增量学习的特性。根据属性的不同建立了 N 棵子决策树，每棵子决策树仅反映了部分属性对分类的影响，当将所有子决策树的属性对分类的影响综合起来考虑，能较好地反映整个问题的实质。

2.2.2 属性关键度多决策树分类算法的设计

属性关键度多决策树模型的基本思想是把大容量数据集分成若干子数据集，在子数据集上运用决策树算法形成不同的子决策树，然后将多棵子决策树的检测结果用属性关键度平均法进行处理，最后得到分类结果。

当多决策树生成后，经过训练数据集的训练后，形成了一个检测模型。从网络上截获的原始 TCP/IP 数据包经过数据预处理模块处理后，再由每棵子决策树 $(TR_1, TR_2, \cdots, TR_n)$ 对这些处理后的 TCP/IP 数据进行判断，然后将判断结果 $(R_1, R_2, \cdots, R_n)$ 进行加权处理，最终得到一个最优的结果。其中 TR_i 表示第 i 棵子决策树，x 为一条待分类的数据包，n[P2P] 中存储的是当前内部节点数据训练时属于 P2P 类的统计数，而 n[Non - P2P] 中存储的是当前内部节点数据训练时属于 Non - P2P 类的统计数，P_i[P2P | x] 为数据包 x 在第 i 棵子决策树中属于 P2P 的比率，而 P[P2P] 为数据包 x 在整个属性集上属于 P2P 的比率。下面为属性关键度多决策树分类算法：

步骤 0：开始。

步骤 1：i 为子决策树的个数，初始化 $i = 1$。

步骤 2：把待分类数据包 x 的属性按照多决策树生成算法进行划分，生成 N 个子集。

步骤 3：repeat

步骤 4：数据包 x 从根节点出发，跟每一个内部节点进行比较后，该包最终到达的叶子节点不管是 P2P 类还是 Non - P2P 类，计算：

$$P_i[\text{P2P} \mid x] = n[\text{P2P}] \times \rho_i / \sum_{\text{P2P}} n[\text{P2P}]$$

步骤 5：$i = i + 1$。

步骤 6：until $i = N$。

步骤 7：$P[\text{P2P}] = \frac{1}{N} \sum_{i=1}^{N} P_i[\text{P2P} \mid x]$。

步骤 8：结束。

根据属性的不同可以建立 N 棵子决策树，每棵子决策树反映了部分属性对分类的影响，当将所有子决策树的属性对分类的影响综合起来考虑，能较好地反映整个问题的实质，可以降低误报率。由于分成 N 棵

子决策树,因此可以并行处理数据包,提高了检测的效率。

3 实验结果

文中实验中数据来源是哈尔滨理工大学网络信息中心研究室,使用 Sniffer 软件在实验室 PC(AMD Athlon(tm)64 X2 Dual Core Processor 4000 + 2.11GHz,1GB 内存)采集实时的网络流量数据。

使用虚警率、漏警率以及检测率三个性能指标作为衡量分类器的分类效果,下面公式(4)~(6)分别是对三个性能指标的定义:

f_1 为被正确分类的 P2P 流量;

f_2 为被错分到 P2P 中的非 P2P 流量;

f_3 为被错分到非 P2P 中的 P2P 流量。

$$虚警率 = \frac{f_2}{f_1 + f_2} \times 100\% \tag{4}$$

$$漏警率 = \frac{f_3}{f_1 + f_3} \times 100\% \tag{5}$$

$$检测率 = \frac{f_1}{f_1 + f_3} \times 100\% \tag{6}$$

在训练分类器实验中使用 Data1,它是定时定量的 P2P 流量,而且数据量比较小。在测试分类器实验中使用 Data2~Data5,主要目的是严格测试虚警率与漏警率。

Data2 含有少量 P2P 流量,主要用来观察漏警率;

Data3 是 P2P 和非 P2P 的混合数据流量,用来测试分类器的学习能力;

Data4 是数据量比较大的混合数据,其中包含前几组数据流量中没有出现的 P2P 应用,用来测试分类器的推广能力;

Data5 是大流量数据,主要观察数据集的增加对分类器分类性能的影响。

采集的测试数据如表 1 所示。

表 1　实验数据集

测试数据集	持续时间	大小 Bytes	P2P 流含量(%)	P2P 种类
Data1	10min	103M	100%	Bit Torrent
Data2	10min	610M	12%	PPLIVE
Data3	10min	753M	51%	迅雷,BitTorrent
Data4	10min	960M	78%	eMule,BitTorrent
Data5	10min	2.18G	65%	Maze,BitTorrent

由表 2 可以看出,对网络流量特征经过选择后,基于多决策树的 P2P 网络流量分类器具有较好的分类效果,平均分类检测率达到 94.53%。

表 2　训练分类器后实验结果

训练数据集	检测率(%)	虚警率(%)	漏警率(%)
Data2	95.6	7.5	11.4
Data3	94.8	11.3	8.5
Data4	94.2	13.6	7.9
Data5	93.5	14.5	9.3

4 结束语

决策树算法是一种广泛使用的数据挖掘分类算法,该算法能够通过训练数据自动生成分类模型——决策树,并且可以利用生成的决策树对未知分类的数据进行预测。针对日益增多的 P2P 流量问题,文中将决策树分类算法应用于 P2P 流量检测,提出了一种基于属性关键度的多决策树 P2P 流量检测模型。

实验证明,提出的 P2P 流量检测模型具有较高的检测率,说明采用决策树算法进行 P2P 流量检测的有效性。

参考文献:

[1] Crowcroft J, Pratt I. Peer to Peer: Peering into the Future[C]//IFIP TC6 Networks 2002 Conference. Pisa, Italy: [s. n.], 2002:111-115.

[2] 赵金生. P2P 业务对我国互联网业务的发展影响[J]. 电信网技术,2006(2):71-73.

[3] Gummadi K P, Dunn R J, Stegan S, et al. Measurement, Modeling and Analysis of A Peer-to-Peer File-Sharing Workload[C]// Proceedings of Multimedia Computing and Networking 2003. New York: [s. n.], 2003:314-329.

[4] 李江涛,姜永玲. P2P 流量识别与管理技术[J]. 电信科学,2005,21(3):57-61.

[5] Zhou L J, Li Z T, Liu B. P2P traffic identification by TCP flow analysis[C]// Proceedings of International Workshop on Networking, Architecture, and Storages. Shenyang: [s. n.], 2006:47-50.

[6] 吴　敏,王汝传. 基于主机的 P2P 流量检测与控制方案[J]. 计算机技术与发展,2009,19(10):26-29.

[7] 於建华,徐艳萍,吴素芹. 基于流传输特性的 P2P 流量识别方法研究[J]. 通信技术,2007,40(11):247-249.

[8] 满桂云. ID3 决策树算法的改进研究[J]. 中国信息科技,2007(13):32-34.

[9] 王春枝,李　涛. 基于双层特征的 P2P 流量检测[J]. 计算机技术与发展,2009,19(7):238-241.

[10] 杨学兵,张　俊. 决策树算法及核心技术[J]. 计算机应用与发展,2007,7(1):43-45.