

第四周

网络流量特征提取和分类研究总结

Wu You

2018 年 6 月 28 日

摘要

目录

1	引言	1
1.1	当前网络环境	1
1.2	网络流量分析技术	1
1.2.1	基于端口的流量识别	1
1.2.2	基于深度包检测的流量识别	1
1.2.3	基于深度流检测的流量识别	2
1.3	网络流量特征提取	2
2	网络安全画像简介	2
3	国内外研究现状	3
3.1	决策树的构建过程	3
3.1.1	信息增益	3
3.1.2	决策树修剪	3
3.1.3	一些特殊属性的处理	4
4	贝叶斯方法	5
4.1	贝叶斯法则	5
4.2	贝叶斯理论的应用	6

1. 引言

1.1. 当前网络环境

由于互联网技术发展迅速，同时也为国家发展带来了巨大的经济效益，我国的互联网规模在一段时期内还将不断扩大。而随着互联网的不断普及，用户人数不断增加，同时产生了多种类的新型网络应用，网络已经覆盖了生产生活的各个方面，社会发展对网络的依赖性也逐渐增强。而网络规模的告诉发展也带来一定问题，首先一些使用新型的网络应用如基于 P2P 或 VoIP 的软件对于网络的占用率较高，影响其他网络应用的运行；其次随着电子商务的兴起，网络中信息价值不断提高，导致网络安全问题日益突出；最后，互联网信息扩散十分迅速，为不良信息的传播提供了条件。[1]可以看出，加强对网络流量的监督和管理是十分重要的。

1.2. 网络流量分析技术

网络流量分析技术是根据不同网络应用流量的特点，实现网络流量的对比、分类，在网络流量管理、信息监测或者网络安全防护工作上，流量分析都是一项重要技术手段。根据分析层面可以将流量分析技术分为三类，即基于端口的分析、深度包检测（Deep Packet Inspection, DPI）以及深度流检测（Deep Flow Inspection, DFI）。

1.2.1. 基于端口的流量识别

早期互联网应用种类较少，不同协议的应用只需根据标准端口号即可进行区分，但互联网环境日渐复杂，出现了多种端口伪装技术，同时新兴的互联网协议，如 P2P 协议使用了动态端口技术导致该方法的准确率大大降低，目前只能作为辅助分析手段。[2]

1.2.2. 基于深度包检测的流量识别

深度包检测技术是一种基于数据包载荷的流量分析技术。由于相同应用的数据包内容有一定的相似性，可以从数据包的内容中提取一定的特征作为规则，并建立规则库，通过内容匹配来进行流量分类。这种方法的优点是识别精度较高，但是由于需要获取数据包内容，涉及了用户隐私问题，同时在高速网络中，分析的实时性无法得到保证。现今网络环境迅速变化，依赖人工建立规则库的开销巨大，且无法对规则库进行及时的更新，而大量的规则也会引起分析速度下降等问题，因此现在 DPI 的研究方向除了算法识别精度的提升外，还包括算法运行速度的提升以及规则集合的精简问题。

1.2.3. 基于深度流检测的流量识别

深度流检测技术是一种基于网络流量整体统计特征的分析技术。该技术主要依赖机器学习算法，可以实现流量的自动、准确分类。但是由于需要提取网络流的整体特征，很难做到流量的实时分析，且训练得到的模型在处理未知类型流量时无法获得良好的效果。目前针对该技术研究主要方向有算法的选择、对比和优化，特征的有效选择以及未知流量的识别等。

无论 DPI 还是 DFI，都存在一些共同的问题需要解决，如数据的采集和处理、特征有效性的验证、实时性的保障等。

1.3. 网络流量特征提取

特征的提取对于流量识别的速度、精度等有直接影响，人工特征提取的方式开支较大，且有效性和时效性无法得到保证，故自动特征提取成为现在研究的重点。基于 DPI 和 DFI 的流量分析所使用的特征有很大区别，因此相应的特征提取技术也有所不同。

基于 DPI 的流量分析主要使用了流量内容的模式特征，除字符串以外，还可以包括比特串、数串以及正则表达式等，这类特征的提取方法是利用数据挖掘技术进行数据包载荷的关联规则分析。下面将介绍几种用于模式特征提取的方法。

文献 [1]中使用了一种基于序列模式挖掘的特征提取方法，改算法是经典 PrefixSpan 算法的改进，该算法直接在数据包的十六进制数序列中进行挖掘，通过模式增长的方式来生成连续序列模式。其中还使用了偏移属性约束，在一定程度上减少了算法的计算量。但没有考虑规则库的筛选和更新问题，文中提出可使用增量序列模式挖掘的方法来实现规则库的更新。文献 [4]中在 AutoSig 算法的基础上进行了一定修改，提出了一种自动协议指纹挖掘算法以用于 P2P 流量的特征提取。实验结果表明该算法可以挖掘出人工未能总结出的协议指纹。文献 [3]中使用了固定比特流算法进行流量特征提取，

基于 DFI 的流量分析主要使用了流量整体的统计特征，Moore 提出了针对网络流量的 249 种统计特征，已经十分全面，DFI 技术所面对的不是特征提取的问题，而是如何选择特征以提高算法的精度、速度，并实现实时检测。

2. 网络安全画像简介

画像技术大多应用于用户分析，可服务于商品推荐、用户行为预测、营销策略的制定等场景，而对于网络安全的状态，我们也可以使用画像技术来进行描述，以便于更好地呈现网络态势，有助于形成更好的防御策略，并实现威胁预警。网络安全画像，实际上就是利用数据对网络的各方面，包括漏洞、弱点、防御能力等，进行定性或定量评估，并将评估结果用直观的、易于理解的方式对用户进行呈现。在安全画像的构建过程中，主要使用数据挖掘技术以及数据融合技术

1.

3. 国内外研究现状

3.1. 决策树的构建过程

决策树算法通过结合训练样例的目标值，对训练样例的各个属性依次分类，每个属性为一个节点，若节点中含有目标属性值不同的样例，则进一步进行划分，最终形成一个树形结构，该树可为二叉树或多叉树。构建树的首个问题是选择最佳的初始分类属性，也就是根节点最佳属性的选择。

3.1.1. 信息增益

对于概念学习，信息熵的定义如下：

$$Entropy(S) = -p_{\oplus} \cdot \log_2 p_{\oplus} - p_{\ominus} \cdot \log_2 p_{\ominus}$$

其中 p_{\oplus} 和 p_{\ominus} 分别为被该属性分为正例和反例的样本所占比例。进一步地，对于目标属性有 c 个取值的概念，信息熵为：

$$Entropy(S) = \sum_{i=1}^c -p_i \cdot \log_2 p_i$$

对任意属性，其信息熵最大值为 $\log_2 c$ 最小值为 0。信息熵描述了样本的均一性。有了信息熵，之后可以对信息增益进行定义，对样本集合 S ，某一属性 A 的信息增益定义如下：

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

该属性刻画了属性 A 对样本 S 的分类效果，信息增益越大，表明样例信息熵降低越多，也就意味着分类后的样本更加一致，在进行最佳分类属性的选择时，以信息增益取最大值为标准。

3.1.2. 决策树修剪

在对样本的学习过程中，由于噪声或特殊样例的存在，通常会造成过拟合的情况发生，即该树在训练样本集合上表现良好，但在对新的实例进行分类时准确率欠佳（泛化精度低），为提高泛化精度，需要额外的处理，主要有两种方式提高泛化精度：

- 提前停止树的生长
- 对树进行后修剪

对于以上两种方法，有一个共同的问题：确定其遵循的准则，也就是何时停止或对树修剪到何种程度，对此问题有以下几种解决方案：

- 用新的样例来评估效果
- 对增长或修剪操作的效果进行估计
- 用编码来衡量

其中，第一种方法最为常用，我们可以将训练样例分为训练集和验证集，比例为 1:1 或 2:1，但验证集要足够大。（k-折交叉验证与此类似，此处应为 2-折交叉验证）下面是两种树的修剪方法：

1. 错误率降低修剪
2. 规则后修剪

错误率降低修剪是将以某一节点为根的树全部移除，并将属于该子树的最常见目标属性值赋予该节点，但仅在可以降低错误率时进行此操作。规则后修剪方法首先令树生长到尽可能拟合训练数据，之后将树转化为等价的规则集合，为根节点到每个叶节点的路径创建规则（ \wedge ），之后对这些规则进行泛化以提高其精度，这一方法使得修剪更为灵活，并且对人来说，规则更加容易理解。

3.1.3. 一些特殊属性的处理

连续值属性

决策树在处理离散值时，可以将其取值范围划分为多个区间，从而像对待离散值属性一样处理该属性。

有大量取值的属性

例如日期这类属性，拥有大量的不同属性值，假设在某一节点以日期为 1 号对样例进行划分，会使得样例被划分为一大一小两部分，此时该属性的信息增益将大于其他属性，但实际上该属性分类效果很差，此时需要用其他标准来代替信息增益，信息增益比率是一个可用的标准。首先定义分裂信息（Splitinformation）：

$$splitinformation(S, A) = \sum_{i=1}^c \frac{|S_i|}{S} \cdot \log \frac{|S_i|}{S}$$

信息增益比率的定义如下：

$$Gainratio(S, A) = \frac{GainS, A}{splitinformation(S, A)}$$

实际上，分裂信息就是 S 关于 A 的熵，当 A 的取值数量较多时，其分裂信息也较大，信息增益比率的值会相应减小，这相当于对取值较多的属性进行了惩罚。

处理缺少的属性

对于缺少某属性的样例，可赋予其最常见值，或按照已观察到的该属性取值比例对这些样例进行赋值。

处理不同代价的属性

4. 贝叶斯方法

贝叶斯方法利用先验知识和观察数据一同决定假设的最终概率，可以对假设做不确定性的预测，该方法也可用于衡量其他算法的最优决策。

4.1. 贝叶斯法则

对某一假设 h 和训练样例 D ，有如下的贝叶斯公式：

$$P(h|D) = \frac{P(D|h) \cdot P(h)}{P(D)}$$

其中， $P(h)$ 是 h 的先验概率，基于现有知识得到， $P(D|h)$ 是在我们认为 h 成立的情况下，观察到 D 的概率，而 $P(D)$ 是 D 的先验概率，等式左边的 $P(h|D)$ 为假设 h 的后验概率，即观察到 D 时， h 成立的概率。极大后验概率假设（MAP）定义为使后验概率最大的假设：

$$\begin{aligned} h_{MAP} &= \underset{h \in H}{\operatorname{argmax}} P(h|D) \\ &= \underset{h \in H}{\operatorname{argmax}} P(D|h)P(h) \end{aligned}$$

有时候我们会给每个假设赋予相同的先验概率，这时，我们只需要根据 $P(D|h)$ 来决定最大可能的假设， $P(D|h)$ 常称为 h 成立时样本 D 的似然度，使该概率最大的假设称为极大似然假设（ML）：

$$h_{ML} = \underset{h \in H}{\operatorname{argmax}} P(D|h)$$

4.2. 贝叶斯理论的应用

对于具有一组属性 a_1, a_2, \dots, a_n 的实例来说, 其极大后验假设为:

$$h_{MAP} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j | a_1, a_2, a_3, \dots, a_n) P(v_j)$$

朴素贝叶斯分类器基于以下假设: 各个属性值的取值是互相独立的。因此, 其极大后验假设可以进一步表示为:

$$\begin{aligned} h_{MAP} &= v_{NB} \\ &= \underset{v_j \in V}{\operatorname{argmax}} P(v_j | a_1, a_2, a_3, \dots, a_n) P(v_j) \\ &= \underset{v_j \in V}{\operatorname{argmax}} \prod_{i=1}^n P(a_i | v_j) P(v_j) \end{aligned}$$

并且

$$P(a_i | v_j) = \frac{n_c + mp}{n + m}$$

参考文献

- [1] 林冠洲. 网络流量识别关键技术研究. PhD thesis, 北京邮电大学, 2011. 1, 2
- [2] 柏骏, 夏靖波, 吴吉祥, 任高明, and 赵小欢. 实时网络流量分类研究综述. 计算机科学, 40(9):8–15, 2013. 1
- [3] 牟澄. 互联网流量特征智能提取关键技术研究. PhD thesis, 北京邮电大学, 2014. 2
- [4] 马婧. 网络流量特征提取与流量识别研究. PhD thesis, 北京邮电大学, 2012. 2