



基于过抽样技术的 P2P 流量识别方法^{*}

钱亚冠¹, 张 旻²

(1. 浙江科技学院理学院 杭州 310023; 2. 杭州电子科技大学计算机学院 杭州 310018)

摘 要: 针对 P2P 类不平衡问题提出将复杂的多类不平衡问题转化为简单的两类不平衡问题, 再通过迭代 SMOTE 过抽样技术丰富 P2P 的概念表达, 从而提高 P2P 流量的识别率。实验结果表明, 该方法可以显著提高诸如 Naïve Bayes 这样的简单模型在 P2P 上的识别率, 由此可证明该方法对于改善 P2P 类不平衡问题的有效性。

关键词: 类不平衡; 过抽样; P2P 流量识别

doi: 10.3969/j.issn.1000-0801.2014.04.016

P2P Traffic Identification Based Over-Sampling Technique

Qian Yaguan¹, Zhang Min²

(1. College of Science, Zhejiang University of Science and Technology, Hangzhou 310023, China;

2. College of Computer Science, Hangzhou Dianzi University, Hangzhou 310018, China)

Abstract: A new approach that transfers the multi-class imbalanced problem into simple two-class imbalanced problem was proposed, and a mended SMOTE over-sampling algorithm to enrich the P2P concept was applied. The experimental results show the approach can efficiently improve the P2P identifying rate using the Naïve Bayes model.

Key words: class imbalance, over-sampling, P2P traffic identification

1 引言

近年来, P2P 技术已被广泛应用于文件共享、视频内容分发、即时通信等网络应用领域。自 2004 年以来, P2P 流量在整个互联网流量中逐渐占据主导地位(60%以上)^[1,2]。P2P 流量的快速增长给网络带宽带来了巨大压力, 其近乎对称的流量模式更加剧了网络的拥塞。同时, 基于 P2P 技术的恶意流量也开始肆虐互联网, 造成带宽的过度消耗, 甚至导致拒绝服务^[3]。因此, 如何快速正确地识别 P2P 流量已经成为当前网络管理者面临的巨大挑战。

互联网流量的识别技术经历了最初的基于 TCP 端口、深度分组检测(deep packet inspection, DPI)到目前兴起的机器学习方法和基于网络行为的识别等技术^[4]。有的 P2P

应用为了躲避检测, 开始采用动态端口、数据分组加密等技术手段, 使得基于 TCP 端口与 DPI 的方法效率越来越低。而基于统计学的机器学习方法却可以克服上述不足, 因而它逐渐显示出在 P2P 流量分类中的优势^[5]。

传统的机器学习方法通常假设目标类是均匀分布的, 而实际的互联网流量中的各种应用的分布是不均匀的。尤其是 P2P 这样的大象流(elephant traffic), 它们按字节数统计在流量上占很大比例, 对网络性能的影响很大, 但从数据流(flow)角度统计却占很少比例^[6]。目前基于机器学习的流量分类方法通常基于数据流的统计信息, 因此占数据流比例很小的 P2P 流量往往难以识别, 分类器倾向于将 P2P 数据流识别为如 WWW 这样的多数类。这种目标类比例严重失衡而导致少数类识别误差增大的问题通常称为类不

^{*} 浙江省网络媒体云处理与分析工程技术中心开放课题(No.2012E10023-14)



平衡(class imbalance)问题,是目前 P2P 流量难以识别的一个重要原因。

网络流量中的众多应用比例极不平衡,流量分类问题面临的是多类不平衡问题^[7]。而 P2P 应用本身在数据流中所占比重很小,又受到其他应用目标类的干扰,本文提出将 P2P 识别中的多类不平衡问题转化为两类不平衡问题的思路,并通过过抽样(over-sampling)方法增加 P2P 流量的比重,消除分类器在学习过程中的偏倚,提高 P2P 的识别率。本文提出改进的迭代 SMOTE(i-SMOTE)过抽样方法来提高 Naïve Bayes 算法的识别率,实验结果证明本文提出的识别框架具有良好的识别性能。

2 类不平衡问题与 SMOTE 过抽样方法

目前基于机器学习的流量分类方法大多利用数据流层面的统计信息。因此像 P2P 这类应用,尽管在字节流上占很大比重,在数据流层面却占很小的比例,与 WWW 应用相比存在严重的不平衡性。这种不平衡性将导致 P2P 很高的误分类率。传统的机器学习分类算法旨在最小化全局分类误差,并假设假正例与假负例的错误代价是相等的,因此偏向于把少数类预测到多数类上,如将 P2P 预测为 WWW。而实际网络管理过程中,可能对于识别类似 P2P 这样的少数类更有价值,因此需要有提高 P2P 识别率的有效方法。为了克服上述类不平衡问题,机器学习界提出重抽样技术来平衡目标类的分布,即对多数类(majority class)进行欠抽样(under-sampling),对少数类(minority class)进行过抽样(over-sampling)。

传统的欠抽样与过抽样技术都具有自身的不足:对多数类欠抽样会导致一些信息的丢失,而对少数类的简单重复抽样在早期的研究中就已发现对于提高分类性能并无太大的帮助^[8]。因此,Chawla N V 等^[9]提出了新的过抽样技术 SMOTE 算法,其基本思想是通过人工合成新的少数类样本来减轻类别的不平衡,解决传统过抽样技术因决策域变小而引起的过拟合现象。SMOTE 算法的基本原理是在相距最近的少数类样本之间进行线性插值,从而生成新的少数类样本。首先根据过抽样倍率 N ,从每个少数类样本 k (默认取 5)个同类最近邻中随机选择 N 个样本;接着将每个少数类分别与它的 N 个选中的样本按式(1)合成 N 个新的少数类样本,并加入到原训练样本集中,形成新的训练样本集。

$$NewSample = x + rand(y[i] - x) \quad (1)$$

其中, $i=1,2,\dots,N$; $rand$ 表示 0~1 的一个随机数; $NewSample$ 表示合成的新样本; x 表示少数类样本; $y[i]$ 表示 x 的第 i 个近邻样本。

3 P2P 流量识别方法

整个 P2P 流量的分类识别方法框架如图 1 所示。

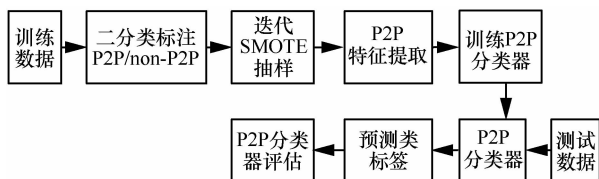


图 1 P2P 流量的分类识别方法框架

步骤 1 将训练数据进行两分类标注,即标注所有的 P2P 数据流后,将其他应用的数据流均标注为非 P2P (non-P2P)。这样就可将多标签分类问题归约到相对简单的二分类问题求解。

步骤 2 采用 i-SMOTE 算法,获得更大的 P2P 数据流样本。原始的 SMOTE 算法只是在原有的少数类样本的基础上进行线性插值获得新的样本,但最新研究表明 P2P 这样的流量少数类具有明显的概念漂移现象^[10],少量的原始样本不能完全表达 P2P 的概念。因此,采用多次迭代 SMOTE 算法的方法,在前一次迭代获得的样本集合上再进行插值运算,使得 SMOTE 算法的输入样本逐渐丰富,以便获得更完整的 P2P 概念表达。通过 i-SMOTE 算法,获得足够的 P2P 样本数,在此基础上进行步骤 3。

步骤 3 特征提取,去除冗余特征,获得维度较低的特征空间。具体的特征提取算法可以采用基于相关性的方法^[11]等。

步骤 4 训练分类器,建立预测 P2P 流量的模型。目前已有很多机器学习的分类模型被尝试用于流量分类,如 Naïve Bayes^[14]、决策树^[13]、支持向量机^[14]、神经网络^[15]等。这些模型被应用于流量分类,具有各自的优缺点。如 Naïve Bayes 具有模型简单、训练时间短的优点,但缺点是对于少数类的识别率低;而支持向量机与神经网络的识别率比较高,但模型复杂、训练与分类时间过长。本文考虑到实际环境中对 P2P 流量识别的实时性要求,认为选择简单的模型更有利于快速获得预测结果,因此选择 Naïve Bayes 模型作为评估模型。通过实验比较分析得出,当 i-SMOTE 方法获得足够的 P2P 样本数时,Naïve Bayes 模型可以对 P2P

获得很高的识别率。i-SMOTE 算法过程如下。

算法:i-SMOTE(迭代 SMOTE 算法)

输入:

训练集中的 P2P 样本集合 S;

训练集中的 P2P 样本数 T;

过抽样倍率 N;

最近邻的个数 k;

迭代次数 M;

输出:过抽样 P2P 样本集合;

(1) $S' = \text{SMOTE}(S, T, N, k)$; //原始 SMOTE 算法见参考文献[9]

(2) for $i \leftarrow 1$ to M

(3) $S' \leftarrow S' \cup S$

(4) $T \leftarrow |S' \cup S|$ //新样本集合的样本数

(5) $S' = \text{SMOTE}(S', T, N, k)$;

(6) end for

(7) return S'

4 实验评估策略

本文提出通过 i-SMOTE 过抽样的方法来提高 P2P 流量的识别率。利用最简单的 Naïve Bayes 模型比较分析 SMOTE 算法和 i-SMOTE 算法过抽样效果:随着 P2P 样本数的逐渐增加,考察它们对识别率的影响。选择最简单的 Naïve Bayes 模型的原因是:在未进行过抽样的情况下,它的识别率非常低。如果过抽样技术能提高这类简单模型的识别效果,则可以证明过抽样技术对于 P2P 识别的有效性。

评估指标采用召回率(recall)与精度(precision)这两个指标: $\text{recall} = TP/P$, $\text{precision} = TP/(TP+FP)$ 。其中,P 为测试集中事先标识为 P2P 的样本数,TP 为分类器正确预测为 P2P 的样本数,TP 为被分类器错误地将 non-P2P 流量预测为 P2P 的样本数。

4.1 实验数据集

本文采用的数据集 1 为剑桥大学 Moore 等提供的公开流量数据集^[6]。该数据集通过连续采集 24 h 的流量数据,并随机抽取 10 个约 28 min 的数据块,在这些数据块上构建出数据流,构成 10 个数据子集 Data1, Data2, ..., Data10。笔者在 10 个数据子集上进行的实验结果非常相似,因此只列出了 Data1 的实验结果。原始 Data1 中共有 12 种流量类型,如 WWW、E-mail、FTP 等,将它们均表示为 non-P2P 数据流,共计 24 524 条,P2P 数据流共计 339 条,占总数的 1.36%。

数据集 2 是从校园网中心的某台交换机上通过端口映射方法获得的流量数据,该交换机汇聚了某幢男生宿舍访问外网的所有网络流量。经过连续 1 h(晚上 21:30-22:30)的连续数据采集,共计获得 325 538 条数据流,其中 P2P 数据流有 18 632 条,占总数的 5.72%。为保护隐私的需要,只截取数据分组的分组头部分,并通过 Tcpdpriv 工具对 IP 地址进行了匿名化处理。

4.2 Naïve Bayes 模型评估

Moore 等^[3]早在 2004 年就已深入分析和应用 Naïve Bayes 模型到互联网流量分类中。通过选择合理的流量特征和核估计方法,Naïve Bayes 模型在全局正确率(accuracy)上达到 96.29%。但他们的工作只是提高了整体的正确率,并没有解决类不平衡的问题,因而对于像 P2P 这样的少数类的识别率提升有限。Naïve Bayes 模型具有简单、计算效率高的特点,与其他复杂模型相比更具有实际应用价值,因此首选它作为评估过抽样技术的效果。

对数据集 1、数据集 2 的原始 P2P 数据采用如下过抽样倍率: $N=100\%$ 、 300% 、 700% 、 $1\ 500\%$ 、 $3\ 100\%$,应用 SMOTE 算法过抽样获得新的 P2P 样本集,抽样结果分别见表 1、表 2。为了便于比较,提出的 i-SMOTE 算法每次迭代采用

表 1 过抽样数据集 1 获得的结果(样本数/所占比例)

抽样	P2P	non-P2P
第 1 次	339/1.36%	24 524/98.94%
第 2 次	678/2.69%	24 524/97.31%
第 3 次	1 356/5.24%	24 524/94.76%
第 4 次	2 712/9.96%	24 524/90.04%
第 5 次	5 424/18.11%	24 524/81.89%
第 6 次	10 848/30.67%	24 524/69.33%

表 2 过抽样数据集 2 获得的结果(样本数/所占比例)

抽样	P2P	non-P2P
第 1 次	18 632/5.72%	306 906/94.28%
第 2 次	37 264/10.83%	306 906/89.17%
第 3 次	55 892/15.41%	306 906/84.59%
第 4 次	74 528/19.54%	306 906/80.46%
第 5 次	93 160/23.29%	306 906/76.71%
第 6 次	111 792/26.70%	306 906/73.30%



固定倍率 $N=100\%$ ，这样获得的 P2P 样本数可与前述 SMOTE 算法保持一致。另外，通过传统的随机过抽样方法产生一个同比例规模的数据集作为比较基准。

采用 10 折交叉验证的方法对不同 P2P 样本数下（见表 1、表 2）的识别率进行评估。特征选择采用 FCBF 算法^[7]。图 2 给出了随机抽样、SMOTE 算法与 i-SMOTE 算法同比例扩大 P2P 的样本数的情况下召回率的对比。可以明显发现 P2P 样本数从开始的 339 条数据流增加到 2 712 条数据流时，即 P2P 比例从 1.36% 增加到 9.96% 时，Naïve Bayes 模型在 i-SMOTE 数据集上获得的 P2P 召回率明显高于 SMOTE 数据集与随机过抽样数据集，前者为 81.6%，后者分别为 31.2% 与 21.8%。同样，当 P2P 样本数增加至 5 424 条，比例增加到 18.11% 时，i-SMOTE 数据集上的召回率达到 98.5%，而 SMOTE 数据集与随机过抽样数据集分别只有 78.5% 与 38.2%。最后当 P2P 的数量比例达到 30.67% 时，SMOTE 数据集与 i-SMOTE 数据集上的召回率均在 97% 以上，而随机过抽样数据集仅为 47.9%。从上述过程可以看出，i-SMOTE 算法与 SMOTE 算法及随机过抽样相比，可以

更快速地提高召回率。同样，可以看到三者在精度上的区别（如图 3 所示）。随着 P2P 样本数的增加，3 种过抽样方法获得的数据集在 P2P 识别精度上都得到了提升，但当 P2P 样本比例到达 30.67% 时，i-SMOTE 数据集上的精度达到了 99.1%，而 SMOTE 数据集上的精度却从 94.7% 跌至 53.6%，甚至低于随机过抽样。图 4、图 5 给出了数据集 2 的 10 折交叉验证的结果，与数据集 1 的验证结果相似。从图 2~图 5 的比较分析中可以得出以下两个结论。

- 通过对 P2P 样本的过抽样，与原始数据相比不论召回率还是精度都可得到提高。
- SMOTE 算法可以使召回率与精度两者同时提高到 90% 以上，而 SMOTE 算法在召回率增长到一定程度时，精度会出现下降。精度的下降意味着 non-P2P 样本被错误地预测为 P2P 的比例增加，即假阳性率增加。传统的随机过抽样方法尽管有所提高，但提高程度有限。

因此，综合召回率与精度这两个评价指标，i-SMOTE 算法比 SMOTE 算法及随机过抽样技术更为有效。

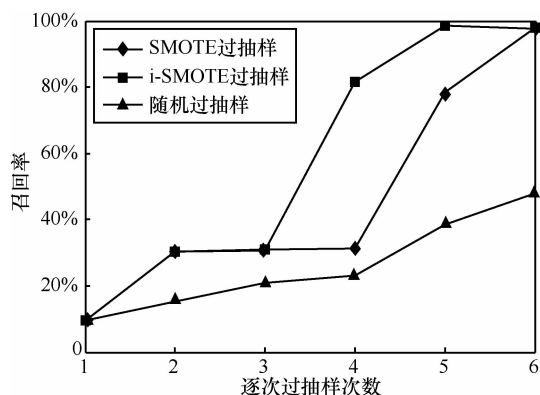


图 2 数据集 1 不同规模的 P2P 样本数的召回率

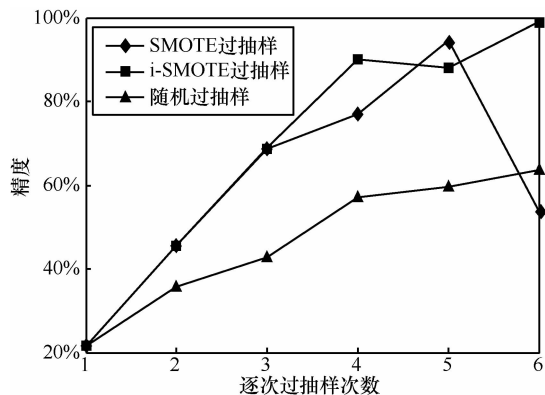


图 3 数据集 1 不同规模的 P2P 样本数的精度

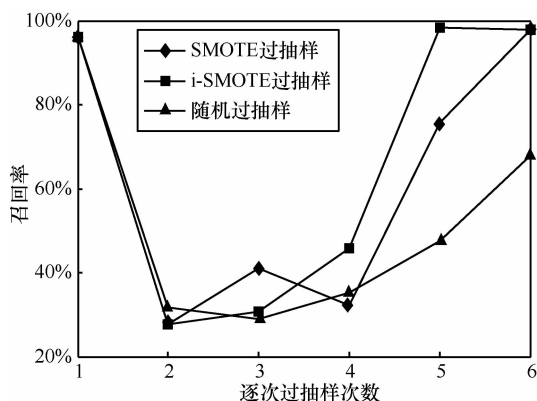


图 4 数据集 2 不同规模的 P2P 样本数的召回率

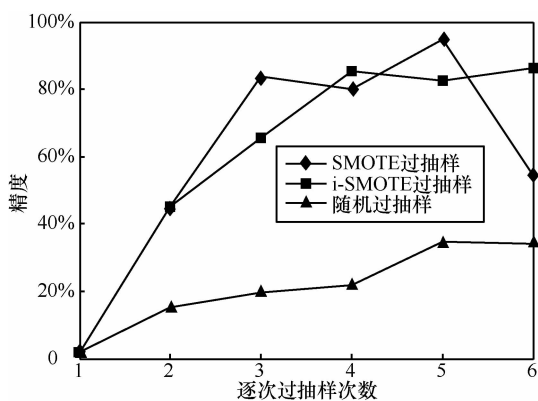


图 5 数据集 2 不同规模的 P2P 样本数的精度

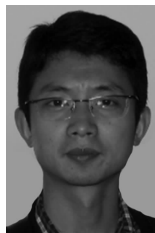
5 结束语

本文通过过抽样技术提高对 P2P 流量的识别率。提出基于迭代的 SMOTE 算法可以比原始的 SMOTE 算法及传统的随机过抽样方法具有更好的表达 P2P 概念的能力。实验结果表明本文提出的基于过抽样的方法可以有效地提高 Naïve Bayes 模型对于 P2P 的识别率。Naïve Bayes 模型由于其简单性,在流量分类中不及 SVM、神经网络等复杂模型的正确率高,通常为研究人员所忽视。但正是 Naïve Bayes 模型的简单性,使得它具有很好的算法效率,容易被应用到实际工作环境。机器学习方法的分类正确率不仅仅取决于分类模型,与数据预处理的质量也有重要关系。本文正是通过改善数据质量的思路,使得 i-SMOTE 方法与简单的 Naïve Bayes 模型相结合实现对 P2P 的高精度识别。

参考文献

- Mochalski K, Schulze H. Ipoque internet study 2008/2009. http://www.ipoque.com/resources/internet-studies/internet-study-2008_2009,2009
- MacManus R. Trend watch: P2P traffic much bigger than Web traffic. http://www.readwriteweb.com/archives/p2p_growth_trend_watch.php,2006
- Sun X, Torres R, Rao S. Preventing DDoS attacks on internet servers exploiting P2P systems. *Computer Networks*, 2010,54(15): 2756~2774
- Dainotti A, Pescapè A, Claffy K C. Issues and future directions in traffic classification. *Network*, IEEE, 2012, 26(1): 35~40
- Gong S F, Chen J. A P2P traffic detection method based on support vector machine. *Applied Mechanics and Materials*, 2012, 198: 1280~1285
- Erman J, Mahanti A, Arlitt M. Byte me: a case for byte accuracy in traffic classification. *Proceedings of the 3rd Annual ACM Workshop on Mining Network Data*, San Diego, California, USA, 2007: 35~38
- Liu Q, Liu Z. A comparison of improving multi-class imbalance for internet traffic classification. *Information Systems Frontiers*, 2012(7):1~13
- Ling C, Li C. Data mining for direct marketing problems and solutions. *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining(KDD-98)*, New York, NY, 1998
- Chawla N V, Bowyer K W, Hall L O, *et al.* SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 2002 (16)
- Wang R Y, Zhang L, Liu Z. Classifying imbalanced internet traffic based PCDD: a per concept drift detection method. *Smart Computing Review*, 2013(2)
- Hall M A. *Correlation-based Feature Selection for Machine Learning*. The University of Waikato, 1999
- Moore A W, Zuev D. Internet traffic classification using bayesian analysis techniques. *ACM SIGMETRICS Performance Evaluation Review*, 2005, 33(1): 50~60
- Xu P, Lin S. Internet traffic classification using C4.5 decision tree. *Journal of Software*, 2009, 20(10): 2692~2704
- Yuan R, Li Z, Guan X, *et al.* An SVM-based machine learning method for accurate internet traffic classification. *Information Systems Frontiers*, 2010, 12(2): 149~156
- Sun R, Yang B, Peng L, *et al.* Traffic classification using probabilistic neural networks. *Proceedings of Natural Computation (ICNC), 2010 Sixth International Conference on IEEE*, Valencia, Spain, 2010
- Moore A W. Dataset. <http://www.cl.cam.ac.uk/research/srg/netos/nprobe/data/papers/sigmetrics/>
- Yu L, Liu H. Feature selection for high-dimensional data: a fast correlation-based filter solution. *Proceedings of the Twentieth International Conference on Machine Learning (ICML 2003)*, Piscataway, NJ, USA, 2003

[作者简介]



钱亚冠,男,博士,浙江科技学院理学院副教授,主要研究方向为互联网流量分类、下一代互联网、机器学习与数据挖掘。



张旻,男,博士,杭州电子科技大学计算机学院讲师,主要研究方向为网络虚拟化、下一代互联网、软件定义网络。

(收稿日期:2014-02-20)