



基于可变特征空间 SVM 的互联网流量分类

钱亚冠¹, 关晓惠², 云本胜¹, 楼琼¹, 马鹏飞¹

(1. 浙江科技学院理学院, 浙江 杭州 310023; 2. 浙江水利水电学院, 浙江 杭州 310018)

摘要: 支持向量机(support vector machine, SVM)是一类具有良好泛化能力的机器学习算法, 适合应用于互联网动态环境下的流量分类问题。目前将 SVM 扩展到流量分类这样的多分类问题的方法主要有 One-Against-All 和 One-Against-One 方法。这些方法都基于单一的特征空间训练 SVM 两分类器, 没有考虑到不同特征对不同流量类的不同区分能力, 因此获得的分离超平面并不是最合理的。为此提出了可变特征空间的 SVM 集成方法, 即为每个两分类 SVM 构建具有最优区分能力的独立特征空间, 单独训练两分类 SVM, 最后再利用 One-Against-All 和 One-Against-One 方法集成为多分类器。实验表明, 与原来的单一特征空间的 One-Against-All 和 One-Against-One 集成方法相比, 提出的方法能有效提高流量分类器分类精度和召回率, 更易获得最优分离超平面。

关键词: 支持向量机; 可变特征空间; 流量分类

中图分类号: TP393.04

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2016132

Internet traffic classification using SVM with flexible feature space

QIAN Yaguan¹, GUAN Xiaohui², YUN Bensheng¹, LOU Qiong¹, MA Pengfei¹

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

2. Zhejiang University of Water Resources and Electric Power, Hangzhou 310018, China

Abstract: SVM is a typical machine learning algorithm with prefect generalization capacity, which is suitable for the internet traffic classification. At present, there are two approaches, One-Against-All and One-Against-One, proposed for extending SVM to multi-class problem like traffic classification. However, these approaches are both based on a unique feature space. In fact, the separating capacity of a special traffic feature is not similar to different applications. Hence, flexible feature space for extending SVM was proposed, which constructs independent feature space with optimal discriminability for each binary-SVM and trains them under their own feature space. Finally, these trained binary-SVM were ensemble by One-Against-All and One-Against-One approaches. The experiments show that the proposed approach can efficiently improve the precision and callback of the traffic classifier and easily obtain more reasonable optimal separating hyper-plane.

Key words: support vector machine, flexible feature space, traffic classification

收稿日期: 2016-01-01; 修回日期: 2016-04-09

通信作者: 钱亚冠, qianyg@zju.edu.cn

基金项目: 国家自然科学基金资助项目 (No.61379118, No.61103200); 浙江省网络媒体云处理与分析工程技术中心开放课题资助项目 (No.2012E10023-14)

Foundation Items: The National Natural Science Foundation of China (No.61379118, No.61103200), Education Department Foundation of Zhejiang Province (No.2012E10023-14)



1 引言

流量分类是互联网领域中的一个重要应用,如何准确地识别出流量的应用类型对于网络管理、流量控制及网络安全等具有重要的意义。由于互联网的复杂性、动态性,在各种应用层出不穷的环境下,如何准确地识别出流量的应用类型目前仍然是个极具挑战的课题。

互联网早期利用 TCP 端口号可以容易地确定流量的应用类型,但随着互联网应用的不断衍生,很多应用开始使用动态端口,甚至使用其他著名端口,如 P2P 应用开始使用 Web 的 80 端口传输数据。这种现状使得基于端口的方法在识别率上显著下降。基于 DPI (deep packet inspection) 的流量分类技术是目前被广泛部署的另一类方法^[1]。该方法通过检测数据分组中的用户数据部分,发现特定应用的特征字符串,实现对流量应用类型的识别。但随着目前用户数据的加密和隐私保护的要求,这种方法也越来越显示出它的不足。

最近,基于流量统计特征的机器学习方法成为流量分类领域的研究热点^[2-5]。所谓的基于机器学习的流量分类方法就是通过某种机器学习算法,从流量训练数据中建立分类模型,从而实现对流量类型的预测。这种方法的优点是可以克服数据加密的限制,同时仅利用 IP 和 TCP 这两层数据分组头部的信息,不受隐私保护的制约。由于互联网流量具有很大的动态性,如果机器学习算法过拟合(over-fitting)训练数据,那么分类模型的泛化能力就会下降,即对未知数据的预测正确率下降。在众多的机器学习算法中,支持向量机(support vector machine, SVM)因具有良好的泛化能力,比其他学习算法更适合于流量分类。

徐鹏等人^[6]提出一种基于 SVM 的流量分类方法,该方法利用非线性变换和结构风险最小化原则将流量分类问题转化为二次优化问题,实验表明该方法具有良好的分类正确率和稳定性。Alice E 等人^[7]将 SVM 应用于流量分类,提出了一个简单的优化算法解决 SVM 最优参数选择的问题。Zhou X S 等人^[8]利用 SVM 实现对产生 P2P 流量的应用程序进行分类。Li Z 等人^[9]选择了对分类影响最大的 9 种流量特征,利用 SVM 技术将网络流量分成了 bulk traffic、interactive、WWW、service、P2P、mail、other 7 类,得到了 95% 以上的整体正确率。但由于 SVM 本质上是一个两分类器(binary classifier),因此将 SVM 应用到流量分类这样

的多分类(multi-class classification)问题时,往往采用 One-Against-All^[10]或 One-Against-One^[11]等方法将两分类器集成为多分类器。但这些方法都在一个共同的特征空间下寻找最优分离超平面,但同一特征在不同类之间的区分能力并不等同^[12]。针对这个问题,本文在 One-Against-All 和 One-Against-One 方法基础上提出可变特征空间(flexible-feature-space, FFS)的方法,实验证明该方法可有效地提高流量分类的正确率。

2 支持向量机

SVM 是一种对线性和非线性数据进行分类的方法。对于非线性可分的数据集,通过非线性映射,把原始训练数据映射到高维空间,在新的空间中搜索最佳分离超平面。假设具有两种不同分类的数据集 $(x_1, y_1), \dots, (x_m, y_m), x_i \in R^n, y_i \in \{-1, +1\}$ 。基本的 SVM 就是寻找一个可以分离两类数据的最优超平面。如果该数据集是线性可分的,则分离超平面可表示为:

$$w^T \cdot x + b = 0 \quad (1)$$

其中, $W=(w_1, w_2, \dots, w_k)$ 是权重向量; b 是一个标量参数。所有的数据实例满足:

$$y_i(w^T x_i + b) \geq 1, \quad i=1, \dots, m \quad (2)$$

如果数据集是线性不可分的,那么通过一个非线性映射函数 $\phi(\cdot)$ 将原始数据映射到高维空间,从而使得在新空间中实现线性可分:

$$w^T \cdot \phi(x) + b = 0 \quad (3)$$

满足上述条件的分离超平面很多,取边缘最大的分离超平面为最优分离超平面,这样的超平面具有最佳的泛化性能。因此,求最优分离超平面的问题转化为如下的凸二次规划问题:

$$\min_{w', b', \xi_i} \|w'\|^2 / 2 + C \sum_{i=1}^m \xi_i \quad (4)$$

$$\text{s.t. } y_i(w' \cdot \phi(x_i) + b) \geq 1 - \xi_i \quad (5)$$

其中, $C > 0$ 为常数,称为惩罚系数,用以控制对错分数据点的惩罚程度; $\xi_i \geq 0$ 称为松弛变量,是为解决样本线性不可分而引入的。利用拉格朗日乘子法和 KKT(Karush-Kuhn-Tucker)条件可求解上述优化问题。

3 可变特征空间的 SVM 集成方法

将多个两分类基本 SVM 集成为能完成多分类的 SVM,通常是在相同的特征空间中搜寻最优的分离超平面。已有

研究表明,同一特征对不同流量类的区分能力是不同的^[12],如数据分组大小的均值可以较好地地区分 SSH 和 P2P 应用,但却不能很好地区分 FTP 和 P2P。单一的特征空间并不适合所有的流量类,会增加搜索最优分离超平面的困难。因此,本文提出可变特征空间的方法来克服这种单一特征空间的局限性。

图 1 给出了不同特征空间下的线性可分性的情况。假设原始特征空间 $F=\{a_1, a_2, a_3, a_4, a_5\}$, 任务是对 C_1, C_2, C_3 3 类流量进行分类。从图 1(a)可以发现,在特征子空间 $\{a_1, a_2, a_3\}$ 下,可以找到合适的超平面分离 C_2, C_3 的实例,但不能分离 C_1, C_2 和 C_1, C_3 的实例。通过改变特征子空间,如图 1(b)所示,选择 $\{a_3, a_4, a_5\}$ 作为特征子集,则可以找到合适的超平面分离 C_1, C_2 的实例,但却不能分离 C_2, C_3 的实例。由此可见,在不同的特征子空间中寻找最优分离超平面的难度不同,对于 C_2, C_3 而言,选择 $\{a_1, a_2, a_3\}$ 作为特征空间更易找到分离超平面;而对于 C_1, C_2 则选择 $\{a_3, a_4, a_5\}$ 作为特征空间则更易线性可分。因此,采用传统上的单一特征空间存在很大的局限性。考虑到 SVM 是典型的两类分类器,利用 One-Against-All 和 One-Against-One 的方式集成为多分类器,可以在单个的 SVM 分类器上采用单独的特征空间,克服单一特征空间的不合理性。除了在原始的特征空间采用这种可变特征空间的方法,也可把它推广到经过非线性变换后的核空间中。假设 $\phi(\cdot)$ 是特征空间 F^m 到 F^n 的非线性映射($n>m$),仍然可以在高维核空间 F^n 中为分类 C_i, C_j 找到合适的特征子空间,使其更容易被线性可分。

3.1 可变特征空间的 One-Against-All 集成分类方法

假设有 k 个流量类,One-Against-All 方法将集成 k 个 SVM 两分类器来实现多分类器的能力。假设给定 m 个流量训练数据实例 $(x_1, y_1), \dots, (x_m, y_m), y_i \in \{\text{HTTP, FTP, mail, } \dots,$

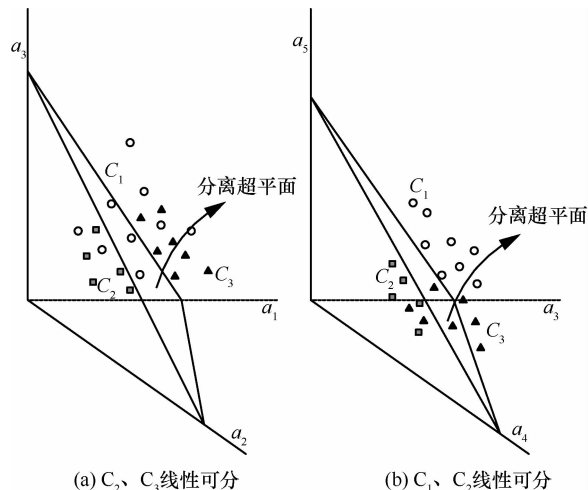


图 1 不同特征空间(也可映射到核空间)下的线性可分性

games}, 这里假设有 k 种流量类型。可变特征空间的 One-Against-All 集成方法如图 2 所示。要识别 k 种流量类型,需要构建 k 个两分类 SVM, 每个 SVM 负责识别一种流量类。每个 SVM 有专门的训练数据,如图 2 所示, SVM_1 的训练数据是通过保留 HTTP 流量的类别标签,将其他流量的类别标签改成 others 的方法构建,这样 SVM_1 只负责识别 HTTP。以此类推,其他 SVM 根据其负责识别的流量类型,用同样的方法构建相应的训练数据。

为了克服单一特征空间的缺陷,本文在每个 SVM 的专门训练数据上抽取特征。考虑到在原始特征空间上存在线性不可分的情况,先将原始的特征空间用多项式核函数 $K(x, x_i) = ((x \cdot x_i) + 1)^d$ 映射到高维空间。具体的特征选择方法采用 Guyon 等人^[13]提出的 SVM-RFE 特征选择算法获得单独的特征空间。该方法是 Wrapper 型特征选择方法,即它选择特征的度量是 SVM 的分类性能,因此该方法产生的特征空间可保证获得合理的分离超平面。该方法的基本原

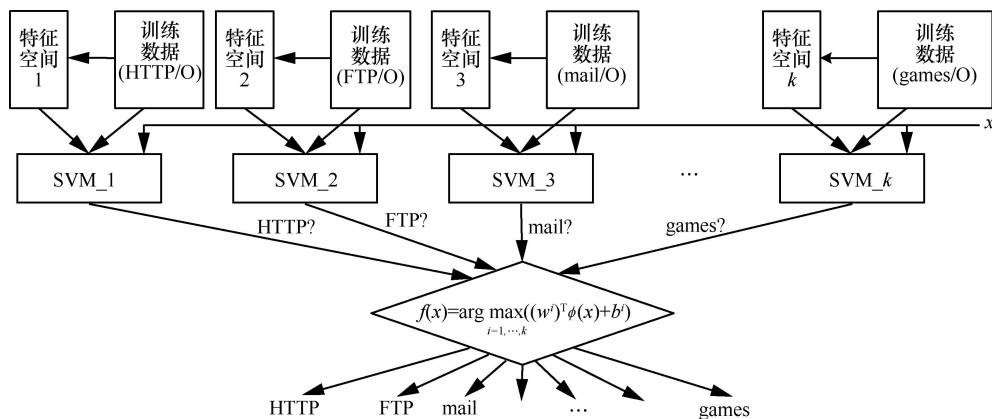


图 2 可变特征空间的 One-Against-All 集成方法



理是根据特征在 SVM 上的分类性能排序,在每一次递归迭代时去除排序在最后的那个特征。具体而言,在训练 SVM 过程中,得到当前的最优分离超平面,计算权向量 $w = \sum_k \alpha_k y_k x_k$, 则第 i 个特征的排序重要性为 $c_i = (w_i)^2$ 。本文提出的可变特征空间方法的优点是:对于特定的 SVM,去除了对于该 SVM 不重要的特征,使得搜索到的最优分离超平面更接近于假设类,从而提高整体的分类精度。

不失一般性,这里仅以 One-Against-All 集成框架中的第 $i(i \leq k)$ 个 SVM 为例说明建模原理,其本质是解决如下凸两次规划问题:

$$\min_{w^i, b^i, \xi^i} \quad \frac{1}{2} (w^i)^T w^i + C \sum_{j=1}^m \xi_j^i (w^i)^T \quad (6)$$

$$\text{s.t.} \quad (w^i)^T \phi(x_j) + b^i \geq 1 - \xi_j^i, \quad y_j = i \quad (7)$$

$$(w^i)^T \phi(x_j) + b^i \leq -1 + \xi_j^i, \quad y_j \neq i \quad (8)$$

其中, $\phi(\cdot)$ 是非线性映射函数, C 是惩罚系数, ξ 是松弛变量。最终通过如下的判决函数判定 x 的分类标签: $f(x) = \arg \max_{i=1, \dots, k} ((w^i)^T \phi(x) + b^i)$, 即取上述 k 个判决函数中的最大值所对应的预测类为最终的分类标签。本文把上述可变特征空间的思路结合到 One-Against-All 集成方法后,将其命名为 One-Against-All+, 具体算法如下所示。

输入 训练数据 $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$, 流量类标签 $TC = \{C_1, C_2, \dots, C_k\}$, 测试数据 $T = \{(x_1', y_1'), (x_2', y_2'), \dots, (x_M', y_M')\}$ 。

输出 预测正确/错误的计数器 $\{r(+1), r(-1)\}$ 和预测类别。

for $i \leftarrow 1$ to N do

for $j \leftarrow 1$ to K do

if (x_i, y_i) 中的 $y_i = C_j$ then

将 (x_i, y_i) 复制到 D_j ; // 为 SVM_j 构建专门的训练数据集 D_j

else

将 (x_i, y_i) 复制到 D_j , 并将 y_j 的标签改为“others”;

end if

end for

for $i \leftarrow 1$ to K do

$D_i' \leftarrow \phi(D_i)$; // 用多项式核函数将 D_i 映射到高维核空间 D_i'

$\Omega_i \leftarrow \text{SVM-RFE}(D_i')$; // 在核空间中得到特征空间 $(|\Omega_i| < n)$

$SVM_i \leftarrow$ 在特征空间 Ω_i 上获得模型; // 求解式 (6)~(8) 的优化问题;

end for

for $i \leftarrow 1$ to M do

for $j \leftarrow 1$ to K do

$f[j] \leftarrow (w^j)^T \phi(x_i') + b^j$; // SVM_j 输出距离值, $x_i' \in T$

end for

$r \leftarrow f[j]$ 中取得最大值的下标 j ;

output C_r ; // 输出预测类别标签

if $C_r = y_i'$ then $r(+1) \leftarrow r(+1) + 1$; // 预测正确, 计数器 $r(+1)$ 加 1

else $r(-1) \leftarrow r(-1) + 1$; end if // 预测错误, 计数器 $r(-1)$ 加 1

end for

3.2 可变特征空间的 One-Against-One 集成分类方法

One-Against-One 方法是另一种把两分类 SVM 集成为多分类器的方法。假设要完成对 k 个流量类的分类任务, 首先为每两个分类构造一个 SVM, 用于判别这两个流量类型, 共需构建 $k(k-1)/2$ 个 SVM 两分类器。对于一个未知流量, 每个 SVM 会输出一种流量类别的预测, One-Against-One 方法通过投票表决的选出最终的预测分类, 从而解决多分类问题。不失一般性, 假设第 i 个分类为 HTTP, 第 j 个分类为 FTP, 那么构建判别 HTTP 或 FTP 的两分类 SVM_{ij} 就是求解如下的凸两次规划问题:

$$\min_{w^{ij}, b^{ij}, \xi^{ij}} \quad \frac{1}{2} (w^{ij})^T w^{ij} + C \sum_t \xi_t^{ij} (w^{ij})^T \quad (9)$$

$$\text{s.t.} \quad (w^{ij})^T \phi(x_t) + b^{ij} \geq 1 - \xi_t^{ij}, \quad y_t = i \quad (10)$$

$$(w^{ij})^T \phi(x_t) + b^{ij} \geq -1 + \xi_t^{ij}, \quad y_t = j \quad (11)$$

对某个未知的流量样本 x 进行测试时, 需要利用 $k(k-1)/2$ 个 SVM 对其进行判别。如果被 SVM_{ij} 判别为属于第 i 类, 则第 i 类的票数加一; 否则, 第 j 类的票数加一。最终得票数最多的类就是 x 的预测类标签。可变特征空间的 One-Against-One 集成方法如图 3 所示。

与 One-Against-All 方法一样, 需要为每个两分类 SVM 准备专门的训练数据。假设训练一个用于区分 HTTP 和 FTP 的 SVM, 训练数据通过如下方式产生: 在原始数据中仅抽取出类标签为 HTTP 和 FTP 的流量数据。同理, 为了训练用于区分 HTTP 和 mail 的 SVM, 只从原始数据中抽取 HTTP 和 mail 流量。假设有 k 个流量类别, 那么共需构建 $k(k-1)/2$ 个训练数据集。同样采用将原始特征空间映射到

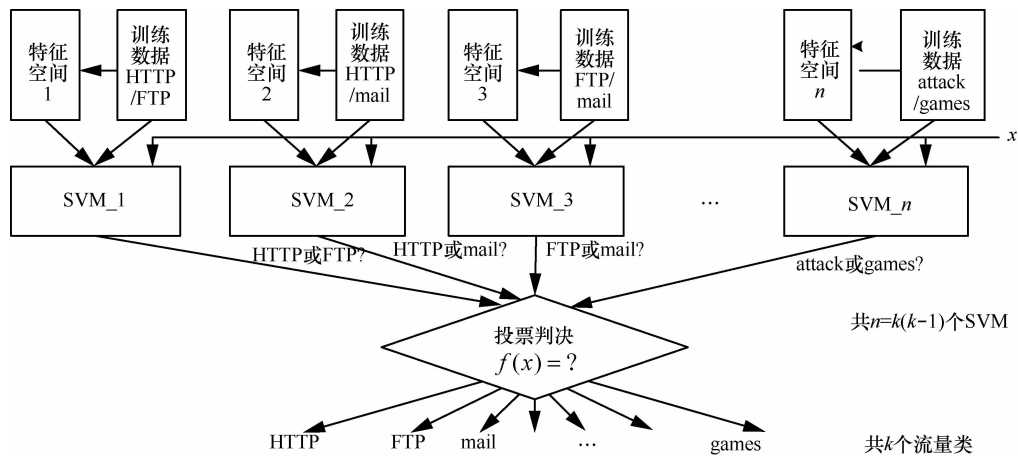


图3 可变特征空间的 One-Against-One 集成方法

高维核空间,再采用 SVM-RFE 特征选择算法为每个 SVM 选取单独的特征空间。将这种可变特征空间的 One-Against-One 集成方法称为 One-Against-One+,具体算法如下。

输入 训练数据 $D=\{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$, 流量类标签 $TC=\{C_1, C_2, \dots, C_K\}$, 测试数据 $T=\{(x'_1, y'_1), (x'_2, y'_2), \dots, (x'_M, y'_M)\}$ 。

输出 预测正确/错误的计数器 $\{r(+1), r(-1)\}$ 和预测类别。

```

for i ← 1 to N do
  for j ← 1 to K-1 do
    for k ← j+1 to K do
      if  $(x_i, y_i)$  中的  $y_i == C_j$  or  $y_i == C_k$  then
        将  $(x_i, y_i)$  复制到  $C_{j,k}$ ; // 为  $SVM_{j,k}$  构建训练数据集  $D_{j,k}$ 
      end if
    end for
  end for
end for
for i ← 1 to K-1 do
  for j ← i+1 to K do
     $D'_{i,j} \leftarrow \phi(D_{i,j})$ ; // 用多项式核函数将  $D_{i,j}$  映射到高维核空间  $D'_{i,j}$ 
     $\Omega_{i,j} \leftarrow \text{SVM-RFE}(D'_{i,j})$ ; // 在核空间中得到特征空间  $(|\Omega_{i,j}| < n)$ 
     $SVM_{ij} \leftarrow$  在特征空间  $\Omega_{ij}$  上获得模型; // 求解 (9-11) 式的优化问题;
  end for
end for
for i ← 1 to M do

```

```

  for j ← 1 to K do

```

```

    for k ← j+1 to K do

```

```

      label ←  $SVM_{j,k}(x_i)$ ; //  $SVM_{j,k}$  输出  $x_i$  的类别预测,  $x_i \in T$ 

```

```

    for s ← 1 to K do

```

```

      if label ==  $C_s$  then  $f[s] \leftarrow f[s] + 1$ ; end if

```

```

    end for

```

```

  end for

```

```

end for

```

```

r ←  $f[s]$  中取得最大值的下标 s;

```

```

output  $C_s$ ; // 输出预测类别标签

```

```

if  $y'_i == C_s$  then  $r(+1) \leftarrow r(+1) + 1$ ; // 预测正确, 计数器  $r(+1)$  加 1

```

```

else  $r(-1) \leftarrow r(-1) + 1$ ; end if // 预测错误, 计数器  $r(-1)$  加 1

```

```

end for

```

4 实验评估策略

本文采用 k -折交叉验证的方法进行实验结果的评价。 k -折交叉验证是将数据随机的划分成 k 个不相交、大小大致相等的子集 $D_1, D_2, \dots, D_k$ 。训练与测试进行 k 次, 在第 i 次迭代时, 子集 D_i 用作测试集, 其余的子集一起用作训练集。分类准确率估计是 k 次迭代准确分类的实例总数除以初始数据中的实例总数, 通常采用 10 折交叉验证。评估指标采用召回率 (recall) 与精度 (precision) 这两个指标:

$$recall = TP/P \quad (12)$$

$$precision = TP/(TP+FP) \quad (13)$$

其中, P 为测试集中事先标识为正例的样本数, TP 为



分类器正确预测为正例的样本数, TP 为被分类器错误的将正例预测为负例的样本数。

4.1 实验数据

本文采用英国剑桥大学 Moore 等人提供的公开流量数据集^[14]作为实验数据。该数据集通过连续采集 24 h 的网络流量,并按 28 min 为间隔随机抽取 10 个数据块,再将流量数据分组构建成数据流(flow),最后得到 10 个数据子集 Data1, Data2, ..., Data10。由于在 10 个数据子集上进行的实验结果非常相似,本文只列出了 Data1 的实验结果。

实验用的第二个数据集是从校园网中心的某台交换机上获得的流量数据,该交换机汇聚了某幢男生宿舍的访问外网的所有网络流量。经过连续 1 h(21:30-22:30)的连续数据采集,共计获得 325 538 条数据流。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。分类标签利用与实验室合作的迪普公司的 DPI 模块完成,并按 Moore 等提出的特征集进行了预处理。

因上述数据集中存在严重的类不平衡情况,采用欠抽样的方法降低 WWW 这类占高比例 (Moore 数据集中占 72.2%) 的流数据,最终的训练数据集中各类流量的比例见表 1、表 2。

4.2 实验分析

英国剑桥大学 Moore 等人^[14]提取出了 248 种网络流特征,但是这些特征有些是不能实时获得的。考虑到过多的特征在 SVM 训练过程中非常低效,而 CFS 这样基于相关的

表 2 类平衡处理后的数据集 2

WWW	mail	QQ	P2P	games	attack
18.7%	19.1%	22.3%	15.7%	6.6%	17.6%

特征选择算法不一定适合 SVM;基于 SVM 的 Wrapper 型算法在特征空间太大,数据很多时也非常低效,为此本文采用目前被大都数参考文献使用,又容易在线提取的特征作为基本的特征子集(见表 3)。本文提出的可变特征空间的方法就是在这个基本特征子集的基础上利用 SVM-RFE 算法提取两分类 SVM 的特征子集,如对于区分 WWW 和 mail 的 SVM,优化的特征空间为: {Dst_port, mean_data_ip_b→a, duration, throughput b→a, mean_data_ip_a→b}。由于篇幅有限,不一一列出所有两分类 SVM 的特征空间。

图 4 和图 5 是 4 种方法在数据集 1 上的流量分类精度和召回率的对比情况。其中 One-Against-One+ 表示改进 One-Against-One 的可变特征空间方法,One-Against-All+ 表示改进 One-Against-All 的可变特征空间方法。为便于比较,4 种方法的 SVM 均采用多项式核函数。从整体观察,本文提出的可变特征空间方法均使比统一特征空间的方法在精度和召回率上都有很大程度的提高。对于如 WWW、mail 这样的比例较高的类,虽然 One-Against-All 和 One-Against-One 方法已经可以达到 85% 以上的精度和召回率,改进的新方法使它们提高到 90% 以上。分类准确率提升幅度最大的是那些比例很小,原本分类准确率很低的少数类,如 attack、intertive 等。如攻击流量 attack,本身包含多种攻击类型的流

表 1 类平衡处理后的数据集 1

WWW	mail	FTP-control	FTP-pasv	attack	P2P	database	FTP-data	multimedia	services	interactive
15.4%	17.5%	6.3%	1.8%	5.2%	14.4%	10.1%	16.7%	3.7%	8.7%	0.2%

表 3 网络流特征子集

序号	网络流(flow)特征	描述
1	Dst_port	服务器端口,连接响应方
2	Src_port	客户端端口,连接发起方
3	max_data_ip	IP 数据分组中最大字节数
4	max_data_wire	以太网数据帧中最大字节数
5	syn_pkts_sent_a→b	客户端到服务器端发送的带 syn 标志的数据分组数
6	syn_pkts_sent_b→a	服务器端到客户端发送的带 syn 标志的数据分组数
7	throughput b→a	从服务器到客户端的吞吐量
8	mean_data_ip_a→b	从客户端到服务器 IP 数据分组的平均字节数
9	mean_data_ip_b→a	从服务器到客户端 IP 数据分组的平均字节数
10	mean_IAT_a→b	从客户端到服务器的数据分组平均到达间隔
11	mean_IAT_b→a	从服务器到客户端的数据分组平均到达间隔
12	duration	TCP 连接持续时间

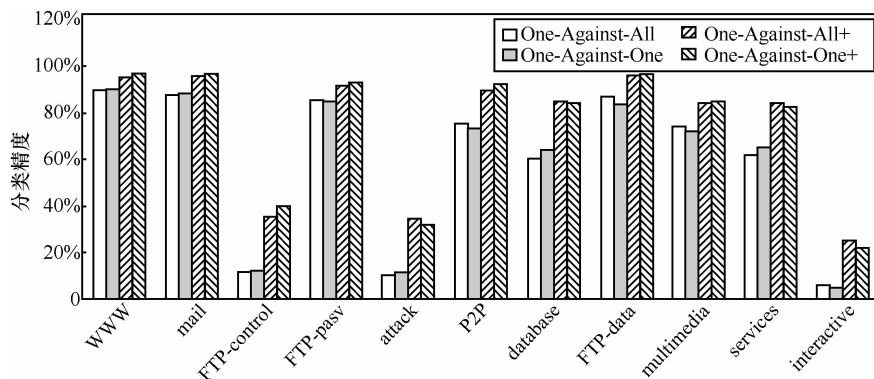


图4 4种方法在数据集1上的分类精度对比

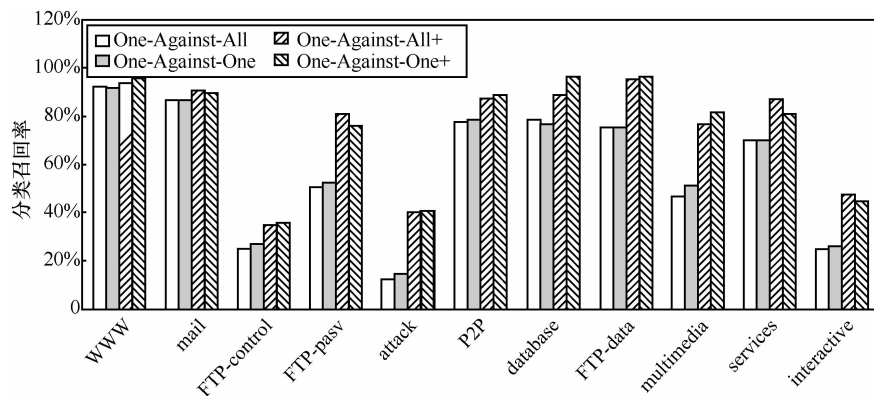


图5 4种方法在数据集1上的分类召回率对比

量(worm,virus等),因此它们的共同特征比较少,如果使用一个所有分类共享的单一的特征空间会使得很多区域叠加,难以找到一个较好的决策分离超平面。改进方法专门为攻击流量的二分类SVM选择特定的特征空间,有利于减少无关特征的干扰。实验数据表明,attack流量的精度从原来的13.4%提高到50.6%(One-Against-All+方法),15.7%提高到51.2%(One-Against-One+方法)。同样,FTP-control,interactive等原来正确率很低的分类也得到了很大的提高。

图6和图7是4种方法在数据集2上的流量分类精度和召回率的对比情况。数据集2是从校网络中心的某台交换机采集到的实际数据,本文同样对数据进行了欠抽样处理,以均衡流量类的分布。数据集2上的流量分类对比结果与数据集1相似,改进的方法使得分类正确率得到了进一步提高。在精度上的提高尤其显著:(One-Against-All+方法)QQ从64.2%提高到83.1%,P2P从72.3%提高到92.6%,games从22.5%提高到40.3%,attack从40.7%提高到67.5%;(One-Against-One+方法)QQ从63.8%提高到84.3%,P2P从66.7%提高到90.1%,games从26.6%提高到39.8%,attack从

32.4%提高到65.2%。在召回率上,改进方法也比原方法有了明显的提高。由此可见,本文提出的方法有助于进一步提高One-Against-All和One-Against-One的分类正确率。

5 结束语

机器学习方法目前应用于流量分类是一个研究热点,SVM由于其良好的泛化能力,非常适合应用于互联网这类高度动态变化的场景。SVM最初是针对两分类问题的,即SVM是典型的两分类器。但互联网流量的应用类型很多,对它们进行分类是典型的多分类问题。传统上将SVM扩展到多分类模型是通过One-Against-All和One-Against-One方法。本文发现不同的流量特征(如数据分组平均大小)对于不同的应用,其区分能力是不同的。因此,传统上采用单一的特征空间来建立这些两分类SVM显然不是最优的。本文提出可变特征空间的方法,在One-Against-All和One-Against-One的基础上,为每个两分类SVM构建独立的特征空间,这样找到的最优分离超平面优于统一的特征空间。通过两个真实的流量数据集,对比分析了各自的分

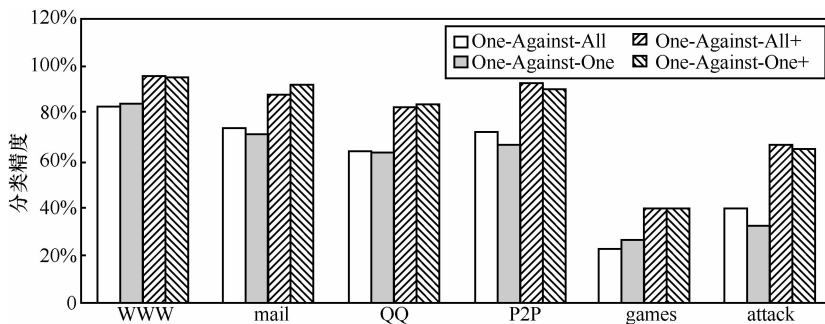


图6 4种方法在数据集2上的分类精度对比

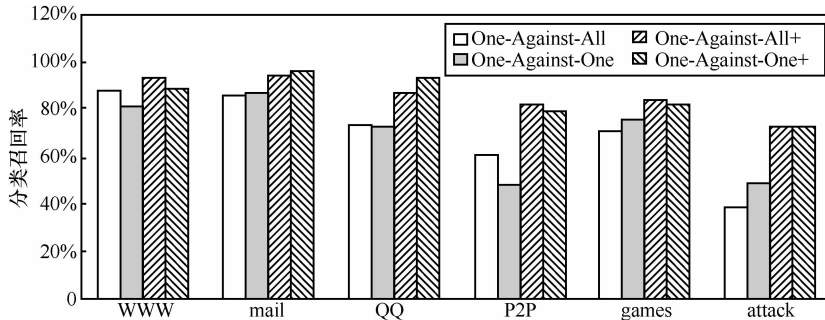


图7 4种方法在数据集2上的分类召回率对比

类正确性。实验结果表明,本文提出的可变特征空间的分类方法可以有效提高原始的 One-Against-All 和 One-Against-One 方法的分类性能。本文提出的基于机器学习的流量分类方法,目前类标签标注仍依赖于 DPI,将来拟研究主动学习等方式来解决大规模类标签标注问题。

参考文献:

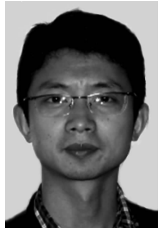
- [1] BUJLOW T, CARELA-ESPANOL V, BARLET-ROS P. Independent comparison of popular DPI tools for traffic classification[J]. Computer Networks, 2015(76): 75-89.
- [2] 钱亚冠, 张旻. 基于过抽样技术的 P2P 流量识别方法 [J]. 电信科学, 2014, 30(4): 109-113.
QIAN Y G, ZHANG M. P2P traffic identification based over-sampling technique[J]. Telecommunications Science, 2014, 30(4): 109-113.
- [3] TONGAONKAR A, TORRES R, ILIOFOROU M, et al. Towards self-adaptive network traffic classification [J]. Computer Communications, 2015(56): 35-46.
- [4] SOYSALA M, SCHMIDT E G. Machine learning algorithms for accurate flow-based network traffic classification: evaluation and comparison[J]. Performance Evaluation, 2010, 67(6): 451-467.
- [5] SINGH H. Performance analysis of unsupervised machine learning techniques for network traffic classification [C]//2015 Fifth International Conference on Advanced Computing & Communication Technologies (ACCT), Feb 21-25, 2015, Haryana, India. New Jersey: IEEE Press, 2015: 401-404.
- [6] 徐鹏, 刘琼, 林森. 基于支持向量机的 Internet 流量分类研究[J]. 计算机研究与发展, 2009, 46(3): 407-414.
XU P, LIU Q, LIN S. Internet traffic classification using support vector machine [J]. Journal of Computer Research and Development, 2009, 46(3): 407-414.
- [7] ESTE A, GRINGOLI F, SALGARELLI L. Support vector machines for TCP traffic classification [J]. The International Journal of Computer and Telecommunications Networking, 2009, 53(14): 2476-2490.
- [8] ZHOU X S. A P2P traffic classification method based on SVM[C]// The 2008 International Symposium on Computer Science and Computational Technology, Dec 20-22, 2008, Washington, DC, USA. [S.l.: s.n.], 2008: 53-57.
- [9] LI Z, YUAN R, GUAN X. Accurate classification of the internet traffic based on the svm method[C]// The IEEE International Conference on Communications, 2007 (ICC'07), June 24-28, 2007, Glasgow, Scotland. New Jersey: IEEE Press, 2007: 1373-1378.
- [10] CHANG C C, LIN C J. LIBSVM: a library for support vector machines [EB/OL]. [2001-07-20]. <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [11] KREBEL H G. Pairwise classification and support vector machines [A]//SCHOLKIPF B, BURGESS C J C, SMOLA A. Advances in kernel methods: support vector learning [M]. Cambridge: The MIT Press, 1999:255-268.
- [12] XIE G, ILIOFOTOU M, KERALAPURA R, et al. Subflow: Towards practical flow-level traffic classification [C]// IEEE INFOCOM 2012, March 25-30, 2012, Orlando, FL, USA. New

Jersey: IEEE Press, 2012: 2541-2545.

[13] GUYONG I, WESTON J, BARNHILL S, et al. Gene selection for cancer classification using support vector machines [J]. Machine Learning, 2002, 46(1-3): 389-422.

[14] MOORE A W. Dataset [EB/OL]. [2009-06-29]. <http://www.cl.cam.ac.uk/research/srg/netos/nprobe/data/papers/sigmetrics/>.

[作者简介]



钱亚冠(1976-),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、下一代互联网和机器学习与大数据处理。



关晓惠(1977-),女,浙江水利水电学院副教授,主要研究方向为机器学习与大数据处理。



云本胜(1980-),男,博士,浙江科技学院理学院讲师,主要研究方向为数据挖掘和服务计算。



楼琼(1987-),女,博士,浙江科技学院理学院讲师,主要研究方向为图像处理、机器学习与计算机视觉。



马鹏飞(1986-),男,博士,浙江科技学院理学院讲师,主要研究方向为运筹优化与机器学习。