基于快速 SVM 的大规模网络流量分类方法*

王 涛[a], 程良伦[b]
(广东工业大学 a. 自动化学院; b. 计算机学院, 广州 510006)

摘 要: 支持向量机方法具有良好的分类准确率、稳定性与泛化性,在网络流量分类领域已有初步应用,但在面对大规模网络流量分类问题时却存在计算复杂度高、分类器训练速度慢的缺陷。为此,提出一种基于比特压缩的快速 SVM 方法,利用比特压缩算法对初始训练样本集进行聚合与压缩,建立具有权重信息的新样本集,在损失尽量少原始样本信息的前提下缩减样本集规模,进一步利用基于权重的 SVM 算法训练流量分类器。通过大规模样本集流量分类实验对比,快速 SVM 方法能在损失较少分类准确率的情况下,较大程度地缩减流量分类器的训练时间以及未知样本的预测时间,同时,在无过度压缩前提下,其分类准确率优于同等压缩比例下的随机取样 SVM 方法。本方法在保留 SVM 方法较好分类稳定性与泛化性能的同时,有效提升了其应对大规模流量分类问题的能力。

关键词: 支持向量机; 大规模流量分类; 比特压缩; 权重 SVM; 分类器; 分类准确率
中图分类号: TP393 **文献标志码**: A **文章编号**: 1001-3695(2012)06-2301-05
doi:10.3969/j.issn.1001-3695.2012.06.080

Large-scale network traffic classification with fast support vector machine method

WANG Tao[a], CHENG Liang-lun[b]
(*a. Faculty of Automation, b. Faculty of Computer, Guangdong University of Technology, Guangzhou* 510006, *China*)

Abstract: SVM has been applied for network traffic classification preliminarily because of its high classification accuracy, stability and generalization. However, scaling up SVM to large-scale network traffic classification is still an open problem because of the high computation complexity as well as long training and prediction time. This paper proposed a bit-reduction based fast SVM. Firstly, it applied the bit-reduction algorithm to reduce the cardinality of the samples by weighting representative examples, and reduced the scale of training dataset with minimum loss of initial sample information. Then it developed SVM trained on weighted samples. The experiment results of large-scale network traffic classification show that bit-reduction SVM produces a significant reduction in the time required for both classifier training and prediction of unknown samples with minimum loss in accuracy. Meanwhile, its results in more accurate classifiers than random sampling based SVM when the dataset are not over-compressed. This method scales up SVM to large-scale network traffic classification with retaining the stability and generalization performance of SVM.

Key words: support vector machine(SVM); large-scale network traffic classification; bit reduction; weighted SVM; classifier; classification accuracy

0 引言

目前,面对互联网海量的各类网络应用流,大规模的流量分类已经成为紧迫的需求。随着日益扩大的互联网用户规模与新型应用的急剧增长,尤其是近年来移动互联网终端用户以及多媒体、P2P、网络游戏等应用大规模兴起,这些应用所产生的庞大网络流量带来极大的网络运行负载。利用网络流量分类手段,管理员可以统计不同应用类型的网络流量,了解用户使用网络的行为,对网络进行容量规划、流量调度,控制缓解网络拥塞。另外,在网络应用趋势分析、流量计费、动态访问控制等方面,网络流量分类都具有极其重大的作用。

传统的流量分类主要基于端口与基于分组深度解析两种方法,其分类的对象往往是常见的几种应用。然而,随着 Internet 底层环境和上层的应用发展以及规模扩大,传统的基于传输层端口的应用识别技术已逐渐不能适应 P2P 和被动 FTP 等新型应用[1]。另外,随着应用负载加密与新型应用的不断涌现[2],通常难以获取数据包负载明文以及未知应用语法与特征,因而导致基于分组深度解析方法的有效性逐步下降。

近年来较多研究人员开始采用基于流量统计特征的机器学习流量分类方法[3]。其中支持向量机(SVM)方法由于其较好的分类稳定性与泛化性,已在网络流量分类领域有初步应用[3]。一方面,SVM 方法训练分类器的复杂度取决于最终支持向量的数量,而不是特征空间的维度,因此,SVM 方法可用于解决高维样本数据的学习分类问题。另一方面,基于结构风险最小化原则,将分类问题转换为在特定约束条件下,寻找最优超平面(optimal hyperplane) 的二次规划问题,从而避免分类

收稿日期: 2011-09-22; **修回日期**: 2011-10-30 **基金项目**: 国家自然科学基金—广东省联合基金重点资助项目(U0935002);广东省重大科技专项资助项目(2009A080207008);广州市科技计划资助项目(2010Z1-D00061);广东省高校优秀青年创新人才培养计划资助项目(LYM11057)

作者简介: 王涛(1983-),男,湖北荆州人,博士,主要研究方向为网络测量、网络流量分类(wangtaosea@msn.com);程良伦(1964-),男,湖北黄石人,教授,博导,主要研究方向为人工智能与模式识别等.

器对样本分布先验概率的依赖,可以有效提高分类器在小样本情况下的分类准确率和稳定性。在较小训练样本集情况下,SVM 方法与决策树、朴素贝叶斯方法相比通常具有较高的分类准确率。

尽管原始 SVM 方法有着计算复杂度与样本特征维度关联度小、分类性能稳定且不依赖于样本分布情况等特点,但却难以满足实际应用中的大规模网络流量分类需求。这是因为:在大规模流量分类问题中,通常应用类型多种多样,网络流的应用类型分布不均衡,流统计特征动态多变,为了获得更好的分类性能,需要提取高维统计特征、构造大规模样本集来训练分类器,这就造成分类器的训练构建时间以及数据存储空间随样本数量以及特征维度增长难以继续保持有效。以本文研究对象 SVM 方法为例,在训练集规模较大时具有计算复杂度高、模型训练与样本预测速度慢的严重缺陷,其计算复杂度为 $O(n^3)$,其中 n 为支持向量的数量,而支持向量数量通常与样本数量成正比。

为提高 SVM 方法应对大规模流量分类问题的能力,本文提出一种基于比特压缩的快速 SVM 方法,即 BRSVM (bit-reduction SVM),可有效加快 SVM 方法分类器训练以及样本预测过程。该方法首先利用比特压缩算法对训练样本集进行处理,将属于同类别且类似的样本聚合为同一子类,并由此构建具有权重信息的新样本集,在损失尽量少原始样本信息的前提下缩减样本空间。进一步,利用基于权重的 SVM 算法训练流量分类器。在大样本集情况下,通过与原始 SVM 算法流量分类实验结果比较,BRSVM 方法在保证分类准确率小幅降低的同时,具有更快的分类器训练速度与样本预测速度。

1 基于机器学习的流量分类研究现状

目前,采用机器学习方法进行流量分类受到越来越多的关注,主要包括有监督与无监督机器学习两种流量分类方法。

1.1 有监督机器学习的流量分类

Roughan 等人[4]提出使用 K-近邻(K-nearest neighbors)、线性判别分析(linear discriminant analysis,LDA)、二次判别分析(quadratic discriminant analysis, QDA)机器学习方法进行网络流量应用分类。该方法共使用了分组层次、流层次、连接层次、流与连接内部特征、同一源目主机间的多条并发流五类特征,同时使用十折交叉认证评价分类方法。然而,实验结果说明该方法随着流量应用类型数量的增加,分类错误率明显上升。

Moore 等人[5]引入有监督的朴素贝叶斯(naïve Bayes,NB)机器学习方法进行流量分类与应用识别。但该方法要求样本各个特征满足条件独立并遵循高斯分布,然而实际应用中的原始网络流量特征很难满足上述条件,因此其分类准确率只有约 65% 。为解决此问题,Auld 等人[6]进一步采用特征选择方法对特征集合进行过滤,并使用核密度估计对朴素贝叶斯方法进行了改进,分类准确率得到提高,达到 95% 以上。然而朴素贝叶斯是一种传统的参数估计方法,依赖训练集样本先验概率分布,然而实际未知流量集的样本分布往往与训练集不同,因此朴素贝叶斯方法无法保证分类性能的稳定性。

王宇等人[7]提出基于 C4. 5 决策树分类器的有监督网络流量分类方法,讨论特征选择和 boosting 增强方法两种改进策略。文中实验结果表明,C4.5 分类器的训练复杂度适中,准确率高且分类速度快。徐鹏等人[8]引入 C4. 5 决策树方法来处理流量分类问题。该方法利用训练数据集中的信息熵来构建分类器,并通过对分类器的简单查找来完成未知网络流样本的分类。理论分析和实验结果都表明,与贝叶斯方法相比,利用 C4.5 决策树来处理流量分类问题在分类稳定性上均具有明显的优势,但决策树方法在高维样本学习时存在复杂度过高的问题。

徐鹏等人[9]提出一种基于支持向量机的流量分类方法。该方法利用非线性变换和结构风险最小化(structural risk minimization,SRM)原则将流量分类问题转换为二次寻优问题,具有良好的分类准确率和稳定性。然而,原始 SVM 方法在面临大样本集时具有训练速度慢、计算复杂度高的问题。

1.2 无监督机器学习的流量分类

Zander 等人[10]提出基于 autoclass 方法对网络流量进行无监督学习分类,该方法使用 EM (expectation maximaization)方法从训练集中得到最佳的聚类簇,并以此训练构建分类器。同时,作者使用从不同的网络位置收集的流量来验证该方法的有效性,获得的分类平均准确率为 86.5% 。然而,此方法需要进一步研究如何确定各个聚类簇与各应用类型之间的映射关系。

Erman 等人[11]引入 EM 聚类方法来处理流量分类问题,通过与 Bayes 的分类方法进行比较,获得了更为准确的分类结果。此类方法无须已标记类型的训练样本,因此具有发现新型网络应用的能力,但此类方法通常需要手工标记各个聚类的应用类型,而且大规模样本聚类时间通常较长,需要较大的计算与存储资源。

可见,现有有监督与无监督流量分类方法在应对大规模流量分类问题时,随着流量应用类型增加、训练集样本规模扩大,其分类准确率、分类速率等通常难以兼顾并达到实际可用的结果。鉴于此,本文提出 BRSVM 方法在保留原始 SVM 较好分类稳定性与泛化性能的同时,可较大程度地提升其应对大规模流量分类问题的能力。

2 BRSVM 算法

2.1 原始 SVM 算法

本节简要说明 SVM 方法的原理[12],以基本的二元分类问题为例。SVM 方法处理二元分类问题的主要思想是:寻找最优分类超平面决策边界,使得训练样本中的两类样本能被准确分类,且此决策边界与各个分类保持最大的边距;对于线性不可分问题,通过非线性化映射(核函数)将低维空间输入向量映射到一个高维特征空间,从而将不可分问题在高维中转换为可分问题,然后在新空间中寻求最优决策边界。SVM 的具体分类方法描述如下:

设训练的样本输入为 $x_i(i=1,\cdots,n)$,对应的期望输出为 $y_i\in\{+1,-1\}$,其中,+1 与 -1 代表两类类别标志。高维空间超平面可表示为 $\boldsymbol{w}\cdot x-b=0$,其中,$\boldsymbol{w}$ 是法向矢量,与超平面正交,$b/\|\boldsymbol{w}\|$ 表示超平面的偏移量。那么,在一般线性不可分条件下,引入松弛变量 ξ 后,SVM 算法即可表示成解决一个具有约束条件的二次优化问题,如下:

$$\begin{aligned}&\text{minimize}:\frac{1}{2}\|w\|^2+C\sum_{i=1}^{n}\xi_i\\&\text{subject to}:y_i\{(\boldsymbol{w}\cdot\varphi(x_i))+b\}\geqslant 1-\xi_i\\&\xi_i\geqslant 0\quad i=1,2,\cdots,n\end{aligned}\tag{1}$$

其中,误差权重 C 为某个指定的常数,实际上起控制对错分样本惩罚程度的作用,实现在错分样本与算法复杂度之间的折中。在式(1)中引入 Lagrange 乘子后,可得到对偶优化问题如下:

$$\begin{aligned}&\text{maximize}\quad \sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i,x_j)\\&\text{subject to}\quad \sum_{i=1}^{n}y_i\alpha_i=0\\&0\leqslant\alpha_i\leqslant C\quad i=1,2,\cdots,n\end{aligned}\tag{2}$$

其中,α_i 是非负拉格朗日乘子。可以证明,在此寻优问题的解中有一部分 α_i 不为 0,它们所对应的训练样本落在两个支持超平面上,这些训练样本即称为支持向量。SVM 通过一个适当的非线性函数 $\Phi(x)$ 将数据由原始特征空间映射到一个新的高维特征空间,定义 $K(x_i,x_j)=\Phi(x_i)^{\mathrm{T}}\Phi(x_j)$ 为核函数。解决此二次优化问题即可得到高维空间中用于分割样本的最优超平面。对样本 x 的分类函数如下:

$$f(x)=\text{sign}\{\sum_{i=1}^{n}y_i\alpha_i K(x_i\cdot x)-b\}\tag{3}$$

其中,符号的正负说明样本的类别。

2.2　BRSVM 算法

SVM 方法可用于训练高精准的分类器,但应用于大样本数据集时具有训练与预测速度慢的缺陷。现有较多针对提高 SVM 算法训练与预测速度的研究,本文提出一种基于样本比特压缩的 BRSVM 方法,可大幅提高 SVM 的训练与预测速度,并保证足够的分类精确率。此算法首先对样本集进行比特压缩处理,然后利用加权支持向量机训练分类器。算法原理如下。

2.2.1　比特压缩

比特压缩技术通常用于降低数据分辨率(data resolution),在文献[13]中也将比特压缩技术用于加快模糊 C 均值聚类算法。将比特压缩用于 SVM 方法包含标准化、比特压缩与聚合三个步骤。

标准化步骤用于确保每个样本特征具有均等的分辨率,所有样本特征数据都首先经过标准化处理,处理后特征值落入区间[0,1]。随后,为了避免在其后比特压缩过程中过多样本信息的丢失,使用一个整数来代表每个标准化后的特征值,计算方法如下:

$$I(v)=\text{int}(Z\times v)\tag{4}$$

其中:Z 是放大标准化后的特征值的比例;int(k)返回 k 的整数部分;$I(v)$将用于后续的比特压缩步骤。

在比特压缩过程中,假设 b 是将压缩的比特数,则压缩过程可表示为

$$I(v)'\leftarrow I(v)>>b\tag{5}$$

其中,$k>>b$ 指将整数 k 右移 b 位,给定一个 m 维特征的样本 $x_i=(x_{i1},x_{i2},\cdots,x_{im})$,经过比特压缩后可表示为 $I(x_i)'=(I(x_{i1})',I(x_{i2})',\cdots,I(x_{im})')$。

聚合步骤将样本按经过比特压缩后 $I(x_i)'$的取值进行聚合,具有相同 $I(x_i)'$的样本落入同一个聚合集 A 中,A 中的样本可能分属不同的类别。为此,对聚合集 A 中属于各个不同类别的样本分别统计数量 W 并计算均值 mean,在计算均值时使用样本特征的原始值$(x_{i1},x_{i2},\cdots,x_{im})$。随后,属于同一类别且具有相同 $I(x_i)'$取值的样本特征值由均值 mean 替代,W 则代表其权重。尽管比特压缩处理的速度很快,但常规聚合处理方法的速率却较低,算法复杂度通常是 $O(n^2)$,其中,n 为样本数量。本文使用哈希表加快样本集聚合速度,并选择全域散列函数(universal hashing),算法复杂度可降到 $O(2n)$,这将在较大程度上加快样本聚合压缩的速度。如表 1 所示为一维样本数据$(x_i,\ y_i)$的比特压缩过程,其中,y_i 代表样本所属类别。

表 1　一维样本集比特压缩处理示例

i	原始样本 $(x_i,\ y_i)$	$I(x_i)$及其二进制表示 $Z=1000$	$I(x_i)'$ 2 位比特压缩	新样本 $(x_j,\ y_j)$	权重
1	(0.0081,1)	8(1000)	2(10)	(0.0087,1)	2
2	(0.0093,1)	9(1001)	2(10)		
3	(0.0102,2)	10(1010)	2(10)	(0.0109,2)	2
4	(0.0117,2)	11(1011)	2(10)		
5	(0.0124,2)	12(1100)	3 (11)	(0.013 ,2)	2
6	(0.0136,2)	13(1101)	3(11)		

2.2.2　权重 SVM 算法

经过比特压缩后,原始样本数据$(x_i,\ y_i)$聚合为具有权重值 W 的新样本数据$(x_j,\ y_j)^w$。假设某个样本 x_j 的权重为 β_i,则基于式(1),权重 SVM 算法可表示为如下带约束条件的二次优化问题:

$$\begin{aligned}&\text{minimize}:\frac{1}{2}\|w\|^2+C\sum_{i=1}^{n}\beta_i\xi_i\\&\text{subject to}:y_i\{(w\cdot\varphi(x_i))+b\}\geqslant 1-\xi_i\\&\xi_i\geqslant 0\quad i=1,2,\cdots,n\end{aligned}\tag{6}$$

其中,约束条件与式(1)中相同。式(6)即可表示为如下对偶形式:

$$\begin{aligned}&\text{maximize}\quad \sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i,x_j)\\&\text{subject to}\quad \sum_{i=1}^{n}y_i\alpha_i=0\\&0\leqslant\alpha_i\leqslant\beta_i C\quad i=1,2,\cdots,n\end{aligned}\tag{7}$$

对比式(2)与(5)可发现,在引入样本权重系数后,只是非负拉格朗日乘子 α_i 的限定条件有所改变。因此,只需要对原始的 SVM 训练过程的边界条件进行简单修改。基于实际的网络流量数据集,本文将比较评价原始 SVM 与 BRSVM 应用于网络流量分类时的性能。

3　实验与分析

3.1　实验数据

本文采用 Moore 等人在文献[5]中所使用的实验数据集来评测快速 SVM 方法应用于流量分类的有效性。该数据集采集自某研究机构的网络出口,此研究机构共有约 1 000 名研究人员、管理人员与技术人员,通过一条千兆的全双工以太网链路与互联网连接。采集的数据为 24 h 内流经网络出口的所有双向流量数据,由于数据量过于庞大,Moore 等人采取分段抽样的方式采集数据,共计采集了 24 h 10 个时间段内的实验数据。同时,Moore 等人在构造实验数据集时,只选取语义完整的 TCP 协议流作为网络流量样本。完整的 TCP 协议流是指以完整的 TCP 三次握手开始与 TCP 三次握手结束的 TCP 协议流量。本文实验选择其中一个高峰时段流量作为样本集,共包含流样本数量 24 668 个,此数据集的流量均通过人工分类。可以看出,WWW 类流量为主要成分。具体构成如表 2 所示。

3.2　流特征集

在 Moore 等人的研究中,对每条网络流提取了 249 个属性

特征[5]，其中有近 100 多项属性特征是经过傅里叶变换而来。在实际网络环境中，由于流的数量巨大，对每条网络流都进行傅里叶变换会带来沉重的计算负载。为满足实际应用要求，减小训练分类器的计算与存储开销、提高分类效率，提取较易获取并具有代表性的 32 项特征来描述每条网络流，如表 3 所示。在这 32 项特征中，最后一项指明流所属的应用类别。其余 31 项网络流统计特征主要分为下面三类：a）分组数量相关特征，即指与网络流中与分组数目相关的统计特征，主要包括前向发送的分组总数以及后向发送分组的总数；b）分组大小相关特征，即指与网络流中分组长度相关的统计特征，主要包括前向发送分组的最大、最小与平均长度以及后向发送分组的最大、最小与平均长度等；c）分组时间间隔相关特征，即指与网络流中时间相关的统计特征，主要包括流的持续时间，前向发送分组的最大、最小与平均到达间隔以及后向发送分组的最大、最小与平均到达间隔等。时间相关特征容易受到网络拥塞状况的影响，但在一定程度上也能够有效区分不同的网络应用。

表 2 Moore_SubSet 数据集概况

类别	应用实例	流数量	百分比/%
WWW	HTTP、HTTPS	18 211	73.8
Mail	Imap、POP2/3、SMTP	4 146	16.8
Attack	Internet worm and virus attacks	122	0.49
P2P	KaZaA、BitTorrent、GnuTella	339	1.37
Database	Postgres、sqlnet Oracle、ingres	238	0.96
Bulk	FTP-data	1 319	5.35
Services	X11、DNS、ldap、ntp	206	0.84
Multimedia	Windows Media Player、Real	87	0.35
Total	– –	24 668	– –

3.3 流量分类方法性能评估策略

针对某一机器学习分类器，模型评估是指评价分类器在未知样本集上处理分类问题的能力，其关键指标是对未知样本的预测准确率。以一个 m 元流量分类问题为例，假设在测试集中存在 N 条流量样本，分别属于 m 种网络应用类型，首先定义以下概念：

TP(true positive)：实际类型为 i 的样本中被分类器正确判定的样本数，记为 TP_i；

FN(false negative)：实际类型为 i 的样本中被分类器误判为其他类型的样本数，记为 FN_i；

FP(false positive)：实际类型为非 i 的样本被分类器误判为类型 i 的样本数，记为 FP_i。

基于以上概念，下面给出评价分类器准确性的三个常用指标，即类准确率(accuracy)、类可信度(precision)以及整体准确率(overall accuracy)的计算方式(式(8)～(10))与描述：

$$\text{accuracy}(i) = A_i = \frac{TP_i}{TP_i + FN_i} \tag{8}$$

$$\text{precision}(i) = T_i = \frac{TP_i}{TP_i + FP_i} \tag{9}$$

$$\text{overall\_accuracy} = OA = \frac{\sum_{i=1}^{m} TP_i}{\sum_{i=1}^{m} TP_i + FN_i} \tag{10}$$

在这三个评价指标中，分类器的整体准确率应用最广，它反映了分类器正确预测样本数占总样本数的比例。类 i 的准确率表示类 i 所有样本中被分类器正确预测的样本所占的比例，类 i 的可信度表示在被分类器判定为 i 类的样本中实际为 i 类的样本所占的比例。类准确率、类可信度反映了分类器对单个应用类型的预测能力。

表 3 流特征集

Number	Abbreviation	Description
1	duration	Connection duration
2	total_pkts	Total number of packets in both forward and backward direction
3	total_bytes	Total number of bytes in both forward and backward direction
4	fwd_pkts	Total number of packets in forward direction
5	fwd_bytes	Total number of bytes in forward direction
6	fwd_pkt_max	Maximum forward packet size
7	fwd_pkt_min	Minimum forward packet size
8	fwd_pkt_avg	Average forward packet size
9	fwd_pkt_var	Standard deviation of forward packet size
10	bwd_pkts	Total number of packets in backward direction
11	bwd_bytes	Total number of bytes in backward direction
12	bwd_pkt_max	Maximum backward packet size
13	bwd_pkt_min	Minimum backward packet size
14	bwd_pkt_avg	Average backward packet size
15	bwd_pkt_var	Standard deviation of backward packet size
16	fwd_iat_max	Maximum forward inter-arrival time
17	fwd_iat_min	Minimum forward inter-arrival time
18	fwd_iat_avg	Average forward inter-arrival time
19	fwd_iat_var	Standard deviation of forward inter-arrival time
20	bwd_iat_max	Maximum backward inter-arrival time
21	bwd_iat_min	Minimum backward inter-arrival time
22	bwd_iat_avg	Average backward inter-arrival time
23	bwd_iat_var	Standard deviation of backward inter-arrival time
24	fwd_push_pkts	Total number of Push packets in forward direction
25	bwd_push_pkts	Total number of Push packets in backward direction
26	fwd_urg_pkts	Total number of Urgent packets in forward direction
27	bwd_urg_pkts	Total number of Urgent packets in backward direction
28	fwd_syn_pkts	Total number of packets with the SYN flag in forward direction
29	fwd_fin_pkts	Total number of packets with the FIN flag in forward direction
30	bwd_syn_pkts	Total number of packets with the SYN flag in backward direction
31	bwd_fin_pkts	Total number of packets with the FIN flag in backward direction
32	flow_class	Class of flow

3.4 分类性能比较分析

本文利用 LIBSVM 工具包[14]实现改进的 BRSVM 算法。实验选择工具包中的 C-SVC 算法，该算法以原始 SVM 算法为基础，经修改后可支持基于权重的 SVM 算法。两种算法中都选择径向基核函数 $k(x,y) = \exp(-\gamma \| x - y \|^2)$ 来实现网络流样本的非线性映射。本文参照 LIBSVM 的使用说明[15]，利用步长搜索策略得到惩罚因子 $C = 512$ 和核参数(kernel parameter) $\gamma = 0.03125$。为了比较原始 SVM 方法与 BRSVM 方法的流量分类性能，将样本集按7∶3比例分割，其中 70% 的样本作为训练集，30% 的样本作为测试集，重复实验三次，取平均值作为实验结果。所有实验在 Windows XP 系统下完成，CPU 为 Intel Core2 Duo E7500，内存 2 GB。

针对 BRSVM 算法中的比特压缩过程，本文实验统一取 $Z = 1000$。取比特压缩位数 $b = 5$ 时，BRSVM 与原始 SVM 算法的流量分类的性能对比如表 4、5 所示。可见，在 $b = 5$ 时，经过比特压缩处理，训练样本的数量缩减为原样本数量的 51.2%，

分类器训练时间由 37.3 s 缩减至 11.7 s,但 BRSVM 方法的整体分类准确率(96.12%)与原始 SVM 方法(96.24%)相比并未明显降低。

表 4　BRSVM 与原始 SVM 算法类准确率对比

Classifier/%	WWW	MAIL	ATT	P2P	DB	FTP(20)
SVM	99.6	96.5	0	39.9	9.6	97.5
BRSVM $b=5$	99.6	96.4	1.9	39.8	11.5	97

Classifier /%	MUL	SER	*OA*	Compression ratio	Train time/s
SVM	38.8	90.2	96.24	1.0	37.3
BRSVM $b=5$	33.3	91.6	96.12	0.512	11.7

表 5　BRSVM 与原始 SVM 算法类可信度对比

Classifier /%	WWW	MAIL	ATT	P2P	DB	FTP(20)	MUL	SER
SVM	97.9	91.5	0	64.8	63.3	96.8	55.3	98.7
BRSVM $b=5$	98.1	90.9	50	59.1	91.7	96.5	55	97.8

进一步,取不同比特压缩位数时,比较 BRSVM 与随机取样时原始 SVM 算法的分类性能,如表 6 所示。可见,即使在比特压缩位数为 0 时,经过聚合后,样本集也有一定程度的缩减(压缩比为 0.952),并且其分类整体准确率基本保持不变。在比特压缩位数 b 值由 1 ~ 5 时,随着 b 值增大,训练样本数量不断缩减(压缩比由 0.913 降到 0.512),BRSVM 方法分类整体准确率依然保持较高水平,并且其分类器训练时间不断缩减(由 37.3 s 缩减到 11.7 s)。同时,在相同的训练集压缩比例下,随机抽取数量相同的样本构成训练集,采用原始 SVM 方法的分类整体准确率低于 BRSVM 方法。这是因为 BRSVM 采用样本聚合的方法,在缩减训练集规模的同时能尽量保留初始样本集信息,而采用随机抽样的方法则将部分训练样本直接去除,减少训练样本数量的同时也丢失了部分样本信息,从而使得其分类整体准确率比 BRSVM 低。除此之外,随着样本集规模降低,其支持向量数量也同步减少,使得分类器的样本预测时间也在减少。

表 6　BRSVM 与随机取样时原始 SVM 算法性能对比

比特压缩位数	压缩比	分类准确率	样本压缩时间	训练时间	测试时间	随机采样
SVM	1.0	96.24%	--	37.3s	3.78s	--
0	0.952	96.19%	0.43s	35.6s	3.75s	96.05%
1	0.913	96.18%	0.48s	32.3s	3.66s	95.63%
2	0.877	96.20%	0.49s	30.6s	3.52s	95.16%
3	0.795	96.15%	0.48s	22.5s	3.15s	93.97%
4	0.663	96.11%	0.48s	15.8s	2.94s	92.76%
5	0.512	96.12%	0.47s	11.7s	2.72s	91.82%
6	0.323	83.41%	0.36s	7.3s	1.88s	87.54%
7	0.117	68.57%	0.36s	4.1s	1.12s	79.22%

在比特压缩位数为 6 或 7 时,可以看到随机取样的原始 SVM 方法的分类整体准确率比 BRSVM 高。这是因为当比特压缩位数过大时,会导致过度聚合,大量的样本将聚为一类,由此产生的新样本集无法为分类器决策边界提供足够信息,使得分类准确率大幅降低。与此同时,在压缩比很低时,随机取样却更能保留初始样本集信息,此时原始 SVM 分类准确率要高于 BRSVM。

4　结束语

本文提出 BRSVM 方法并用于大规模网络流量分类,与传统 SVM 方法相比,BRSVM 通过对训练样本集进行比特压缩与聚合,有效缩减了训练样本规模,从而较大程度地加快分类器的训练速度以及样本预测速度,并且其分类准确率与完整样本集下 SVM 方法保持相近。同时,在样本没有过度压缩的情况下,基于相同的样本压缩比例,BRSVM 可更大程度地保留原始样本集信息,其分类准确率要高于随机取样时的 SVM 方法。本文 BRSVM 方法一定程度上解决了原始 SVM 方法用于大规模流量分类时速度慢、计算复杂度高的不足,下一步工作将研究 BRSVM 方法流量分类器的在线更新机制。

参考文献:

[1] MADHUKAR A, WILLIAMSON C. A longitudinal study of P2P traffic classification[C]//Proc of the 14th IEEE Int'l Symposium on Modeling, Analysis, and Simulation. 2006.

[2] CALLADO A, KAMIENSKI C, SZABO G, *et al*. A survey on internet traffic identification[J]. IEEE Communications Surveys & Tutorials, 2009, 11(3): 37-52.

[3] NGUYEN T, ARMITAGE G. A survey of techniques for Internet traffic using machine learning[J]. IEEE Communications Surveys & Tutorials, 2008, 10(4): 56-76.

[4] ROUGHAN M, SEN S, SPATSCHECK O, *et al*. Class-of-service mapping for QoS: a statistical signature-based approach to IP traffic classification[C]//Proc of ACM/SIGCOMM Internet Measurement Conference (IMC). 2004.

[5] MOORE A W, ZUEV D. Internet traffic classification using Bayesian analysis techniques[C]//Proc of ACM International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS). 2005.

[6] AULD T, MOORE A W, GULL S F. Bayesian neural networks for Internet traffic classification[J]. IEEE Trans on Neural Networks, 2007, 18(1): 223-239.

[7] 王宇,余顺争. 网络流量的决策树分类[J]. 小型微型计算机系统, 2009, 30(11): 2150-2156.

[8] 徐鹏,林森. 基于 C4.5 决策树的流量分类方法[J]. 软件学报, 2009, 20(10): 2692-2704.

[9] 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.

[10] ZANDER S, NGUYEN T, ARMITAGE G. Automated traffic classification and application identification using machine learning[C]// Proc of the 30th IEEE Conference on Local Computer Networks. 2005.

[11] ERMAN J, ARLITT M, MAHANTI A. Traffic classification using clustering algorithms[C]//Proc of SIGCOMM Workshop on Mining Network Data. New York: ACM Press, 2006: 281-286.

[12] WITTEN I H, FRANK E. Data mining: practical machine learning tools and techniques[M]. 2nd ed. Amsterdam: Elsevier Inc, 2005.

[13] ESCHRICH S, KE J, HALL L, *et al*. Fast accurate fuzzy clustering through data reduction[J]. IEEE Trans on Fuzzy Syst, 2003, 11(2): 262-270.

[14] CHANG C, LIN C. LIBSVM: a library for support vector machines (version 2.3)[EB/OL]. (2001)[2010-07-04]. http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

[15] HSU C W, CHANG C C, LIN C J. A practical guide to support vector classification[EB/OL]. (2003)[2010-07-05]. http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.