DOI: 10.7667/PSPC152144

基于改进随机森林算法的电力业务实时流量分类方法

许勇刚[1]，张建业[2]，龚小刚[3]，姜 珂[4]，周 欢[4]，殷继英[5]

(1. 北京中电普华信息技术有限公司，北京 100085；2. 国网新疆电力公司，新疆 乌鲁木齐 830018；
3. 国网浙江电力公司，浙江 杭州 310007；4. 华北电力大学控制与计算机工程学院，北京 102206；
5. 国家开发投资公司，北京 100034)

摘要：为了更有效地对电力业务系统安全接入过程中日渐增多的流量进行实时分类，提高电力系统的业务处理速度，提出了一种基于改进随机森林算法的电力业务实时流量分类方法。在分析电力业务安全接入实时流量特征的基础上，改进传统随机森林算法，基于分类间隔加权对随机森林进行修剪来提高分类实时性；对新的样本数据进行数据剪辑来提高分类的准确性。在此改进算法的基础上设计了电力业务安全接入实时流量分类流程。最后以某省电力公司安全接入实时流量分类为例，验证了所提方法的准确性和实时性。

关键词：随机森林；数据剪辑；分类间隔；电力业务；流量分类

A method of real-time traffic classification in secure access of the power enterprise based on improved random forest algorithm

XU Yonggang[1], ZHANG Jianye[2], GONG Xiaogang[3], JIANG Ke[4], ZHOU Huan[4], YIN Jiying[5]

(1. Beijing China Power Information Technology Co., Ltd., Beijing 100085, China; 2. State Grid Xinjiang Electric Power Co., Wulumuqi 830018, China; 3. State Grid Zhejiang Electric Power Co., Hangzhou 310007, China; 4. School of Control and Computer Engineering, North China Electric Power University, Beijing 102206, China; 5. State Development Investment Co., Beijing 100034, China)

Abstract: This paper aims to classify the growing number of real-time traffic during the secure access process of the power business system more effectively and to improve the speed of business processing of the power system. A real-time traffic classification method of the power business based on improved random forests algorithm is proposed. On the basis of analyzing characteristics of real-time traffic in secure access of the power business, traditional random forests algorithm is improved. This paper prunes random forests based on margin weight to improve real-time performance of classification and does data-editing for the new sample data to improve accuracy performance of classification. Based on this improved algorithm, a process of real-time traffic classification in secure access of the power business is designed. At last, an instance of a province's real-time traffic classification in secure access of the power enterprise is used to validate the feasibility and efficiency of the method proposed.

Key words: random forests; data editing; classification margin; power business; traffic classification

0 引言

随着电力系统不断发展[1-5]，电力系统内外网交互日益频繁、规模增大、业务种类繁多、用户数攀升、行为日趋复杂。如何管理网络访问控制、流量入侵检测、网络规划建设，提升内网边际安全是当前电力系统内网建设急需解决的问题。实时流量分类技术能够按照业务类型对在线网络流量分类，有效地减少安全接入业务的处理时间，同时定期分析表现特殊的流量以了解网络流量的发展态势，为网络优化提供决策支持。

基于流统计特征的网络流量分类技术是当前较常用的实时流量分类技术之一，它主要根据流量的某些属性，例如平均包间隔时间、平均包长等统计信息，借助机器学习的分类方法将流映射到不同的流类型。目前，在流量分类中使用较为广泛的分类技术主要有：贝叶斯、决策树、支持向量机(SVM)[6]、随机森林(Random Forests)等。其中贝叶斯和决策树

是单分类器技术中比较有代表性的技术，但是单分类器由于自身的限制，其性能提升达到了无法超越的瓶颈[7]，于是使用多个元分类器进行分类，综合分类结果形成最终结果的多分类器组合的思想应运而生。随机森林就是在这个背景下产生的一种多分类器组合。随机森林的应用广泛：生物信息学方面，文献[8]等人使用随机森林算法研究了沙滩细菌密度与其他变量的影响关系；生态学方面，文献[9]利用随机森林算法研究土地的覆盖面积，并发现随机森林算法与其它组合算法相比训练更快；遗传学方面，Diaz-Uriarte 等人利用随机森林算法进行基因识别[10]；医学方面，文献[11]利用利用随机森林技术对肺部 CT 图像进行肺结节的自动检测。

电力企业新增业务[12]的不断涌现使得新增业务的端口更加具有随机性甚至被调用，这些都使得传统方法在电力业务安全接入实时流量分类中存在诸多不足。随着机器学习技术的不断成熟，基于流统计特征的流量分类方法成为流量分类的重要手段，而随机森林算法因其训练速度快、分类结果好、通用性广等特点最近几年在各领域分类问题上广泛使用。本文结合电力业务安全接入流量特点，提出基于分类间隔加权对随机森林进行修剪和对新的样本数据进行数据剪辑的改进随机森林算法，实现一种适用于电力系统的，分类速度快、准确性高、扩展性强的实时流量分类方法。

1　电力系统安全接入物理拓扑图

电力企业安全接入平台的物理成分[13]如图 1 所示：主要包括接入终端、安全接入网关、安全认证系统、访问控制器等。

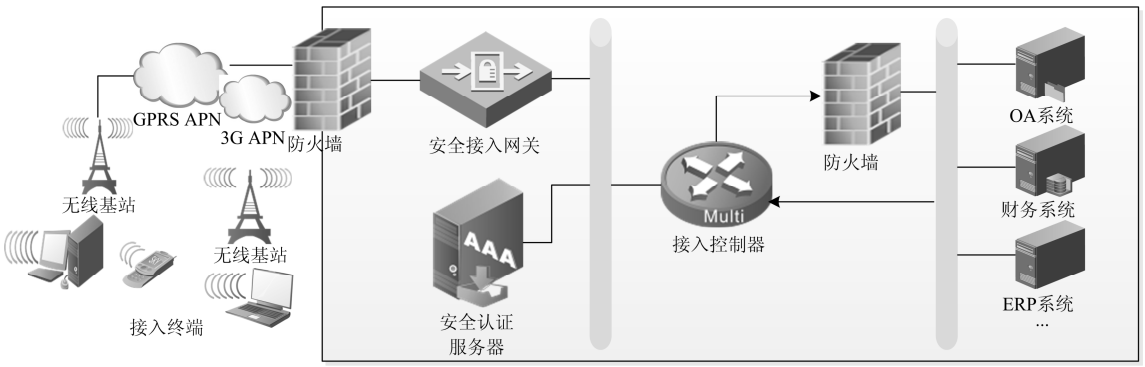

图 1 安全接入平台物理拓扑图

Fig. 1 Security access platform's physical topology diagram

安全接入主要是从终端接入、数据传输以及内网数据保护这三个方面解决了电力企业终端接入的安全性问题，为终端安全接入电力企业内网提供了一种可靠的解决方案。安全接入的流程主要包括如下步骤：

① 终端向安全接入网关发送接入请求；

② 终端和网关握手协商，建立控制通道后交换认证参数；

③ 多因素认证之后交换预共享主密钥，建立数据通道；

④ 安全接入网关解密数据包，转发给接入控制器；

⑤ 接入控制器过滤包之后，选择合适的发送端单向传输装置将包传输给相应的内网应用系统；

⑥ 应用系统接收并处理请求，将处理结果传输回安全接入网关；

⑦ 安全接入网关接收到应用系统的相应信息后封装成数据包，发送给移动终端；

⑧ 最后终端发送 FIN 消息断开连接并清空缓存。

在安全接入整个过程中，安全接入网关有着举足轻重的作用，既用于转发接入、访问请求，又用于将应用系统的返回的信息传回相应的接入终端，是企业内网与移动专网的唯一接口。因此对接入流量进行业务分类的过程应部署在安全接入网关设备中，判断其性能的标准是分类的准确性和实时性，分别如下介绍。

(1) 准确性评价指标

准确性是指在实验或调查中某一实验指标或性状的观测值与其真值的接近程度，是流量分类技术的关键评价指标。本文提出以召回率(recall)和精度(precision)两项指标来评价分类结果的准确性。召回率和精度的计算方法[14]为

$$recall = \frac{TP}{TP + FN} \tag{1}$$

$$precision = \frac{TP}{TP + FP} \tag{2}$$

式中：TP、FN、FP、TN 分别代表真正(true positive)、假负(false negative)、假正(false positive)、真负(true negative)。表 1 阐明了四者之间的关系。

表 1 TP、FN、FP、TN 之间的关系

Table1 Relationships among TP, FN, FP and TN

Classified as	X	$\bar{X}$
X	TP	FN
$\bar{X}$	FP	TN

(2) 实时性评价指标

实时性是指实时系统必须对外来事件在限定时间内做出反应，能够反映流量分类技术在线、快速识别流量业务类型的能力。本文提出以固定流量在加了实时流量分类方法实现包之后的安全接入网关中的停留时间与未加之前的停留时间的时间差 Δt 来评价分类的实时性。

$$\Delta t = t_1 - t_0 \tag{3}$$

式中：t_1 代表的是固定流量在加了实时流量分类方法实现包的安全网关中的停留时间；t_0 代表固定流量在初始状态下的安全网关中的停留时间。

2 改进随机森林算法的实时流量分类方法

2.1 改进的随机森林算法

2.1.1 随机森林算法基本原理

随机森林是一个由一系列决策树分类器 $h(x, \theta_k), k = 1, \cdots, n$ 组成的分类器集合，其中 θ_k 表示独立同分布的随机向量，且每个决策树都为输入变量 x 的类别归属进行预测[15]。随机森林通过 Bagging 方法生成相互之间有差异的不同训练样本集，采用分类回归树(Classification and Regression Trees，CART)作为元分类器构建集成分类器，用简单多数投票结果作为分类结果。其中 Bagging 方法是 bootstrap aggregating 的缩写，其主要思想是用算法训练多轮，每轮的训练集从初始训练集中随机抽取(又放回)得到，训练之后可得到一个预测函数序列，最终的预测函数对分类问题采用投票方式。

随机森林的具体过程如下：

① 给定训练集 S，测试集 T，特征维数 F。确定参数：使用到的分类回归树的数量 t，每棵树的深度 d，每个节点使用到的特征数量 f;

② 从 S 中通过 Bagging 方法有放回的抽取 t 个训练集 $S(i)$;

③ 每一个训练集 $S(i)$用于构建一棵分类树，t 个训练集产生 t 个分类树。单棵树的生长过程为：在树的每个内部节点处，从 F 个特征中随机挑选 f 个特征作为候选特征，按照节点不纯度最小的原则从 f 个候选特征汇总选择一个最优特征对节点进行分裂生长。终止条件：每一棵数的每个叶子节点的不纯度达到最小。

④ 统计建好的 t 棵分类树中每一棵树的投票结果，投票数最多的那一类即为未知样本的预测类别。

2.1.2 随机森林算法的改进策略

大量的电力系统业务如输变电状态监测、移动作业平台、供电电压监测、营销一体化缴费平台等需要通过电力通信网进行传输，使得电力系统对于通信网的依赖性在不断增大。因此电力通信网要具有很高的实时性、安全性和准确性，才能保证电力系统的正常运行。电力系统中的很多业务，如用电信息采集业务、电力营销等也都有实时性和准确性的需求。在电力系统安全接入业务流量分类问题中，针对实时性和业务分类准确性的需求，提出如下改进措施。

(1) 基于分类间隔加权对随机森林进行修剪

随机森林在做分类决策时，树的数目过多会使分类时间过长，影响分类的实时性；同时每棵树在参与最终决策时的权重都设置成一样的，这可能忽略了不同树对于样本判别的重要性会不同的情况。基于此，本文提出增加基于分类间隔加权对随机森林进行修剪的过程，减少树的数目的同时，增加对分类间隔贡献度较大的树的权重，既提高了分类的准确性，也提高了分类的实时性。

在集成分类器的研究中，分类间隔(margin)作为一个研究要素，在分类器集成中扮演了重要的角色。对于给定一个样本和投票方式的情况下，集成分类器中的分类间隔被定义为集成分类器在该样本上正确分类的票数与判为其他类的最大投票数之间的差值。集成分类器中分类间隔的具体定义如下所述。

给定一个集成分类器 $H_k = \{h_1, h_2, \cdots, h_k\}$ 和样本集合 $S = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$，样本 (x_n, y_n) 所对应的分类间隔定义为

$$mg(x_n, y_n) = av_k I(h_k(x_n) = y_n) - \max_{j \neq y_n} av_k I(h_k(x_n) = j) \tag{4}$$

其中：$I(\bullet)$ 为指示函数，统计投票数；$av_k(\bullet)$ 为取平均值函数。

由集成分类器中分类间隔的定义引出随机森林中分类间隔的定义如下所述。

对于一个给定的样本 (x_n, y_n)，在随机森林 H_k 的每一棵树 h_k 上，分类间隔的计算公式为

$$mg(x_n, y_n, H_k) = \frac{1}{|H|}\left(\sum_{i=1}^{|H|} I(h_i(x_n) = y_n) - \sum_{i=1}^{|H|} \max_{j \neq y_n} I(h_i(x_n) = j)\right) \quad (5)$$

其中，$|H|$ 是随机森林中决策树的数目。

(2) 对新的样本数据进行数据剪辑

在随机森林的建立过程中，不断加入新的置信度高的样本对于提高分类模型的性能和泛化能力具有重要意义。因此随机森林利用带标记样本训练得到各元分类器，组成森林，然后对无标记数据进行预测，将置信度高的样本加入到训练集合中，然后利用新的训练数据重新进行分类器的训练。但是一些被错误标记样本的存在，使得这些错误样本会影响分类模型的性能。因此，在传统的随机森林算法的训练过程中增加了对新增样本进行数据剪辑的过程，减少其中错误标记样本的数目，提高了分类的准确性。

基于最近邻规则的 Depuration 技术是一种应用原型选择的数据剪辑技术，它分为 RemoveOnly 和 RelabelOnly 两个部分。 其中，RelabelOnly 仅将样本进行移除操作，而 RelabelOnly 仅将样本标签进行修正[16]。文献[16]中通过实验证明： Depuration 的剪辑效果仅与 RelabelOnly 相当，且二者都没有 RemoveOnly 效果好。因此，本文只选择使用 Remove Only 操作进行数据剪辑操作。

2.2 基于改进随机森林算法的实时流量分类方法

图 2 是改进随机森林算法分类方法的流程图。主要分为随机森林的建立过程、基于分类间隔加权的修剪过程和新样本数据进行数据剪辑的过程。

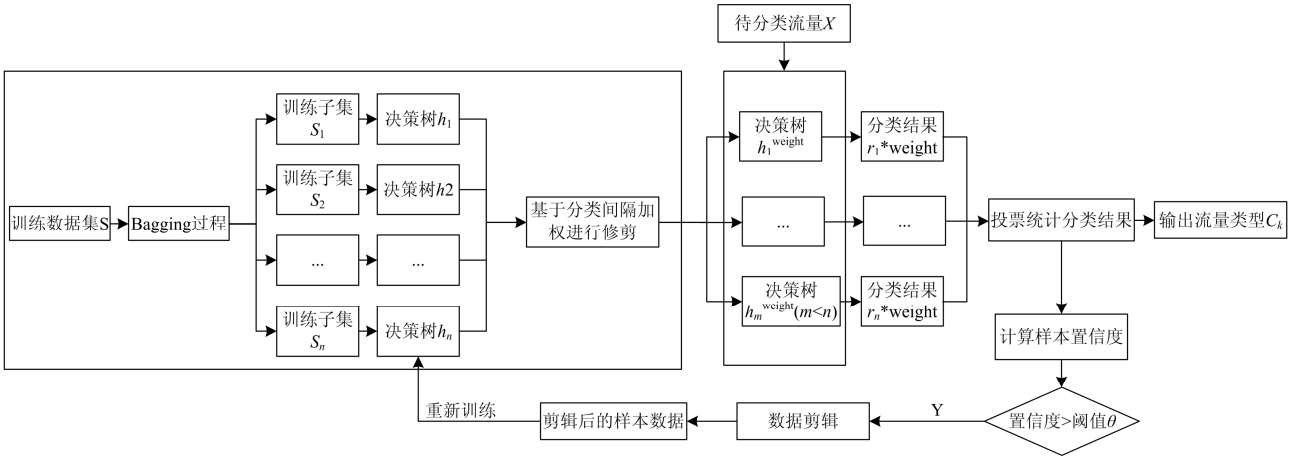

图 2 改进随机森林算法分类流程图

Fig. 2 Classification flowchart based on improved random forest algorithm

2.2.1 基于分类间隔加权的修剪过程

基于分类间隔加权[17]的修剪过程采用后向递归消除的方式，在每一次迭代中，删除对分类间隔影响最小的树，并根据重要性排序结果对各棵树进行加权。

假设初始的随机森林为 H，基于分类间隔加权修剪随机森林的具体过程如下：

① 计算森林中每一棵树对随机森林分类间隔的重要性，即通过把该棵树从当前森林中移除，计算 margin 的改变量；

② 根据每棵树 margin 的改变量对数进行排序，改变量越小说明该棵树对森林的重要性越小；

③ 挑选其中改变量最小的树 $h_{\min}$ 进行删除，随机森林缩减为 $H' = \{H / h_{\min}\}$；

④ 重新对 H' 中每一棵树计算分类间隔，并以分类间隔作为量化指标对树进行加权，权重进行归一化处理。得到一个带权重的分类器子集序列 H^k_{weight}；

⑤ 重复①-④，知道达到某个停止准则，即森林中树的个数达到一定的值。

在上述过程中，计算某棵树 h_i 的权重是通过把 h_i 从随机森林 H 中移除，计算 margin 的平均改变量得到的， margin 的平均改变量通过度量函数 $f(h, H, S)$ 计算得到。

对于一棵给定的决策树 h_i，其度量函数的定义为

$$f(h_i, H, S) = \underset{(x,y)\in S}{av}\left(mg(x, y, H) - mg(x, y, \{H \setminus h_i\})\right) \quad (6)$$

其中： $av(\bullet)$ 表示求平均值； $mg(x, y, H)$ 表示未删除任何树时样本 (x, y) 的分类间隔； $mg(x, y, \{H \setminus h_i\})$ 表示删除了树 h_i 后样本 (x, y) 的分类间隔。

2.2.2 新样本进行数据剪辑的过程

通过训练得到的随机森林分类器 $H_k=\{h_1,h_2,\cdots,h_k\}$，其中对于分类器集合 H 中的每个决策树 h_i，其余 n-1 个决策树所组成的分类器集合，称为 h_i 的对等分类器集合，记做 H_i。使用对等分类器集合 H_i 对无标签数据执行多数投票的方式进行预测，使用标签的一致性表示样本的置信度，选择置信度大于默认阈值的新样本，加入到决策树 h_i 的训练子集 S_i 中，重新训练得到新的决策树。对森林中的每棵树都执行上述过程，直到所有的决策树不再发生变化。

数据剪辑[18]操作的具体过程如下：

① 为新标记训练集 S' 中每个样例，按最近邻规则从 $S\bigcup S'$ 中选取 k 个近邻；

② 观察 k 个近邻中是否有 k' 个近邻的标记相同；

③ 若有，则保留此新样例；若没有，则该信仰里被识别为“可疑”的错误标记样例，将其从 S'中移除。其中，当 k 和 k' 设为 3 和 2 时[19]，实际剪辑效果最好。

2.3 改进随机森林算法的实时流量分类方法流程

基于改进随机森林算法的实时流量分类方法由四个模块组成：流量采集模块、流量特性统计模块、流量分类模块和分类结果处理模块。图 3 表示基于改进随机森林算法的实时流量分类方法的整体流程。

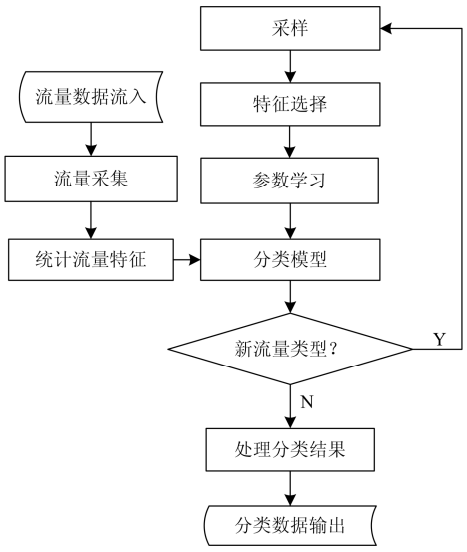

图 3 实时流量分类方法的整体流程图

Fig. 3 Overall flowchart of real-time traffic classification method

Step1：流量采集模块采集数据流量；

Step2：将采集到的流量交由流量特性模块，统计采集到的数据包的 IP 分组层特性和传输层特性，然后对数据包进行流汇聚，即把源 IP 地址、目的 IP 地址、源端口、目的端口和传输层协议相同的包划分为一个流[14]，再统计每个流的特性；

Step3：根据事先已经建立好的实时流量分类方法，按照特征属性进行业务分类[15]；

Step4：如果分类结果是已知流量类型，分类结果处理模块将分好的业务类别打包封装发送给相应的业务系统；如果分类结果是未知流量类型，那么则需重新进行样本学习的过程，来学习新流量类型的特征。

3 实验与结果分析

3.1 实验过程

本文以某省电力公司 2015 年 7 月-9 月的安全接入实时流量分类为例来验证基于改进随机森林分类算法的实时流量分类方法的准确性和实时性。实验的基础数据如表 2 所示。共收集了 1000 个样本集，600 个是学习样本，后 400 个是测试样本，每一个样本记录的是一个小时内的流量数据。

表 2 实验基础数据

Table 2 Basic data of experiment

硬件环境	Inter E7400
软件环境	Windows 7
实验工具	Weka version 3.6

(1) 利用学习样本构建初始随机森林(ntree=80)，并计算此时森林中每一棵树的重要性。从当前森林中移除重要性最低的一棵树，更新森林，重新计算剩余每一棵树对 margin 的重要性来赋予不同的权重，保存当前的森林及森林中每棵树的权重。计算加权的森林子集在测试样本上的分类准确率。重复以上操作，直至森林中的树的数目降低至 20 棵。该实验重复 20 次，计算森林规模从 80 到 20 之间每一次的测试准确率均值表 3(实验参数见表 3)。

表 3 实验参数

Table 3 Parameters of experiment

分裂节点分裂属性个数	$\sqrt{d}$ (d 为特征个数)
森林初始规模	80
森林最终规模	20
k	3
k'	2

(2) 选用最佳的森林规模时的分类模型，使用 JAVA 语言开发基于改进随机森林算法的实时流量分类方法的实现包，并将其嵌入到 Weka 系统中。

将测试样本一分为二，每组 200 个。用本文提出的改进随进森林算法和 Weka 系统中的传统随机森林算法、朴素贝叶斯算法、支持向量机算法(SVM)对测试样本进行分类，最后将实验结果进行对比。

(3) 将实现包嵌入到安全接入网关中，将测试样本一分为二，每组 200 个。统计加了实时流量分类方法实现包后，测试样本在安全接入网关里停留的时间，得到每组中 200 个样本的平均停留时间，并与之前的作对比。

3.2 结果分析

图 4 为测试准确率与森林规模关系图，表明了森林规模从 80 到 20 之间测试准确率的变化趋势。其中横坐标代表随机森林规模，纵坐标代表 20 次重复实验的平均测试准确率。

由图 4 中可以看出，随着森林规模的减小，测试的准确率会随之改变。在森林规模大于 30 时，测试准确率下降幅度不大，森林规模小于 30 时，测试准确率会受到较大影响。在森林规模为 33 时，测试准确率为 91.9%，虽然略低于当森林规模为 78 时的准确率(92.1%)，但是由于森林规模大幅减少，分类的时间也大幅缩短，有效提高了系统的实时性，故将森林规模为 33 时的随机森林模型作为最佳的分类模型，用于之后的实验。根据表 4 的实验数据，分别得到改进随机森林算法与传统随机森林算法、朴素贝叶斯算法、支持向量机(SVM)算法的精度和召回率的对比图，分别为图 5 和图 6。

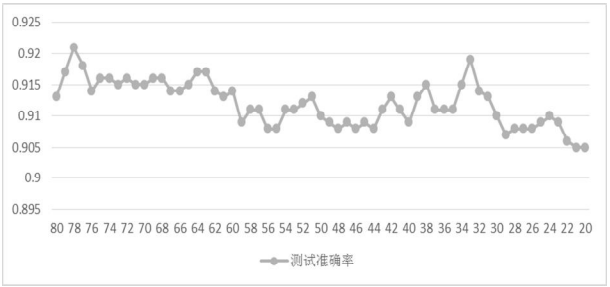

图 4 测试准确率与森林规模关系图

Fig. 4 Chart of test accuracy and forest-scale

表 4 分类方法的准确性

Table 4 Accuracy performance of the classification method

样本集序号	传统随机森林		朴素贝叶斯		SVM		改进随机森林	
	精度	召回率	精度	召回率	精度	召回率	精度	召回率
1	90.50%	92.03%	85.50%	87.35%	89.50%	91.46%	92.50%	94.89%
2	90.00%	91.87%	87.50%	89.78%	90.00%	91.87%	91.50%	93.75%

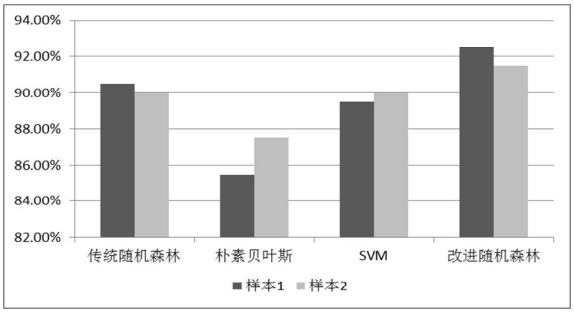

图 5 精度对比图

Fig. 5 Accuracy comparison chart

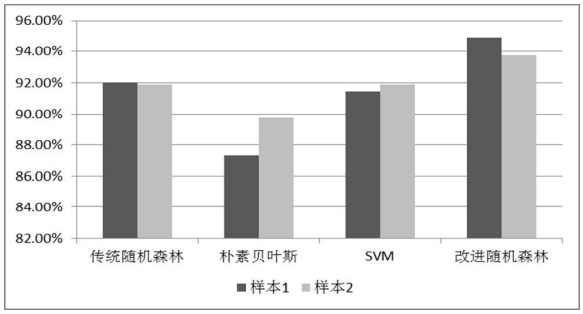

图 6 召回率对比图

Fig. 6 Recall comparison chart

经上述实验结果证明，与传统随机森林算法、朴素贝叶斯算法、支持向量机(SVM)算法相比，该方法的识别精度分别提高了 1.94%、6.36%、2.51%，召回率分别提高了 2.58%、6.49%、2.89%。满足电力业务安全接入实时流量分类的准确性的要求(每个样本经过安全接入网关的平均时间见表表)。

表 5 每个样本经过安全接入网关的平均时间

Table 5 Average time of each sample through the secure access gateway

样本集序号	未加实时流量分类方法实现包	加了实时流量分类方法实现包	时间 Δt
1	9.3 ms	10.2 ms	0.9 ms
2	9.5 ms	10.3 ms	0.8 ms

图 7、图 8 分别为 1 号样本集、2 号样本集中每个样本的前后停留时间对比图。

经实验结果证明，采用本文所提出的实时流量分类方法实现包之后的安全接入网关中的停留时间与未加之前相比，增幅分别为 9.68%和 8.42%，但都在 1 ms 之内，满足电力业务安全接入实时流量分类的实时性的要求。

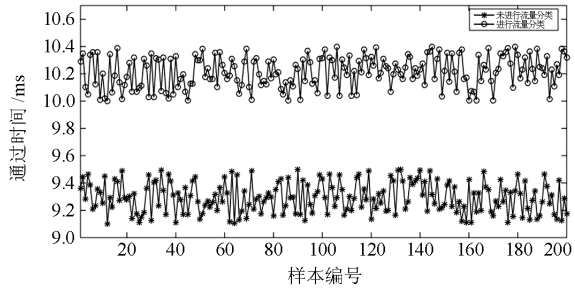

图 7 1 号样本集每个样本的前后停留时间对比图

Fig. 7 Residence time before and after comparison chart of each sample in the 1st sample set

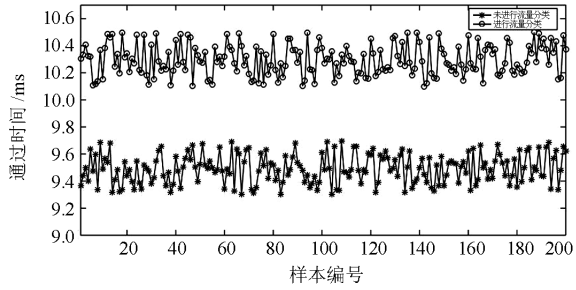

图 8 2 号样本集每个样本的前后停留时间对比图

Fig. 8 Residence time before and after comparison chart of each sample in the 2nd sample set

4 结论

本文在传统随机森林算法的基础上，增加了基于分类间隔加权的修剪过程和新样本进行数据剪辑的过程，提出一种基于改进随机森林算法的电力业务安全接入的实时流量分类方法，可以对电力业务安全接入的流量按业务类型进行实时分类。经实验结果验证其准确性高、实时性好。

参考文献

[1] 杨贵, 吕航, 袁志彬, 等. 智能变电站过程层网络流量控制和同步方法研究与实现[J]. 电力系统保护与控制, 2015, 43(11): 70-74.
YANG Gui, LÜ Hang, YUAN Zhibin, et al. Research and realization of intelligent substation process level network flow control and synchronization method[J]. Power System Protection and Control, 2015, 43(11): 70-74.

[2] 赵昆, 邹昱, 邢颖, 等. 电力系统实时动态监测主站系统检测评估方法研究[J]. 电力系统保护与控制, 2014, 42(10): 72-76.
ZHAO Kun, ZOU Yu, XING Ying, et al. Detection and evaluation on power system real time dynamic monitoring master station system[J]. Power System Protection and Control, 2014, 42(10): 72-76.

[3] 孟建良, 刘德超. 一种基于 Spark 和聚类分析的辨识电力系统不良数据新方法[J]. 电力系统保护与控制, 2016, 44(3): 85-91.
MENG Jianliang, LIU Dechao. A new method for identifying bad data of power system based on Spark and clustering analysis[J]. Power System Protection and Control, 2016, 44(3): 85-91.

[4] 党存禄, 张宁, 邵冲. 电力系统无功优化研究综述[J]. 电网与清洁能源, 2014, 30(1): 8-14, 26.
DANG Cunlu, ZHANG Ning, SHAO Chong. Review of reactive power optimization in power system[J]. Power System and Clean Energy, 2014, 30(1): 8-14, 26.

[5] 王惠中, 侯璟琨, 赵凯, 等. 基于云计算的电力系统扩展短期负荷预测[J]. 电网与清洁能源, 2014, 30(6): 1-4, 10.
WANG Huizhong, HOU Jingkun, ZHAO Kai, et al. Extended short-term load forecasting in power system based on the cloud computing[J]. Power System and Clean Energy, 2014, 30(6): 1-4, 10.

[6] 律方成, 金虎, 王子建, 等. 基于主成分分析和多分类相关向量机的 GIS 局部放电模式识别[J]. 电工技术学报, 2015, 30(6): 225-231.
LÜ Fangcheng, JIN Hu, WANG Zijian, et al. GIS partial discharge pattern recognition based on principal component analysis and milticlass relevance vector machine[J]. Transactions of China Electrotechnical Society, 2015, 30(6): 225-231.

[7] 曹正凤. 随机森林算法优化研究[D]. 北京: 首都经济贸易大学, 2014.
CAO Zhengfeng. Study on optimization of random forests algorithm[J]. Capital University of Economics and Business, 2014.

[8] PARKHURST D F, BRENNER K P, DUFOUR A P, et al. Indicator bacteria at five swimming beaches — analysis using random forests[J]. Water Research, 2005, 39(7): 1354-1360.

[9] GISLASON P O, BENEDIKTSSON J A, SVEINSSON J R. Random forests for land cover classification[J]. Pattern Recognition Letters, 2006, 27(4): 294-300.

[10] DÍAZ-URIARTE R, DE ANDRES S A. Gene selection and classification of microarray data using random forest[J]. BMC Bioinformatics, 2006, 7(1): 1.

[11] LEE S L A, KOUZANI A Z, HU E J. Random forest based lung nodule classification aided by clustering[J].

Computerized Medical Imaging and Graphics, 2010, 34(7): 535-542.

[12] 汪强, 徐小兰, 张剑. 一种新的智能变电站通信业务安全隔离技术的研究[J]. 电力系统保护与控制, 2015, 43(17): 139-144.
WANG Qiang, XU Xiaolan, ZHANG Jian. A new method of smart substation communication service security isolation technology[J]. Power System Protection and Control, 2015, 43(17): 139-144.

[13] 吴克河, 崔文超, 何健平. 电力企业移动安全接入平台[J]. 计算机系统应用, 2014, 23(7): 31-36.
WU Kehe, CUI Wenchao, HE Jianping. Wireless security access platform in power utilities[J]. Computer Systems & Applications, 2014, 23(7): 31-36.

[14] 柏骏, 夏靖波, 吴吉祥, 等. 实时网络流量分类研究综述[J]. 计算机科学, 2013, 40(9): 8-15.
BO Jun, XIA Jingbo, WU Jixiang, et al. Survey on real-time traffic classification[J]. Computer Science & Applications, 2013, 40(9): 8-15.

[15] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.

[16] JIANG Y, ZHOU Z H. Editing training data for KNN classifiers with neural network ensemble[J] // Advances in Neural Networks–ISNN 2004. Springer Berlin Heidelberg, 2004: 356-361.

[17] SÁNCHEZ J S, BARANDELA R, MARQUÉS A I, et al. Analysis of new techniques to obtain quality training sets[J]. Pattern Recognition Letters, 2003, 24(7): 1015-1022.

[18] YANG F, LU W, LUO L, et al. Margin optimization based pruning for random forest[J]. Neurocomputing, 2012, 94: 54-63.

[19] 谢永芳, 蒋有为, 唐明珠. 一种基于数据剪辑的半监督最邻近分类算法[C] // Proceedings of the 2011 Chinese Control and Decision Conference (CCDC). 2011: 41-45.
XIE Yongfang, JIANG Youwei, TANG Mingzhu. A semi-supervised K-nearest neighbor algorithm based on data editing[C] // Proceedings of the 2011 Chinese Control and Decision Conference (CCDC). 2011: 41-45.

收稿日期：2015-12-09；　修回日期：2016-02-05

作者简介：

许勇刚(1974-)，男，本科，高级工程师，研究方向为信息安全。E-mail: xuyonggang@sgitg.sgcc.com.cn

(编辑　姜新丽)