基于机器学习的网络流量特征选择

孙 振
(上海大学通信与信息工程学院 上海 200072)

摘 要：互联网技术水平不断提高的同时，也带来了日益复杂的网络安全问题，传统地利用端口检测和深度包检测等流量识别技术已经难以应对当下日趋复杂的网络环境。伴随着机器学习理论的成熟，机器学习方法已经成功的应用于图像识别、声音辨别、医疗等各个领域，机器学习使用计算机模拟人类的活动，通过学习现有的知识，建立有效的学习模型，进一步对未知的数据进行预测或者分类。将机器学习方法应用在网络流量识别领域，首先对网络流量识别的研究现状和机器学习作了相关的介绍，其次基于3种机器学习分类算法，对比分析了不同特征选择算法对网络流量识别准确率的影响，提出了改进的特征选择算法，并经过实验验证了改进后特征选择算法的有效性。

关键词：机器学习；流量识别；特征选择；对称不平衡性
中图分类号：TN91 **文献标识码**：A **国家标准学科分类代码**：510.50

Research of network flow feature selection based on machine learning

Sun Zhen
(School of Communication and Information Technology, Shanghai University, Shanghai 200072, China)

Abstract: With the development of Internet technology, it also brings an increasingly complex network security issues. The traditional traffic recognition technology such as port detection and deep packet inspection has been difficult to deal with the current increasingly complex network environment. As the theory of machine learning become mature, it has been successfully applied in many subject areas such as Image or voice recognition and medical fields. Machine learning methods simulate the human cognitive pattern by computers. The target of machine learning is to establish learning model by studying the existing knowledge and use the learning model to class or predict unknown data. In this research, machine learning methods were applied in internet traffic identification. Firstly, we introduced the research status of internet traffic identification and the relevant concepts of machine learning. Secondly, the main work is to research and compare the influence of different feature selection to identification accuracy based on three machine learning classification algorithms. The author proposed an improved feature selection algorithm and verified the effectiveness of this algorithm by experiments.

Keywords: machine learning ; traffic identification ; feature selection; symmetrical uncert

1 引 言

网络技术的飞速发展主要造成了两个方面的问题：网络安全威胁以及网络资源不合理，这给网络服务质量[1](QoS)造成了巨大的挑战。网络流量识别技术基于数据流以及报文信息判断流量具体所属的上层应用类别，这对于网络监管、网络安全以及实现访问控制和内容审计等[2]都具有基础性的重要作用。流量识别技术主要经历了3个阶段的发展：端口检测、深度包检测以及深度流检测，当前研究的热点在于基于机器学习方法的深度流检测技术。

机器学习方法应用于网络流量识别领域的前提是提取网络流量统计特征进而构建多样本数据集(包含流量类别标签)并将之作为机器学习的先验知识，机器学习通过学习这一先验知识构建出一个分类模型，使用已构建的分类模型达到对未知类型流量的分类。然而数据集中流量的统计特征有些是与类别是无关的，同时特征彼此之间也可能存在一定的冗余，这些无关和冗余特征会严重影响到分类的速度和精度，因此在进行机器学习构建分类模型之前对数据集中的特征进行特征选择是非常有必要的。

本文对机器学习中的特征选择阶段做了相关的研究。CFS算法基于特征与类别以及特征彼此间的相关性对特征进行选择，信息增益算法仅考虑单个特征对分类的贡献值并结合后续学习算法设置合适阈值完成特征选择，针对信息增益算法没有考虑到特征之间的冗余性，本文引入对称

收稿日期：2016-10

不平衡性改进信息增益算法，希望改进后的算法通过进一步删除冗余特征以达到提高分类准确率的目的，最后在 WEKA 平台上、分别基于 3 类学习算法下对比了 3 种特征选择方法的分类效果。

2　流量识别与机器学习

2.1　流量识别技术综述

根据 RFC3917，一个流是指在一个特定的时间间隔内通过网络中的观测点的 IP 分组集合，属于同一个流的分组具有相同的特性集合[3]。

网络流量识别是指利用网络数据流以及数据流中报文的某些信息将网络上的流分成既定的（比如长短流、快慢流）或者基于应用类型的不同类别[4]。流量识别技术介绍如下。

1）基于端口号检测：这种方法是检测网络数据包的传输层端口号，根据 IANA（internet assigned numbers authority）发布的端口号与网络应用的映射确定具体应用类别[5]。该方法易于实现，时间复杂度低，但同时分类精确度得不到保障。

2）深度包检测（deep packet inspection，DPI）：应用层载荷存在唯一标识应用类别的特征字符串，通过对网络流数据包与已知的网络应用特征数据库进行匹配，从而确定网络流应用类别[6]。该方法速度快、准确率高，缺点是可扩展性低，具有一定的滞后性，由于是基于应用层，可能会侵犯个人隐私，存在一定的安全隐患且不能识别加密流量。

3）深度流检测（deep flow inspection，DFI）：不同的网络应用表现出不同的流量统计特征，深度流检测便是基于一条流的多个数据包之间呈现的某种统计规则或者本身状态的一种识别方法[7]，这种方法不需要深入到应用载荷部分，避免了侵犯用户隐私的风险且可识别加密流量，当前针对 DFI 技术的研究热点是如何利用机器学习方法提高 DFI 流量分类的速度和精度。

2.2　机器学习综述

机器学习（machine learning，ML）是近年来兴起的一门融合多个领域的交叉学科，其目标是通过学习已有的先验知识构建学习机，同时使用该学习机对未知的数据进行分类或者预测，并在这一过程中通过自我学习不断完善已构建的学习机，进而提高学习机的分类性能。

根据学习形式的不同，机器学习可以概括为 3 个类别：监督学习、无监督学习、半监督学习。

1）有监督学习是一种人工参与的学习方式，通过已给出的先验数据进行训练并建立预测模型，先验数据不仅包括输入的样本集合 $X=(x_1,x_2,\cdots,x_n)$，而且涵盖所有样本所属的类别标记 $Y=(y_1,y_2,\cdots,y_n)$，y_i 表示对应样本 x_i 的类别，有监督学习使用 $<x_i,y_i>$ 关系作为输入，建立满足 X、Y 关系的近似表达式 f 作为分类器模型，并使用该分类器对未知样本进行分类。经典的有监督学习算法比如朴素贝叶斯、C4.5 决策树、支持向量机、K 近邻等[8]。有监督学习流程如图 1 所示。

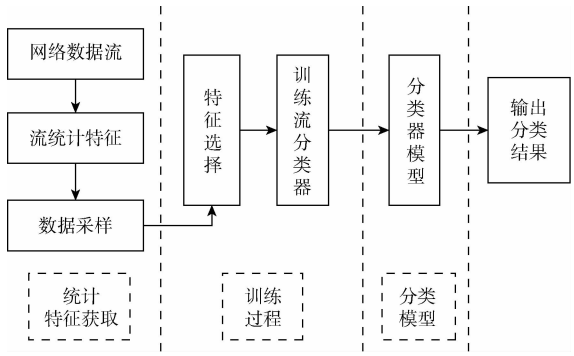

图 1　有监督机器学习流程

2）无监督学习不需要确定数据样本集的类别标签，根据聚类原则，计算样本之间的相似度，并以此对样本进行分组，每一个分组都认为是一个类别。与有监督学习不同的是，无监督学习并不直接给出数据的具体类型，之于网络流量识别上，得到的是流量的类型 1，类型 2，…，却不知道这些类型具体对应的是哪种网络应用，因此无监督学习方法并不适用于网络流量识别，常用的无监督学习算法比如 K-means 聚类、KMM 分割[9] 等，无监督学习流程如图 2 所示。

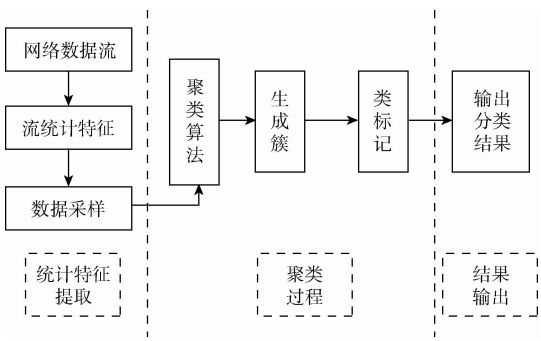

图 2　无监督学习流程

3）半监督学习[10] 是有监督学习与无监督学习相结合的一种学习方法，它主要考虑如何利用少量的标记样本和大量的未标记样本进行训练和分类的问题，这种学习方式更贴近于人类的学习过程，使分类器可以如人一般的触类旁通、举一反三，然而半监督学习目前主要用于处理人工合成数据，使用范围局限在实验室内，也就是说其理论价值还没有在现实应用中具体体现出来，另外，这种学习的抗干扰性比较弱。

2.3　特征选择概念

特征选择是指从原始特征数据集中找出对流分类效果最优的特征子集，特征选择可以删减掉不相关和冗余的特征，降低数据集特征维度，简化分类器模型，同时提高分类器的准确率，典型的特征选择流程如图 3 所示。

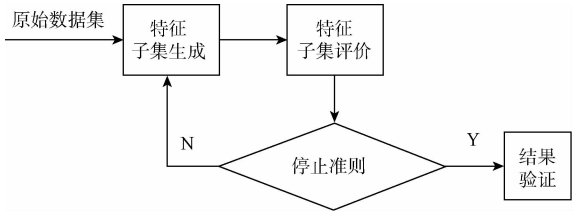

图 3　特征选择流程

生成特征子集是指对原始数据集按照特定的搜索策略进行搜索,搜索策略分为完全搜索(比如分支定界算法等)、随机搜索(比如遗传算法[11]等)和启发式搜索(比如 SFFS、SBFS 等);特征子集评价是指根据评价函数对生成的特征子集进行评价,根据评价函数和学习算法的关系,可以将评价函数分为独立评价函数和非独立评价函数;停止准则指设定的阈值或迭代次数,循环至满足停止准则,输出最优特征子集;结果验证也即验证输出特征子集的分类效果,通过分析后续分类器的建模时间以及准确率等来验证特征子集的有效性。

根据特征选择算法与分类算法的关系,将其分为过滤式(Filter)、封装式(Wrapper)与嵌入式(Embedded)3 种。

过滤式特征选择算法认为特征选择和机器学习两个过程是彼此独立的,利用特征内在的独立信息选出特征子集,常用评价标准包括距离度量、一致性度量、信息度量以及关联度量[12]。这种方法具有较好的通用性,缺点是选择出的特征子集未必是与后续的机器学习算法相匹配的最优特征子集。

封装式特征选择算法直接对所选的特征子集利用后续的机器学习算法来进行训练,根据分类器的结果评价特征子集的优劣,这种方法优点是对不同的分类算法得到的特征子集更有针对性,缺点是筛选的过程要进行大量的计算,不适合用于高维度的特征选择,且通用性较低。

嵌入式特征选择算法是指将特征选择嵌入到机器学习算法中,特征选择和训练过程同时进行,由分类模型决定是否使用特征,比如人工神经网络[13]。

2.4　算法评估标准

为评价机器学习算法的优劣引入混淆矩阵,如表 1 所示。混淆矩阵的每一列代表样本的预测类别,每一行代表样本的真实类别,比如 N_{ij} 的取值也即实际类别为 i 却被分类器预测为类别 j 的样本数目。

表 1　混淆矩阵

预测类别				实际类别
Class 1	Class 2	…	Class n	
N_{11}	N_{12}	…	N_{1n}	Class 1
N_{21}	N_{22}	…	N_{2n}	Class 2
⋮	⋮	⋮	⋮	⋮
N_{n1}	N_{n2}	…	N_{nn}	Class n

基于混淆矩阵给出如下 3 个指标的定义,其中 TP 表示某一类别中正确分类的样本数目,FN 表示某一类别中被错误分类为其他类别的样本数量,FP 表示被分类器错误分类为某一类别的样本数量。

1)整体准确率 $Accuracy$:

$$Accuracy = \frac{\sum TP}{\sum (TP + FN)} \tag{1}$$

2)召回率 $Recall$(类别 i):

$$\mathrm{Re}call = \frac{TP_i}{TP_i + FN_i} \tag{2}$$

3)精度 $Precision$(类别 i 准确率):

$$\mathrm{Pr}ecision = \frac{TP_i}{TP_i + FP_i} \tag{3}$$

4)误报率($FP\%$)、漏报率($FN\%$):

$$FP\% = \frac{\sum FP}{\sum (FP + TP)} \tag{4}$$

$$FN\% = \frac{\sum FN}{\sum (FN + TP)} \tag{5}$$

上述指标仅从分类准确性的角度评价分类算法,实际上评价算法的好坏应从多方面进行考量,比如实时性、可扩展性、鲁棒性等等。

3　特征选择算法及其改进

3.1　CFS 算法

CFS(correlation-based feature subset)特征选择算法根据特征之间的冗余度搜索特征子集,目的是找到与类别相关度高且特征之间相关度低的特征,以此构建出特征子集,这种方法在找到与类别强相关的基础上有效地消除了冗余特征和无关特征。CFS 对特征子集 s 的评价函数表述为:

$$R_s = k \cdot \bar{r}_{cf} / \sqrt{k + k \cdot (k-1) \cdot \bar{r}_{ff}} \tag{6}$$

式中:R_s 表示相关系数,k 为特征子集含有特征的个数,$\bar{r}_{cf}$ 是各个特征与类之间相关系数的平均值,$\bar{r}_{ff}$ 特征之间相关系数的平均值,以 BestFirs 搜索策略为例,算法流程描述如表 2 所示。

该算法易于实现,时间复杂度为 $O(nm^2)$,其中 m 是特征的个数,n 为样本的数量,运行时间较短。

3.2　信息增益算法

信息增益(info gain,IG)算法基于信息度量标准,通过计算每个特征对分类的信息大小来判断特征的重要程度,属于过滤式特征选择分类,其结果是基于单个特征对分类的贡献对特征进行由大到小的排序,通过结合后续的学习算法选定合适的阈值,进而得出特征子集。信息增益的计算公式如下。

$$I(X,Y) = H(X) - H(X|Y) \tag{7}$$

其中 $H(X)$表示随机变量 X 的信息熵,$H(X|Y)$表示

表 2　基于关联规则的特征选择流程

输入：$X=\{x_1,x_2,\cdots,x_k\}$ //含 k 个特征
输出：特征子集 X_l
步骤 1：X_l 初始为空，从 X 中取出一个特征添加到 X_l 使得 $C(X_l)$ 最大，记录其评估值 $c_{\max}$；
步骤 2：如果 X 中还有未被添加到 X_l 的特征，则执行步骤 3；
步骤 3：从 X 中取出一个新的特征加入到 X_l 使得 $C(X_l)$ 最大，记录其值 $c_{\max}$，如果 $c_{\max}$（$c_{\max}$ 则 $c_{\max}=c_{\max}$，返回步骤 2；否则结束算法，$c_{\max}$ 即为最佳评估值，此时得到的 X_l 即是最优特征子集；
步骤 4：如果 X 为空，则直接返回 X_l 为最优特征子集。

已知随机变量 Y 后 X 的条件信息熵，数学表达式如下。

$$H(X)=-\sum_i P(x_i)\log_2(P(x_i)) \tag{8}$$

$$H(X|Y)=-\sum_j P(y_j)\sum_i P(x_i|y_j)\log_2(P(x_i,y_j)) \tag{9}$$

信息增益值越大，表示特征对分类的贡献越大，也即该特征对分类越重要[14]，算法流程描述如表 3 所示。

表 3　信息增益算法流程

输入：样本数据集
输出：特征信息增益值排序
步骤 1：输入数据集，初始化零矩阵 I 用于存放特征权重，计算样本特征数 n；
步骤 2：计算类别的信息熵 $H(Y)$；
步骤 3：for $i=1:n$ 计算已知特征 X_i 后 Y 的条件信息熵 $H(Y\|X_i)$ 及其信息增益值 $I(X_i,Y)=H(Y)-H(Y\|X_i)$；
步骤 4：根据信息增益值降序排列特征；
步骤 5：结束。

从信息增益算法的流程中可以看出，信息增益算法在进行特征选择时只考虑了特征与类别之间的相关性，忽略了特征之间的冗余度。

3.3　基于对称不确定性改进信息增益算法

上述的 CFS 特征选择算法既考虑到特征与类别之间的相关性，同时也计算了特征之间的冗余性，然而信息增益特征选择算法仅根据计算特征对于类别的信息增益值并进行排序，结合学习算法选定阈值进而确定特征子集，这种方法筛选出的特征彼此之间是有可能存在冗余的，针对信息增益算法的缺点，本文基于对称不确定性（symmetrical uncert，SU）改进信息增益特征选择算法，进一步计算特征彼此之间的冗余度，删除冗余度较高且信息增益较低的特征，并基于三种机器学习算法在 weka 平台上做了相关实验，以期得到更精简的特征子集和更高的分类精确度。

SU 是由前面讲述的信息增益与信息熵计算特征之间的相关性，假定 X 和 Y 代表两个特征，则其相关性计算公式如下：

$$SU(X,Y)=\frac{2I(X,Y)}{H(X)+H(Y)} \tag{10}$$

$SU(X,Y)$ 的取值范围为[0,1]，取值越大则代表特征 X 和特征 Y 之间的相关性越高，当取值大于设定的阈值时，表示特征 X 与特征 Y 之间存在冗余，此时删除两者之间信息增益值较低的特征，在经过信息增益特征选择算法得到的特征子集上基于对称不确定性进一步筛选出冗余的特征，最后认为得到的特征子集即为最优特征子集。改进后的特征选择算法流程如图 4 所示。

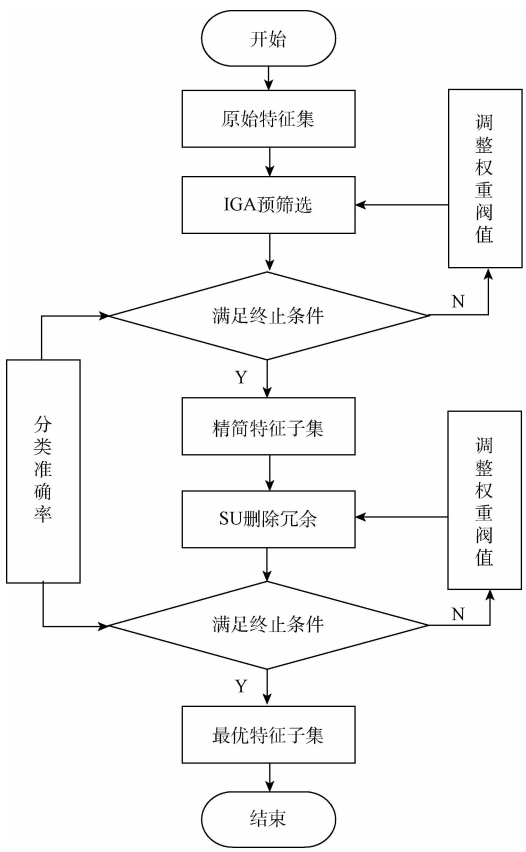

图 4　基于对称不确定性改进信息增益特征选择算法流程

4　改进特征选择算法的分类性能研究

4.1　数据挖掘工具 WEKA 简介

怀卡托智能分析环境（waikato environment for knowledge analysis，WEKA）是由来自新西兰怀卡托大学的 Witten 教授主导开发的一款开源的数据挖掘工作平台，

集合了大量能承担数据挖掘任务的机器学习算法,包括对数据进行预处理、分类、回归、聚类、关联规则以及在新的交互式界面上的可视化[15]。

WEKA 主界面为 weka GUI 选择器,可供用户选择的主要有 4 种应用程序。

1)探索者(Explorer),系统提供的最容易使用的图像用户接口,通过选择菜单填写表单即可调用 weka 的所有功能。

2)知识流(KnowledgeFlow),知识流界面弥补了探索者界面的缺陷,它使用增量方式的算法处理大规模数据集,按照一定顺序将代表数据源、预处理工具、学习算法、评估手段以及可视化模块的各构件组合在一起,形成数据流。

3)实验者(Experimenter),通过该界面,用户可以让处理过程实现自动化且更容易使用不同参数设置分类器和过滤器。

4)简单命令行(Simple CLI),命令行界面可以直接执行 weka 命令。

在本课题研究中使用 Explorer 探索者界面来完成实验工作。

4.2　摩尔数据集

摩尔(Moore)数据集是由剑桥大学教授 Moore 在一个网络中心分若干个时间段采集而来,以完整的 TCP 双向流为研究对象,共采集 377 526 个流样本,分为 Entry01～Entry10 共 10 个数据集,定义了 248 个流特征,这些特征按照"序号,简称,详细描述"的形式来定义,第 249 项表明流对应的应用类别,涵盖了常见网络应用的 10 种类别,比如 WWW、FTP、P2P 等,部分特征及其描述如表 4 所示。

表 4　摩尔数据集描述

序号	简称	详细描述
1	Server Port	服务器端端口号
2	Client Port	客户端端口号
3	min_IAT	流中包到达时间间隔的最小值
4	q1_IAT	流中包到达时间间隔的第一四分位数
⋮	⋮	⋮
249	Classes	流量应用类型

4.3　实验与结果分析

在 10 个摩尔数据集中,选取 entry03 数据集合作为训练数据集,并对此数据集分别使用 CFS 特征选择算法、信息增益特征选择算法以及基于对称不确定性改进的信息增益特征选择算法进行特征筛选,对得到的 3 组特征子集均采用朴素贝叶斯、J48 决策树和支持向量机(SVM)3 种机器学习算法建立分类模型,WEKA 设置为十折交叉验证模式,其余的 9 个数据集作为测试数据集,对测试结果取平均值,不同方法所得特征如表 5 所示,3 种机器学习算法下测试所得分类整体准确率如图 5 所示。

表 5　3 种特征选择方法得到的特征

选择算法	选取的特征标识号	特征数
CFS	4,72,78,108,113	6
IGA	4,107,108,201,98,194,94,200,193,198,96	11
IGA-SU	4,107,201,194,200,198	6

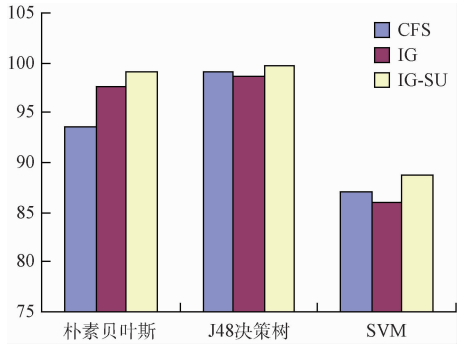

图 5　3 种特征选择方法的整体分类准确率

通过上述表格特征数的对比以及分类整体准确率的对比,不难发现:在 J48 和 SVM 两种机器学习算法下,CFS 算法选出的特征子集比信息增益算法选出的特征子集有着较高的准确率,然而基于对称不确定性改进信息增益算法后得到的特征子集,在 3 种机器学习算法下都比 CFS 算法得到的特征子集有着更高的整体准确率,同时改进后的特征选择算法在降低特征子集维数上也有着良好的表现,这对于面对更高维数据集时,在节约算法运行时间、降低计算复杂度都有着很重要的现实意义。

5　结　　论

本文针对基于机器学习的网络流特征选择做了一定的研究,提出了基于对称不确定性改进信息增益特征选择算法,分别对 3 种特征选择方法得到的特征子集在 3 种机器学习算法建立的分类器上做整体准确率的比较,实验结果表明改进后的特征选择方法对整体分类准确率和降低特征维数都具备了一定的可行性和有效性。

参考文献

[1]　林闯,李寅,万剑雄. 计算机网络服务质量优化方法研究综述[J]. 计算机学报,2011,34(1):1-14.

[2]　王海忠. 基于决策树的网络流量分类系统的设计与实现[D]. 北京:中国科学院大学,2014.

[3]　李天枫,王劲松,王立学,等. 基于 IPFIX 的大规模网络异常流量检测机制研究[J]. 天津理工大学学报,2015,31(3):1-5,11.

[4] 陈亮,龚俭,徐选. 应用层协议识别算法综述[J]. 计算机科学,2007,34(7):73-75.
[5] 王一萍,宋广军. 一种基于端口检测的主机防护系统的研究[J]. 微计算机信息,2009,25(12):80-82.
[6] 潘志浩,杨博文,曹炳尧. 基于网络处理器的深度包检测系统的研究[J]. 微计算机信息,2009,25(27):115-116.
[7] 张潇晓. 网络流量分析关键技术研究与系统实现[D]. 长沙:国防科学技术大学,2012.
[8] 李荣雨,程磊. 基于 SVM 最优决策面的决策树构造[J]. 电子测量与仪器学报,2016,30(3):342-351.
[9] 胡桂香,李宁,邢艳肖,等. 基于 KMM 与超像素的 SAR 海面暗斑分割算法[J]. 国外电子测量技术,2016,35(6):101-108.
[10] SEEGER M. Learning with labeled and unlabeled data[R]. Technical Report, Institute for Adaptive and Neural Computation, University of Edinburgh, 2001:1-7.
[11] XING H, LU C, ZHANG Q. Frequency modulated weak signal detection based on stochastic resonance and genetic algorithm[J]. Instrumentation, 2016, 3(1):4.
[12] 崔文玲,潘静,何改云,等. 基于类心和特征加权的特征选择算法[J]. 电子测量技术,2015,38(3):26-29.
[13] 焦敬品,李勇强,吴斌,等. 基于 BP 神经网络的管道泄漏声信号识别方法研究[J]. 仪器仪表学报,2016,37(11):2588-2596.
[14] WU G, XU J. Optimized approach of feature selection based on information gain[C]. IEEE International Conference on Computer Science and Mechanical Automation (CSMA), 2015:157-161.
[15] 陈慧萍,林莉莉,王建东,等. WEKA 数据挖掘平台及其二次开发[J]. 计算机工程与应用,2008,44(19):76-79.

作者简介

孙振,工学硕士,主要研究方向为网络信息安全、嵌入式软件设计等。
E-mail:sunzhen1013@163.com

(上接第 130 页)

[18] 韩丹,方之龙. 应急通信建设现状和技术分析[J]. 中国应急救援,2016(1):43-46.
[19] 谢小军,梁本仁. 基于 4G 的应急通信在安徽电力的应用[J]. 通信技术,2013,46(4):84-86.
[20] 行业应用[EB/OL]. [2016-9-30] http://e.huawei.com/cn/solutions/industries.
[21] ALHAD K, KHALID A. A real world evaluation of push to talk service over IMS and LTE for public safety systems[C]. IEEE 10th International Conference on Wireless and Mobile Computing, Networking and Communications, WiMob 2014, 2014:365-370.
[22] 桑逾方,施玮. 不同制式指挥调度语音对讲终端互通的研究[J]. 数字通信世界,2014(1):1-7.
[23] RAZA A. LTE network strategy for smart city public safety[C]. IEEE International Conference on Emerging Technologies and Innovative Business Practices for the Transformation of Societies, EmergiTech 2016,2016:34-37.
[24] KUMBHAR A,GUVENC I. A comparative study of land mobile radio and LTE-based public safety communications[C]. Proceedings of IEEE Southeastcon,2015.
[25] 何晨光,魏守明,苏阳,等. 我国警用通信专网与公网比较研究[J]. 警察技术,2015(3):25-27.

作者简介

闫复利,1982 年出生,学士学位,工程师,摩托罗拉系统(中国)有限公司上海分公司工程师,上海大学在职工程硕士研究生,主要研究方向为无线通信。