一种基于有监督局部决策分层支持向量机的异常检测方法

徐琴珍　　杨绿溪
(东南大学信息科学与工程学院　南京　210096)

摘　要：该文针对包含多种攻击模式的高维特征空间中的异常检测问题，提出了一种基于有监督局部决策的分层支持向量机(HSVM)异常检测方法。通过 HSVM 的二叉树结构实现复杂异常检测问题的分而治之，即在每个中间节点上，通过信息增益准则构建有监督学习所需的训练信号，监督局部决策；在每个嵌入中间节点的二分类支持向量机(SVM)的训练过程中，以局部决策边界对特征的敏感度为依据，选择入侵检测的局部最优特征子集。实验结果表明，该文提出的异常检测方法能够在训练信号的局部决策监督下构建具有良好稳定性的检测学习模型，并能以更精简的特征信息实现检测精确率和检测效率的提高。

关键词：异常入侵检测；分层支持向量机；特征信用度；有监督局部决策
中图分类号：TP391　　**文献标识码**：A　　**文章编号**：1009-5896(2010)10-2383-05
DOI: 10.3724/SP.J.1146.2010.00321

A Supervised Local Decision Hierachical Support Vector Machine Based Anomaly Intrusion Detection Method

Xu Qin-zhen　　Yang Lü-xi
(*School of Information Science and Engineering*, *Southeast University*, *Nanjing* 210096, *China*)

Abstract: This paper dedicates to propose a supervised local decision Hierachical Support Vector Machine (HSVM) learning model for anomaly intrusion detection in high dimensional feature space. The binary-tree structure of HSVM presents a "divide-and-conquer" algorithm for complex anomaly intrusion detection problem, i.e., the training signal for supervising local decision at each internal node is constructed according to information gain criterion. The embedded SVMs at internal node are trained on local optimized feature subsets standing on the sensitivity degrees of a margin to features. The experimental results suggest that the proposed anomaly intrusion detection method can gain learning model with better stability under the local decision supervisal of training signals. Further, it also achieves competitive detection accuracy and higher detection efficiency with condensed feature information.

Key words: Anomaly intrusion detection; Hierachical Support Vector Machine (HSVM); Feature credit; Supervised local decision

1　引言

入侵检测技术是网络安全防护系统构成的重要环节，通过从计算机网络系统中收集的若干关键信息分析网络中是否存在入侵行为。根据不同的入侵检测分析方法，网络入侵检测技术可分为滥用检测和异常检测两类[1]。滥用检测技术已经广泛应用于绝大多数商用网络入侵检测系统，对已知的攻击模式能实现高效检测，而对未知的攻击模式无法做出预测；而异常检测技术通过建立主体的正常行为模型，发现异常行为，从而能对未知攻击作出预测，异常检测作为一个开放性研究课题，已经受到越来越广泛的关注。

在机器学习任务中，给定样本特征集的情况下，异常检测问题可以将看作高维特征空间中的多分类预测问题：从给定的各维特征数据中学习到检测所需的最佳特征信息组合，构建决策超曲面，实时准确地预测出“正常”或“攻击”访问。针对异常检测问题的常用机器学习方法包括：(1)基于符号式学习模型的检测方法，学习结果可以表示成明确的推理规则集。例如，文献[2]提出了一种基于数据挖掘技术的 RIPPER 规则算法，通过遗传算法(GA)与 RIPPER 算法相结合的检测方式，抽取有效特征和构建检测规则集；Cheng 等人[3]提出了一种基于有监

2010-03-29 收到，2010-06-29 改回
国家自然科学基金(60702029，60902012)，国家科技重大专项(2009ZX03003-004)，国家 973 计划项目(2007CB310603)和东南大学科研启动费(4004001041)资助课题
通信作者：徐琴珍　summer@seu.edu.cn

督决策树与无监督贝叶斯聚类法相结合的异常检测方法，实现检测率的提高和误检率的下降。(2)基于非符号式学习模型的检测方法，学习结果以权值、系数或其他数值序列的形式存储。例如，基于人工免疫系统的检测方法：Jamie 等人[4]提出的以多级信息源为输入数据的人工免疫系统入侵检测方法，Dasgupta 等人[5]提出的运用基于免疫算法的技术检测和描述网络入侵模式。基于神经网络的方法：Thomas 等人[6]运用通过多种检测方法的组合实现入侵检测精确率的全局优化，有监督学习的神经网络模型用于调整检测某种特定入侵方式的学习方法的权重，即用于衡量多种检测方法集合中某种方法的有效性；基于支持向量机(SVM)的方法：Charles 等人[7]采用支持向量机与线性判决分析相结合的方法来提高支持向量机的异常检测精确率和线性判决分析的速度。此外，基于非符号式学习模型的异常检测方法还包括基于遗传算法的检测法、基于隐马尔可夫过程的异常检测技术、基于粗糙集的异常检测方法等[1]。

这些方法为异常检测提供了卓有成效的技术支持，并为进一步的研究提供了坚实的理论和实践基础，同时也遇到了共同的问题：包含多种攻击的异常检测问题是一个具有高维特征空间的高度复杂的多分类问题；检测所需的最佳特征信息组合往往无法先验获得，而冗余的特征信息往往会意外增加学习算法搜索解空间的复杂度，降低学习的效率；此外，冗余的特征信息还可能增加决策曲面的复杂度，影响学习结果的泛化性，甚至造成“维度灾难”。为此，本文提出了一种基于有监督局部决策的分层支持向量机(HSVM)异常检测方法。通过 HSVM 的树型结构在训练信号的监督下实现复杂异常检测问题的“分而治之”，并在每个层次上，为当前的局部决策曲面选择最优的特征信息子集，简化问题空间，降低学习模型的复杂度，从而提高检测的泛化性和效率。

2 支持向量机

支持向量机是由 Vapnik 等学者最早提出的一种基于结构风险最小化思想的机器学习方法[8]，它集成了机器学习领域的最大间隔超平面、Mercer 核、凸二次规划等多项技术，在包括异常检测在内的若干挑战性应用场景中表现出了优良的性能[9]。支持向量机中最简单的模型是针对线性可分情况下的最大间隔分类器，即给定线性可分样本：$S=\{(x_i,y_i)\mid i=1,2,\cdots,l\}$，其中 x_i 为第 i 个观测样本，$y_i\in\{+1,-1\}$ 为 x_i 对应的类别，求解最优化问题：$\min\|w\|_2^2$, s.t. $y_i(w\cdot x_i+b)\geq 1$，$i=1,2,\cdots,l$，得到最大几何间隔为 $\gamma=1/\|w\|_2$ 的最大间隔超平面 (w,b)。

在本文研究的异常检测问题中，特征空间无法线性分开，需要引入松弛变量 ξ_i 和惩罚项参数 C，需要求解的最优化问题转化为：$\min\|w\|_2^2+C\sum_{i=1}^{l}\xi_i^2$, s.t. $y_i(w\cdot x_i+b)\geq 1-\xi_i$，$i=1,2,\cdots,l$。针对这一情况，可以引入核函数 $K(\bullet)$ 在隐式定义的特征空间中实现线性可分，通过拉格朗日定理可以将问题表述成对偶形式[8]：$\max\ W(\alpha)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j K(x_i,x_j)$，$\text{s.t.}\ \sum_{i=1}^{k}\alpha_i y_i=0$，$0\leq\alpha_i\leq C$，$i=1,2,\cdots,l$。由此得到的决策规则为 $f(x)=\sum_{i=1}^{l}\alpha_i y_i K(x_i,x)+b$，通过对 $f(x)$ 的符号判别实现支持向量的二分类功能。

在本文提出的方法中，支持向量机用于实现复杂决策过程中的局部最优决策，即作为 HSVM 中间节点上的嵌入学习模块。

3 基于 HSVM 学习模型的异常检测方法

3.1 HSVM 学习模型的结构

HSVM 的整体结构与二叉树类似(如图 1 所示)，中间节点实现局部决策，叶节点标识类别。区别在于在每个中间节点上嵌入了可以提取相对复杂特征信息组合的 SVM。

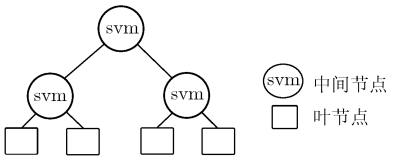

图 1 HSVM 学习模型示例

本文提出基于 HSVM 学习模型的异常检测方法，主要出于以下 3 方面的考虑：首先，二叉树结构的性质使模型能够根据局部决策训练信号将复杂的异常检测问题分解，而后在不同的层次上以相对降低的复杂度解决子问题；其次，与简单的决策树中间节点相比，嵌入 SVM 模块的节点能够在局部决策中提取更加有效的特征信息；此外，HSVM 相比于其它多分类支持向量机(如 DAGSVM, 1-V-1 SVM, 1-V-R SVM)而言，具有更高的检测效率。

3.2 训练信号的生成

对于二分类问题(如在异常检测中只需区分是正常访问或是攻击访问的情况)，训练样本中的类别标识可直接作为二叉树节点分裂时的训练信号。而对于类别数大于 2 的多分类问题(例如异常检测中，

除了检测出正常访问和攻击访问外，还需预测具体的攻击种类)，在中间节点分裂时，需要在训练信号的监督下，将包含多类的训练样本划分为两个子集，因此需要为中间节点的分裂构建局部决策训练信号。

在中间节点上，通过合适的准则构建训练信号，可以增加学习模型检测的稳定性。本文结合学习模型的二叉树结构，通过信息增益准则构建训练信号。信息增益准则是生成决策树时采用的节点分裂准则之一[10]，在 c4.5 算法中，可以通过信息增益量选取决策树中间节点分裂所需的有效特征。与之不同的是，在本文的 HSVM 树中，信息增益准则用于选择生成的训练信号，而非特征。

设中间节点上的样本集合 X 由 k 类访问模式(包括正常访问和各类攻击访问)组成：$X=\{c_1,\cdots,c_k\}$，则该样本集的信息熵为

$$\mathrm{entro}(X)=-\sum_{j=1}^{k}\left|c_j\right|/|X|\times\log_2\left(\left|c_j\right|/|X|\right) \tag{1}$$

其中 $|X|$ 表示样本集 X 中的访问样例总数，$|c_j|$ 表示 X 中包含的第 j 类访问样本数。任意给定一训练信号 T (T 将 k 类访问模式划分为正例和反例两大子类，一般分别以+1 类、−1 类表示)，将 X 划分为两个互不相交的子样本集 $X_1=\{c_{i_1},\cdots,c_{i_m}\}$ 和 $X_2=\{c_{j_1},\cdots,c_{j_n}\}$，其中 $1\le i_m,j_n\le k$，$1\le m,n\le k-1$，$m+n=k$，则该训练信号的期望信息为

$$\mathrm{entro}_T(X)=\sum_{i=1}^{2}\left|X_i\right|/|X|\times\mathrm{Entro}\left(\left|X_i\right|\right) \tag{2}$$

产生的信息增益为

$$\mathrm{gain}_T(X)=\mathrm{entro}(X)-\mathrm{entro}_T(X) \tag{3}$$

需要选择的训练信号为能够产生最大信息增益的 T^* 为

$$T^*=\arg\max_T \mathrm{gain}_T(X) \tag{4}$$

3.3 局部决策曲面上的特征选择

SVM的决策曲面边界对某一特征的敏感程度体现了该特征对分类决策的影响程度，因此，在每个中间节点的局部决策曲面训练中，选择相对有效的特征子集在一定程度上有利于促进HSVM结构的简化和异常检测泛化性的增强。给定样本集 $X=\{(x_i,y_i),i=1,2,\cdots,l\}$，和核函数 $K(\bullet)$，其中 $x_i\in\boldsymbol{R}^N$ 为第 i 次观测样本，对应的训练信号为 $y_i\in\{-1,+1\}$，则SVM决策边界的平方倒数为

$$w^2=\sum_{ij}\alpha_i\alpha_j y_i y_j K(x_i,x_j) \tag{5}$$

决策边界对于给定样本集 X 上的第 n 维特征 x_j^n 的敏感程度定义为[11]

$$D_n(w^2)=\sum_i\left|\sum_j\alpha_i\alpha_j y_i y_j\frac{\partial K(x_i,x_j)}{\partial x_j^n}\right| \tag{6}$$

若 $K(\bullet)$ 为高斯核函数，则决策边界对第 n 维特征的敏感程度为

$$D_n(w^2)=\sum_i\left|\sum_j\frac{1}{\sigma^2}\alpha_i\alpha_j y_i y_j K(x_i^n,x_j^n)(x_i^n-x_j^n)\right| \tag{7}$$

在Sindhwani等提出的基于最大输出信息的特征选择方法[11]中，第 n 维特征的信用度的衡量需要考虑两个因素：(1)SVM决策边界对于该特征的敏感度；(2)单个二分类SVM的信用度。由于多分类SVM一般由多个二分类SVM按照一定规则组合完成多分类任务，在训练过程中，所有的二分类SVM都依据特征信用度值在相同的特征子集上训练学习，从而在一定程度上导致了每个二分类SVM依然包含了部分冗余的特征信息，而这些冗余的特征信息却可能是影响其它二分类SVM决策的重要信息。对于多分类的异常检测而言，各维特征的重要性对于不同的局部决策曲面往往会随子问题而变化，即各二分类SVM局部决策时所需要的特征子集可能不同。为此，我们结合HSVM检测模型分而治之的树型结构，改进了基于最大信息输出的特征选择方法，针对每个中间节点上局部决策曲面的不同情况灵活地选择不同的特征子集，使之更适用于多分类的异常检测问题。

为了实现局部决策曲面上特征自组织选择的差异性，各维特征在不同局部决策过程中的重要性可以直接以式(7)计算的局部决策边界对该维特征的敏感度值来衡量，控制特征子集中成员的选择，同时还需要控制特征子集的规模。本文在Sindhwani等人的方法上作了改进，以分类边界对特征的敏感度值的累积量比率来优化特征的自组织选择。决策边界对特征的敏感度值的累积量比率定义为

$$\mathrm{Sr}=\sum_{i=1}^{m'}D_{n_i}\Big/\sum_{i=1}^{N}D_{n_i},\quad m'\in\{1,2,\cdots,N\} \tag{8}$$

其中 D_{n_i}，$n_i\in\{1,2,\cdots,N\}$ 为敏感度值序列 $\{D_n,n=1,2,\cdots,N\}$ 经降序排列后的结果，m' 为 Sr 达到给定的阈值 Sr^* 时选用的最佳特征子集的维数。式(8)在形式上与样本固有维数的计算类似，但有着本质的区别。样本固有维数以近邻样本点的协方差矩阵特征值的累积量为基础，为每个样本计算近邻点及其协方差矩阵的特征值，从而得出每个样本的固有维数，样本集的最终固有维数通过投票决定[12]；而 m' 的计算则依赖于边界对各维特征敏感度值 D_n，计算复杂度较固有维数的计算要低。

4 异常检测结果及分析

为验证本文提出的异常检测方法，实验选择入侵检测研究人员广泛使用的 KDD Cup 1999 入侵检测数据库中的 corrected 观测数据集[13]。

4.1 数据预处理

corrected 数据含 311029 例样本，每个观测样本含 41 维特征，在数据的预处理中，我们根据符号类别标签将访问样例标示为 4 类攻击和 1 类正常访问：正常访问样例类别标示为 1，共 60593 例；dos 攻击类别标示为 2，共 229853 例；u2r 攻击类别标示为 3，共 70 例；r2l 攻击类别标识为 4，共 16347 例；probe 攻击类别标示为 5，共 4166 例。由于 u2r 攻击样例稀少，我们将 KDD Cup 中 kddcup.data_10_percent 数据包中包含的 52 例 u2r 攻击并入到实验数据中。为改善样例的极度不平衡状况，实验从 corrected 数据集中随机抽取除 u2r 攻击外的 7878 例访问样本，与 corrected 和 kddcup.data_10_percent 中的 122 例 u2r 攻击样例构成含 8000 例访问样本的入侵检测数据集，1/3 作为训练样本，2/3 作为测试样本。实验给出的异常检测数值结果为 200 次实验结果的平均值。

4.2 数值结果对比及分析

图 2 为在训练集上生成的 HSVM 异常检测模型示例。为了在局部决策训练中最大限度地保留有用特征信息，Sr^* 阈值设为 1，即在每个中间节点上构建的特征子集中仅剔除对局部边界的敏感度值为 0 的特征。

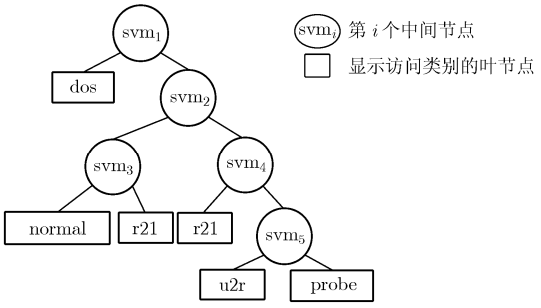

图 2 HSVM 异常检测模型示例

与图 2 相对应的各中间节点 svm_i 上特征选择的情况(特征子集规模 N，信息增益 IG)，以及训练信号 TS 对训练样本的局部决策监督情况如表 1 所示。以第 1 行数值为例说明表格中各项的关联：svm_1 的 TS 为[2]/[1 3 4 5]，表示该节点选择了当前样本下具有最佳信息增益 0.8467 的训练信号，该训练信号将第 2 类访问模式(dos 攻击)标示为正例，将 1、3、4、5 类访问模式(正常访问，u2r 攻击，r2l 攻击，probe 攻击)标示为反例。svm_1 在该训练信号监督下完成当前节点上的局部二分类决策训练，根据各维特征对局部决策边界的敏感度值，在给定的 Sr^* 下从 41 维特征中选择了 21 维对决策边界有贡献的特征构成当点节点上的特征子集。在 svm_3 和 svm_5 这两个中节点上的子样本仅包含两类访问模式，因此可以不必计算信息增益，直接构建训练信号。从图 2 及其对应的表 1 所示的训练情况说明，在不同的局部决策过程中，各 SVM 所需要的特征子集的规模和特征子集中的成员都会随着局部决策任务的变化而变化，改进的特征选择方法更好地适应了决策曲面上特征自组织选择的差异性需求。

表 1 中间节点的训练情况

svm_i	N	IG	TS(+1/−1)
1	21	0.8467	[2]/[1 3 4 5]
2	13	0.8693	[1]/[3 4 5]
3	3	–	[1]/[4]
4	27	0.9984	[4]/[1 3 5]
5	26	–	[4]/[5]

为进一步说明本文提出的异常检测方法的有效性，我们将检测结果与多种异常检测方法进行了对比：多分类支持向量机(DAG-SVM、1-v-1-SVM 和 1-v-r-SVM)[14]异常检测方法；采用启发式方法构建训练信号的支持向量机树(CSVMT)检测方法[12]；基于主分量分析法实现特征信息抽取后结合 k 近邻法实现异常检测的方法(PCA-KNN)，其中主分量分析的特征值的累积量和参照 Sr^* 取为 1；基于径向基神经网络(RBF)的异常检测方法。对比的指标包括：需要训练的二分类 SVM 数 n_{svm}，每个二分类 SVM 构建局部决策曲面需要的平均特征数 n_f，异常检测精确率 p_d 及其方差 p_{std}，虚警率 p_f 以及测试时间 t(以 HSVM 的测试时间为单位 1)，平均数值结果如表 2 所示。

由表 2 所示 HSVM 与其他异常检测方法的数值结果对比可知：(1)与多分类支持向量机相比，由于 DAG-SVM 和 1-v-1-SVM 在两两配对的攻击种类间训练二分类 SVM，需要的 SVM 数为 $k(k-1)/2$ 个，而 1-v-r-SVM 需要为某一类访问模式和剩余访问模式训练二分类 SVM，因此需要 k 个二分类 SVM，而 HSVM 则可以根据特征空间复杂度的不同，自适应地调整 SVM 的数量；(2)与具有类似分层结构的 CSVMT 相比，由于 CSVMT 采用启发式方法构建训练信号，具有很大的随机性，由检测精确率的方差对比可知，HSVM 在最大信息增益训练信号的监

表 2 不同异常检测方法的结果对比

方法	n_{svm}	n_f	p_d (%)	p_{std} (%)	p_f (%)	t
DAG-SVM	10.00	41.0	95.71	0.39	2.48	3.89
1-v-1-SVM	10.00	41.0	95.70	0.36	2.42	5.99
1-v-r-SVM	**5.00**	41.0	94.90	0.36	1.91	1.78
CSVMT	6.70	41.0	93.32	4.45	2.25	1.78
RBF	–	41.0	94.51	0.83	1.70	10.36
PCA-KNN	–	28.00	95.48	0.41	1.94	8.72
HSVM	6.17	**18.4**	**96.84**	**0.28**	**1.61**	**1.00**

督下构建的检测模型具有更好的稳定性；(3)与进行特征信息抽取的 PCA-KNN 检测方法相比，HSVM 所需的平均特征维数较小，且测试时无需进行特征坐标的转换，能够以更精简的特征信息实现异常检测；(4)此外，与包括 RBF 在内的其他检测方法相比，HSVM 获得了与其他方法相当甚至更优越的异常检测精确率和较低的虚警率；从检测率的方差还可以看出 HSVM 具有更好的稳定性；从检测效率看，HSVM 也能更好地符合实时快速检测的要求。

5 结束语

本文针对包含多种攻击模式的高维特征空间中的异常检测问题，提出了一种基于有监督局部决策的 HSVM 异常检测方法。通过 HSVM 的二叉树结构实现复杂异常检测问题的分而治之，通过信息增益准则构建中间节点分裂所需的训练信号，监督局部决策，提高检测方法的稳定性和局部决策的有效性；在检测模型的中间节点上，以局部决策边界对特征的敏感度为依据，自适应地优化入侵检测的局部最优特征子集(包括特征的选择和特征子集规模的调整)，以优化的特征子集训练中间节点上的 SVM。实验结果表明，本文提出的异常检测方法能够在在训练信号的局部决策监督下构建具有良好稳定性的异常检测学习模型，并能以更精简的特征信息实现检测精确率和检测效率的提高。

参 考 文 献

[1] Tsang Chi-ho, Kwong Sam, and Wang Han-li. Genetic-fuzzy rule mining approach and evaluation of feature selection techniques for anomaly intrusion detection[J]. *Pattern Recognition*, 2007, 40(9): 2373–2391.

[2] Helmer G, Wong J S K, and Honavar V, *et al*.. Automated discovery of concise predictive rules for intrusion detection [J]. *Journal of Systems and Software*, 2002, 60(3): 165–175.

[3] Cheng Xiang, Png Chin-yong, and Lim Swee-meng. Design of multiple-level hybrid classifier for intrusion detection system using Bayesian clustering and decision trees [J]. *Pattern Recognition Letters*, 2008, 29(7): 918–924.

[4] Jamie T and Uwe A. Information fusion in the immune system[J]. *Information Fusion*, 2010, 11(1): 35–44.

[5] Dasgupta D and Gonzalez F. An immunity-based technique to characterize intrusions in computer networks[J]. *IEEE Transactions on Evolutionary Computation*, 2002, 6(3): 281–291.

[6] Thomas C and Balakrishnan N. Improvement in intrusion detection with advances in sensor fusion [J]. *IEEE Transactions on Information Forensics and Security*, 2009, 4(3): 542–551.

[7] Charles J J, Das A, Lee B, and Seet B. CARRADS: cross layer based adaptive real-time routing attack detection system for MANETS [J]. *Computer Networks*, 2010, 54(7): 1126–1141.

[8] Cristianini N and Shawe-Taylor J. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. New York: Cambridge University Press, 2000: 93–122.

[9] Hernández-Pereira E, Suárez-Romero J A, Fontenla-Romero O, and Alonso-Betanzos A. Conversion methods for symbolic features: a comparison applied to an intrusion detection problem[J]. *Expert Systems with Applications*, 2009, 36(7): 10612–10617.

[10] Quinlan J R. C4.5: Programs for Machine Learning [M]. San Mateo, California: Morgan Kaufmann publishers, 1993: 17–26.

[11] Sindhwani V, Rakshit S, Deodhare D, Erdogmus D, Principe J C, and Nivogi P. Feature selection in MLPs and SVMs based on maximum output information[J]. *IEEE Transactions on Neural Networks*, 2004, 15(4): 937–948.

[12] 徐琴珍, 杨绿溪. 基于改进的混合学习模型的手写阿拉伯数字识别方法[J]. 电子与信息学报, 2010, 32(2): 433–438.
Xu Qin-zhen and Yang Lu-xi. An improved hybrid learning model based handwritten digits recognition approach [J]. *Journal of Electronics & Information Technology*, 2010, 32(2): 433–438.

[13] KDDCup 1999 Data, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, 2010.

[14] Hsu C W and Lin C J. A comparison of methods for multiclass support vector machines [J]. *IEEE Transactions on Neural Networks*, 2002, 13(2): 415–525.

徐琴珍：女，1977 年生，讲师，研究方向为智能信息处理、图像处理、混合学习模型、入侵检测.

杨绿溪：男，1964 年生，教授，博士生导师，研究方向为通信信号处理、移动通信中的 MIMO 空时信号处理、中继协作通信与网络编码、盲信号处理与阵列信号处理.