doi:10.11959/j.issn.1000-436x.2015127

入侵检测中基于 SVM 的两级特征选择方法

武小年[1,2,3]，彭小金[1]，杨宇洋[1]，方堃[1]

(1. 桂林电子科技大学 信息与通信学院，广西 桂林 541004；
2. 桂林电子科技大学 广西无线宽带通信与信息处理重点实验室，广西 桂林 541004；
3. 桂林电子科技大学 广西信息科学实验中心，广西 桂林 541004)

摘 要：针对入侵检测中的特征优化选择问题，提出基于支持向量机的两级特征选择方法。该方法将基于检测率与误报率比值的特征评测值作为特征筛选的评价指标，先采用过滤模式中的 Fisher 分和信息增益分别过滤噪声和无关特征，降低特征维数；再基于筛选出来的交叉特征子集，采用封装模式中的序列后向搜索算法，结合支持向量机选取最优特征子集。仿真测试结果表明，采用该方法筛选出来的特征子集具有更好的分类性能，并有效降低了系统的建模时间和测试时间。

关键词：入侵检测；特征选择；支持向量机；Fisher 分；序列后向搜索

中图分类号：TP393 **文献标识码**：A

Two-level feature selection method based on SVM for intrusion detection

WU Xiao-nian[1,2,3], PENG Xiao-jin[1], YANG Yu-yang[1], FANG Kun[1]

(1.School of Communication and Information, Guilin University of Electronic Technology, Guilin 541004, China;
2. Guangxi Wireless Broadband Communication and Signal Processing Key Laboratory, Guilin University of Electronic Technology, Guilin 541004, China;
3.Guangxi Experiment Center of Information Science, Guilin University of Electronic Technology, Guilin 541004, China)

Abstract: To select optimized features for intrusion detection, a two-level feature selection method based on support vector machine was proposed. This method set an evaluation index named feature evaluation value for feature selection, which was the ratio of the detection rate and false alarm rate. Firstly, this method filtrated noise and irrelevant features to reduce the feature dimension respectively by Fisher score and information gain in the filtration mode. Then, a crossing feature subset was obtained based on the above two filtered feature sets. And combining support vector machine, the sequential backward selection algorithm in the wrapper mode was used to select the optimal feature subset from the crossing feature subset. The simulation test results show that, the better classification performance is obtained according to the selected optimal feature subset, and the modeling time and testing time of the system are reduced effectively.

Key words: intrusion detection; feature selection; support vector machine; Fisher score; sequential backward selection

1 引言

特征选择是根据某种准则从原始特征集中选择部分最有区分类别能力的特征[1]。在入侵检测系统中，特征选择通过从网络数据集中筛选出对分类器分类性能影响最重要的最优特征子集，降低特征的维数，减少计算量，提高入侵检测系统效率。随着网络的高速发展，网络流量越来越大，如何优化特征选择方法，提高检测速度和检测精度，是入侵检测系统要解决的关键问题。

依据评价标准是否依赖学习算法，特征选择算法可分为过滤和封装 2 种模式[2]。过滤模式独立于

收稿日期：2014-05-28；**修回日期**：2014-06-22

基金项目：广西自然科学基金资助项目（2012GXNSFAA053224）；广西无线宽带通信与信号处理重点实验室 2014 年开放基金资助项目（GXKL0614110）

Foundation Items: The National Natural Science Foundation of Guangxi Province (2012GXNSFAA053224); The Key Laboratory Open Foud Preject of Broadband Wireless Communication and Signal Processing of Guangxi Province in 2014 (GXKL0614110)

学习算法，计算量较低，且能有效去除噪声特征；而封装模式需要预先确定分类器算法，再利用分类器对特征集合进行评价，其分类性能较好，但计算代价大。如何将过滤模式和封装模式结合，进一步提高分类性能是目前特征选择算法的研究重点。针对过滤模式，文献[3]提出基于马氏距离特征排序和穷举搜索的算法，以提高入侵检测系统的分类准确率；文献[4]采用信息增益来过滤噪声和冗余特征，从而实现降维和减轻数据处理的难度；对于封装模式而言，文献[5]通过结合 Kappa 系数和模糊神经网络来选取最优特征集合；文献[6]则提出基于贝叶斯(Bayes)网络分类器的封装算法，以达到改善系统检测性能的目的；文献[7]将过滤模式和封装模式进行结合，在蛋白质无序区域预测和基因选择中筛选特征，提高检测速度和检测精度。

本文提出一种基于支持向量机 (SVM, support vector machine)的两级特征选择方法，该方法将过滤模式和封装模式相结合，通过构造一个基于检测率与误报率比值的特征评测值，以之作为特征选择的评价指标，先采用过滤模式中的 Fisher 分和信息增益分别过滤噪声和无关特征，并进行特征子集的交叉合并；再基于筛选出来的交叉特征子集，采用序列后向搜索算法，结合 SVM 选取最优特征集合。

2 相关工作

2.1 Fisher 分

Fisher 分[8]属于过滤模式，其基本思想是使不同类样本间的距离最大化，同时最小化同类样本间的距离，并以两者的比值作为特征的 Fisher 分值。以 FS_k 表示第 k 维特征的 Fisher 分值，则

$$FS_k=\frac{S_{b,k}}{S_{w,k}}=\frac{(\overline{m_{1,k}}-\overline{m_k})^2+(\overline{m_{2,k}}-\overline{m_k})^2}{\delta_{1,k}^2+\delta_{2,k}^2} \tag{1}$$

其中，$S_{b,k}$ 为类间离散度，表示不同类样本间的距离；$S_{w,k}$ 为类内离散度，表示同类样本间的距离；$\overline{m}_{1,k}$、$\overline{m}_{2,k}$ 和 $\overline{m}_k$ 表示正类、负类和全部样本第 k 维特征的均值；$\delta_{1,k}^2$、$\delta_{2,k}^2$ 则表示正类和负类样本第 k 维特征的方差。特征的 Fisher 分值越大，特征的分类能力则越强，对分类的贡献也越大。

2.2 信息增益

信息增益是一种有效的基于样本信息的过滤特征选择算法，能用来度量某个特征与类别的相关性。特征的信息增益值越大，表示它与类别之间的相关程度越高，其类别区分能力也越强。若将特征 A 的信息增益定义为原来的信息需求与新的需求之间的差[9]，并以 $IG(A)$表示，则

$$\begin{aligned}IG(A)&=Info(D)-Info_A(D)\\&=-\sum_{i=1}^{m}p_i\mathrm{lb}(p_i)-\sum_{j=1}^{v}\frac{|D_j|}{|D|}\times Info(D_j)\end{aligned} \tag{2}$$

其中，$p_i=|D^*|/|D|$ 表示任意样本属于类 C_i 的概率，$|D^*|$ 是属于类 C_i 的样本数，$|D|$ 是总样本数，m 是样本类别数，v 为划分子集 D_j 的个数。

2.3 序列搜索算法

序列搜索算法属于封装模式，通常包含序列后向搜索和序列前向搜索 2 种。

序列后向搜索一般从特征全集出发，每次从当前特征集中剔除一个或若干个对其评价函数贡献最小的特征，使剔除该特征后特征子集的评价指标达到最优。该方法考虑了特征间的依赖性，分类性能较高。

序列前向搜索则是每次选择一个特征或若干个特征加入到当前特征子集，使特征子集的评价指标达到最优，是一种贪心算法。与序列后向搜索相比，其计算量较小，但最终结果依赖初始选择的特征。

2.4 支持向量机

在入侵检测中，特征选择是从网络数据集中筛选出对分类器分类性能影响最重要的最优特征子集。而分类器则是一种机器学习程序，其通过自动学习，可将数据划分到已知的不同类别中，分类器实质是数学模型。针对模型的不同，目前有多种分支，包括 Bayes 分类器、BP 神经网络分类器、决策树、SVM 等。Bayes 给出了最小化误差的最优解决方法，可用于预测和分类；但在使用中，它需要知道证据的确切分布概率，而实际上并不能给出证据的确切分布概率，只能以某种假设去逼近 Bayes 的要求。BP 神经网络具有较好的自适应性、容错性、外推性及较好的模式识别性能等优势；但其训练时间长、需大量的训练数据、不能保证最佳结果；另外，当训练集过小，且采集的数据有误时，会影响其效果。与其他分类算法相比，决策树具有速度快、准确性高的优点，但其深度优先搜索，在处理大训练集数据时缺乏伸缩性；其预测性在有噪声时会受影响；且其还可能产生子树复制和碎片问题。

SVM[10]是以统计学习理论为基础的一种机器学习算法，具有严格的数学理论基础，其用于分类的实质是寻找一个最优分类超平面。SVM 通过将网络数据分组提取的特征映射到高维空间，并在高维空间寻求一个能实现数据分类的最优分类超平面，将低维空间线性不可分的数据转化到高维空间从而实现可分。SVM 可以在有限训练样本下，较好地解决小样本、非线性、高维数和局部极小点等一些实际问题，并使分类器的泛化能力达到最优。基于大量已标记的数据训练出分类模型，SVM 可利用分类模型实施后续的分类检测。本文采用 SVM 作为分类器进行数据分类处理。

在 SVM 将高维空间中的内积运算转换为低维空间时，其需要选择合适的核函数进行计算，避免维数灾难。常用核函数有线性核函数、多项式核函数、RBF 核函数和 Sigmoid 核函数等[11]。本文采用性能最优的 RBF 核函数，其不仅可以将样本映射到一个更高维的空间，而且参数较少，能够降低模型的复杂性。

3　两级特征选择方法

如何选择合适的特征子集以提高入侵检测系统的效率，是人们一直在探究的问题。特征选择中的过滤模式的优点在于计算量较低，能有效去除噪声特征，但其不足在于其筛选出来的特征子集仍然存在一些冗余特征，特征的分类能力不够好。而封装模式筛选出来的特征其分类性能较好，但其计算代价大。本文给出一种基于支持向量机的两级特征选择方法，其将过滤模式与封装模式相结合，先采用过滤模式，以 SVM 建立分类模型进行特征选择，降低特征维度，筛选特征子集；再采用封装模式进一步寻优，有效降低计算开销，提高分类性能。

该方法的基本过程如下：分别采用 Fisher 分和信息增益对原始特征集合进行计算，并根据特征评测值去除无关和噪声特征；将通过 Fisher 分和信息增益计算后筛选出的特征子集进行合并，选出交叉特征子集；再采用序列后向搜索算法，以 SVM 为分类器建立分类模型，以特征评测值作为评价指标，在交叉特征子集中筛选出最优的特征子集。

3.1　相关定义

为形式化描述两级特征选择方法，给出如下声明和相关定义。

给定原始特征向量 $F=\{F_1,F_2,\cdots,F_i,\cdots,F_m\}$，其中，$m$ 为特征维数，F_i 为特征向量中的第 i 维特征，且 $1\leqslant i\leqslant m$。

以 FS_i 表示特征向量 F 中 F_i 的 Fisher 分值，以 *FS_Sub* 表示采用 Fisher 分计算并筛选出来的特征子集；以 IG_i 表示特征向量 F 中 F_i 的信息增益值，以 *IG_Sub* 表示采用信息增益计算并筛选出来的特征子集。

在特征选择中，检测率和误报率是进行特征评价的 2 个重要指标。特征选择的目的是期望筛选出具有较强分类能力的特征，这些特征的检测率应尽可能高，其误报率应尽可能低。以特征的检测率和误报率的比值作为特征评价的准则，可以筛选出合理的特征子集。

定义 1　特征评测值是指某一特征或一组特征的检测率与误报率的比值。以 *F_DS* 表示特征评测值，则

$$F\_DS_i = DR_i / FR_i \tag{3}$$

其中，i 表示第 i 维特征或第 i 组特征；DR_i 表示第 i 维(组)特征的检测率；FR_i 表示第 i 维(组)特征的误报率。若某特征的检测率越高，误报率越低，其特征评测值越大，其分类能力更好。

3.2　特征过滤与特征交叉合并

在特征过滤时，人们常通过设置阈值来过滤掉噪声和无关特征。但阈值的设置多数凭经验，没有统一的标准，且随机性很大。为合理地过滤噪声特征和无关特征，本文以 SVM 建立分类模型，以特征评测值为评价指标，根据测试、计算出的特征或特征子集的特征评测值变化情况过滤特征。

不同的特征选择算法，从算法自身的特性方面可以选择出相适应的最优特征子集。但这些筛选出来的特征子集，在面向具有不同特性的数据集时，因缺乏完备性，其最终的分类结果存在不足。将不同特征选择算法筛选出来的特征子集进一步合并，并在合并的特征子集基础上重新筛选，将能够获取更优的特征子集。

Fisher 分和信息增益均属于过滤模式。Fisher 分从距离的角度评价特征与类别之间的相关性，信息增益则从概率统计的角度度量特征与类别的相关性。本文将 Fisher 分和信息增益 2 种特征选择算法进行组合，分别根据自身特性筛选特征子集，再将这些特征子集进行交叉，进一步筛选特征子集，使最终确定的特征子集具有更好的适应性。

3.2.1　特征过滤

由于分别采用 Fisher 分和信息增益进行特征过

滤的方法相同，下面仅给出针对 Fisher 分的特征过滤方法。

1) 设置一个空的特征集合 Ω。

2) 针对原始特征向量 F，采用式(1)计算各维特征 F_i 的 Fisher 分值 FS_i。

3) 对所有特征计算的 Fisher 分值 FS_i 进行降序排序，形成一个特征序列 *FS_Sort*。

4) 从 *FS_Sort* 中，顺序依次取出一个特征，将其加入到 Ω 中。

5) 针对 Ω 中的特征子集，以 SVM 建立分类模型，测试、计算其特征评测值 F_DS_i；记录并保存 F_DS_i 值及对应 Ω 中的特征子集。

6) 重复步骤 4)和步骤 5)，直到 *FS_Sort* 中的所有特征被全部加入到 Ω 中。

7) 在计算出的所有 F_DS_i 中，搜索最大的 F_DS_i 值。

8) 将最大 F_DS_i 值对应的 Ω 中的特征子集作为筛选过滤出来的特征子集，构成特征子集 *FS_Sub*。

按照同样的方法，针对原始特征向量 F，采用式(2)测试、计算特征的信息增益值 IG_i，并以特征评测值筛选并构成特征子集 *IG_Sub*。

3.2.2 特征交叉合并

特征交叉合并是将采用不同特征选择算法筛选出来的特征子集合并在一起，并将不同特征子集中存在的相同特征选取出来，构成交叉特征子集。交叉特征子集中的特征，具有从不同角度进行特征分类的能力。在面向具有不同特性的数据集时，交叉特征集中的特征将具有更好的适应性。

以 *F_Cross* 表示交叉特征子集，则

$$F\_Cross=FS\_Sub \cap IG\_Sub$$

3.3 特征二次优化选择

交叉特征集已经去除了噪声特征，但其仍然存在一些冗余特征。基于交叉特征集，特征优化选择采用序列后向搜索算法，以 SVM 为分类器建立分类模型，并以特征评测值作为评价指标，进一步执行特征优化选择。其算法如下。

输入：$F\_Cross=\{F_1,F_2,\cdots,F_i,\cdots,F_k\}$，其中，$k$ 为特征个数；

输出：优化特征子集 *F_P*；

```
Begin
{
   利用 F_Cross 建立 SVM 模型，测试其特征评测值，并将其赋予 F_DS_Max;
  F_P = F_Cross;
  While（k>0）
  {
     Max = 0;
     For i=1 to k Do
        {
        F_Cross'= F_Cross – { F_i };    //每次从特征集中去除一维不同特征进行计算；
        利用 F_Cross'建立 SVM 模型，测试其特征评测值 F_DS_i;
        If（F_DS_i>F_DS_Max）          //找最大的特征评测值；
        {   F_DS_Max = F_DS_i ;
            Max = i;      }
     }
     If（Max≠0）
     {
        F_Cross = F_Cross–{F_i};  //获取当前一轮循环中最优的特征子集；
        F_P = F_Cross;
        k--;
     }
     Else
        Break;     //不能筛选出比上一轮循环更优的特征子集则停止搜索；
   }
 }
End
```

3.4 两级特征选择算法

两级特征选择算法通过采用过滤模式中的 Fisher 分和信息增益分别进行特征的初次过滤，去除噪声和无关特征，降低了特征维数；再通过对 2 个筛选出来的特征子集进行合并，挑选出交叉特征子集；为去除交叉特征子集中的冗余特征，采用序列后向搜索算法进行优化选择，获得具有最优性能的特征子集。算法流程如图 1 所示。

算法在特征过滤和特征二次优化选择中，都采用特征评测值作为以 SVM 建立分类模型进行特征筛选的评价指标。

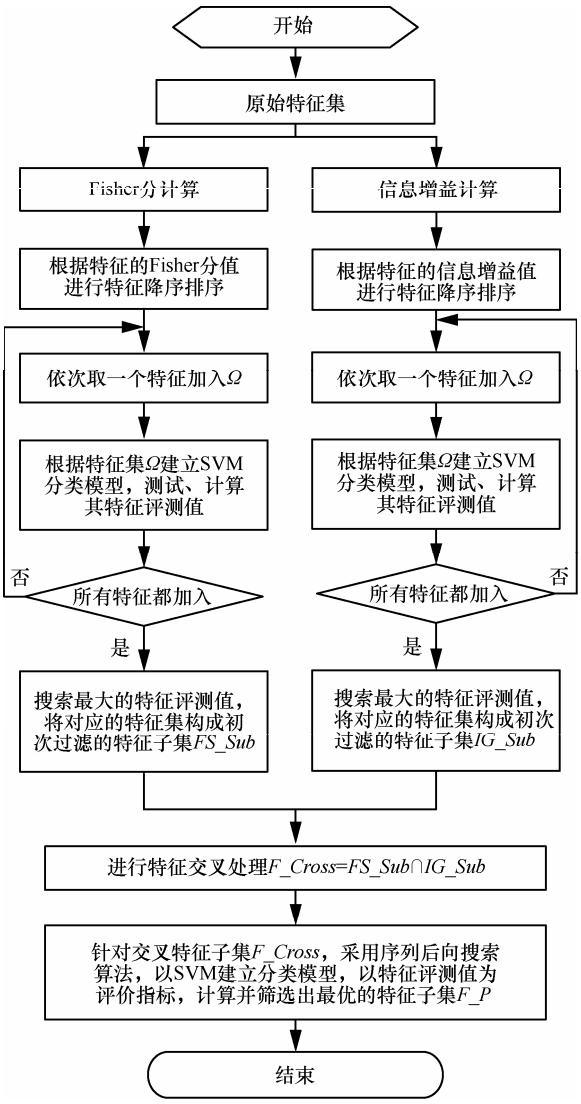


图 1　两级特征选择算法流程

4　仿真实验及结果分析

4.1　实验设置

本文在 Kddcup99[12]数据集给出了类别标号的 10%训练子集上进行仿真实验。实验测试的计算机操作系统为 Windows 7，处理器为 AMD Athlon(tm)X2 DualCore-64 2.10 GHz，内存为 2 GB。实验中，在 Matlab R2011b 环境下实现特征选择算法，并采用 LibSVM（libsvm-mat-3.1）作为训练和测试工具，核函数采用 C-SVM 中的 RBF 核函数，其参数设置为：惩罚系数 c 取 1.2，RBF 函数的核半径 g 取 2.8。

Kddcup99 数据集中的不同特征度量方法不同，数据类型也不同，为便于模型训练，本文对数据集进行预处理，包括对字符类型数据的量化，以及对数据的标准化和归一化处理。字符类型数据的量化主要对协议类型、服务类型和标志位进行数值化，如将协议中的 TCP、UDP、ICMP 分别赋值为 1、2、3 等。

4.2　实验结果与分析

实验 1　特征过滤及特征交叉合并。

特征过滤主要用于去除噪声和无关特征。在测试中，从 10%的 Kddcup99 训练子集中，以正常与异常数据比约为 4:1 的比率随机抽取约 6 000 条记录作为训练集，再随机抽取不相交的 2 300 多条记录作为测试集。针对原始 41 维特征，分别采用 Fisher 分和信息增益计算不同特征的 Fisher 分值和信息增益值，并分别降序排序。基于排序的 Fisher 分特征序列和信息增益特征序列，依次选取特征构成特征子集，以 SVM 建立分类模型，计算其特征评测值，其计算结果如图 2 和图 3 所示。

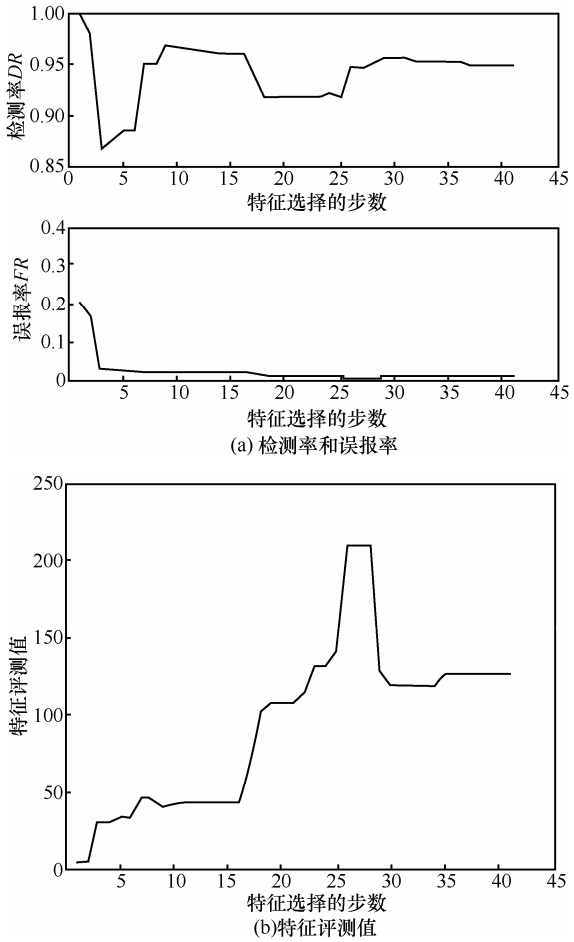


图 2　采用 Fisher 分计算的特征评测值结果

基于图 2 和图 3 的计算结果，选取特征评测值最大的特征子集，则分别采用 Fisher 分和信息增益进行特征过滤筛选出来的特征子集 *FS_Sub* 和 *IG_Sub*，如表 1 所示。

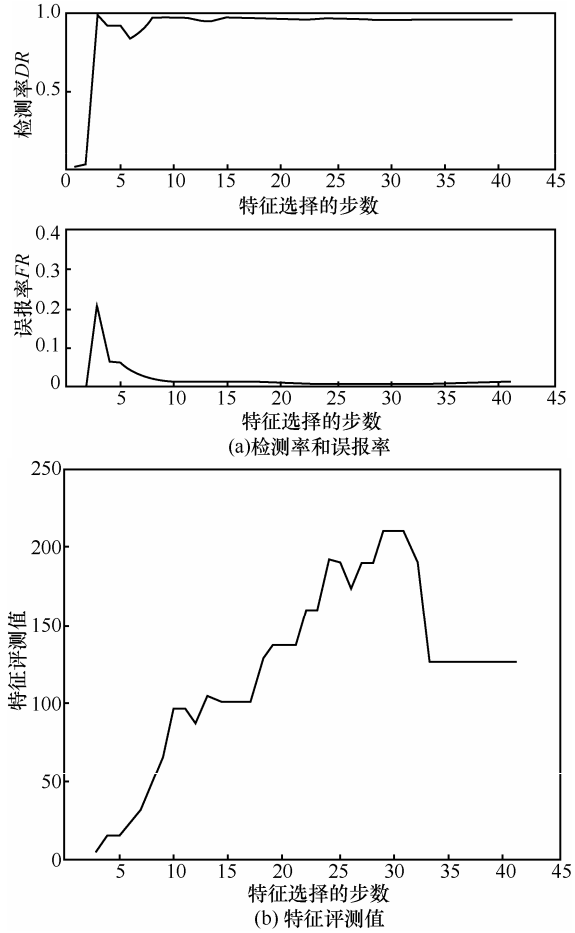


图 3 采用信息增益计算的特征评测值结果

基于表 1 中的特征子集 *FS_Sub* 和 *IG_Sub*，进行特征交叉合并，选取其交集，即 *F_Cross*= *FS_Sub* ∩*IG_Sub*，则特征子集 *F_Cross* 结果如表 2 所示。

实验 2 特征二次优化选择。

特征二次优化选择即采用序列后向搜索对特征子集 *F_Cross* 进一步优化，以 SVM 建立分类模型，依次选取特征构成特征子集，以特征评测值作为评估指标，去除冗余特征，得到优化特征集合。该测试采用同样的方式随机抽取约 9 000 条记录作为训练集，抽取 2 300 多条记录作为测试集。在实验中，采用多组不同的数据进行测试，获取了对应的特征子集，通过对这些特征子集的进一步测试评估，确定并选取最优的特征子集 *F_P*，其包含了 13 维特征，分别为 2、3、4、10、12、22、23、28、32、33、34、37、39。

实验 3 特征对比测试。

针对本文筛选的特征，在同样的实验环境下，与 GFR 方法[13]、F-Score 方法[14]和 IIG 方法[15]对比测试算法的建模时间、测试时间、检测率、误报率和分类准确率，测试结果如表 3 所示。其中，GFR 是一种采用序列后向搜索进行封装的方法；F-Score 为对 Fisher 分算法的改进算法；IIG 为对信息增益的改进算法。在测试中，同样从 10%的 Kddcup99 训练子集中，随机抽取约 26 000 条记录作为训练集，以与训练集不相交的 12 000 多条记录作为测试集。

在上述的测试中，均采用 SVM 分类。而 SVM 分类实际上包括了训练和分类 2 个完全不同的过程。训练时间对应表 3 中的建模时间，分类时间为表 3 中的测试时间。在训练阶段，其采用的核方法中需要用到求解二次规划的问题。根据文献[16]的分析，训练计算复杂度在 $O(N^3+L\times N^2 + d\times L\times N)$和 $O(d\times L^2)$之间，空间复杂度为 $O(L^2)$；而分类的时间复杂度为 $O(d\times N)$；其中，n 是支持向量个数，l 是训练集样本个数，d 为每个样本的原始维数。由此可见，算法的性能与支持向量的个数、训练集样本个数和样本的原始维数有关。同时，所选择的算法不同，其性能也会不同。本文方法先利用过滤算法计算量低的优点，去除噪声特征降低特征维数，降低计算量，再在此基础上进行特征优化选择。表 3 的测试结果显示，在同样的测试环境下，本文方法具有较好的性能。

本文方法以特征评测值作为特征筛选的评价指标，通过二次特征选择提高算法的分类能力。相对于仅采用 Fisher 分或信息增益的特征选择来说，

表 1 特征子集 *FS_Sub* 和 *IG_Sub*

特征集合	特征维数	所含特征
FS_Sub	26	336 33 23 12 24 4 29 34 2 27 40 28 41 39 26 37 10 35 25 38 30 22 5 32 14
IG_Sub	28	5 3 23 24 33 36 2 32 4 12 29 10 34 39 26 37 27 38 25 40 28 22 41 35 30 6 1 11

表 2 特征子集 *F_Cross*

特征集合	特征维数	所含特征
F_Cross	25	2 3 4 5 10 12 22 23 24 25 26 27 28 29 30 32 33 34 35 36 37 38 39 40 41

表 3　　**特征选择算法对比测试结果**

特征选择方法	特征维数	建模时间/s	测试时间/s	检测率/%	误报率/%	分类准确率/%
GFR	19（2 4 8 10 14 15 19 25 27 29 31 32 33 34 35 36 37 38 40）	5.10	2.26	97.83	0.41	99.26
F-Score	19（2 4 6 12 23 24 25 26 29 30 31 32 33 34 35 36 37 38 39）	7.20	2.90	96.60	0.91	98.62
IIG	12（2 3 5 6 8 10 12 23 25 36 37 38）	5.84	2.28	97.36	1.19	98.53
所有特征	原始 41 维特征	7.63	4.89	97.23	0.23	99.29
本文方法	13（2 3 4 10 12 22 23 28 32 33 34 37 39）	4.48	1.73	98.43	0.31	99.45

表 4　　**特征选择算法稳定性对比测试结果**

特征选择方法	特征维数	建模时间平均偏移值	测试时间平均偏移值	检测率平均偏移值	误报率平均偏移值	分类准确率平均偏移值
GFR	19	0.299 2	0.093 6	0.267 2	0.038 4	0.078 4
F-Score	19	0.445 6	0.119 2	0.192 8	0.018 4	0.036 8
IIG	12	0.298 4	0.076	0.264	0.026 4	0.053 6
所有特征	41	0.246 4	0.068 8	0.274 4	0.024 8	0.059 2
本文方法	13	0.210 4	0.046 4	0.255 2	0.031 2	0.048

本文方法的二次特征优化选择，保证了所筛选出来的特征子集具有更优的分类能力；而相对于仅采用序列后向搜索的特征选择来说，本文方法的初次特征过滤，有效去除了噪声特征，避免了噪声特征在后续特征选择中的影响。表 3 的测试结果显示，在同样的测试环境下，相对于其他算法，本文所述的方法具有较好的分类能力，其检测率、误报率和分类准确率都有较好的表现。

实验 4　特征分类稳定性测试。

在同样的测试环境下，采用同样的方法选取不相交的 5 组训练集和测试集，进一步测试本文方法与 GFR 方法、F-Score 方法、IIG 方法和全部特征集在不同数据集下的特征分类稳定性。稳定性是指特征子集的分类测试结果不随测试数据的变化而剧烈变化。在此。本文将稳定性由实际测试值与平均值之间的偏移值描述，即实际测试值与平均值之间差的绝对值。测试过程如下。

1) 测试上述 5 种方法筛选出来的特征子集分别在 5 组不同数据集上的建模时间、测试时间、检测率、误报率和分类准确率。

2) 针对 5 个测试指标，计算不同方法分别在 5 组不同数据集上测试的平均值。

3) 计算出不同方法在 5 组不同数据集上，针对每个测试指标的偏移值。

4) 针对 5 组不同数据集，计算不同方法在每个测试指标的平均偏移值。平均偏移值越小，表明特征子集在面向不同的数据集时分类能力越稳定，其分类效果越好。测试结果如表 4 所示。

本文方法将过滤模式和封装模式相结合实现特征的优化选择，其筛选出来的特征子集具备不同特征选择算法的特性，在面向不同数据集时适应性更好。表 4 的测试结果也表明本文方法具有较好的稳定性。

5　结束语

针对入侵检测中的特征选择问题，本文提出一种基于 SVM 的两级特征选择方法。该方法将过滤模式与封装模式相结合，以 SVM 建立分类模型，根据特征评测值对分别采用 Fisher 分和信息增益计算的特征进行特征过滤，并选取两者的交叉特征子集；再采用序列后向搜索算法，以特征评测值作为评价指标，基于 SVM 建立的分类模型在交叉特征子集中筛选出最优的特征子集。仿真测试结果表明，本文方法具有较高的分类检测率和较低的误报率，以及良好的分类性能，其能够有效提高入侵检测系统的安全性。在将来的工作中，将进一步扩展和完善特征评测值的计算方法，并针对 SVM 在训练样本数量增加情况下的系统开销问题，寻求有效的解决途径，获得更好的分类检测性能。

参考文献：

[1] ZHANG Y, YANG A, XIONG C, *et al*. Feature selection using data envelopment analysis[J]. Knowledge-Based Systems, 2014, 64:70-80.

[2] LEE M C. Using support vector machine with a hybrid feature selection method to the stock trend prediction[J]. Expert Systems with Applications, 2009, 36(8): 10896-10904.

[3] YONGLI Z, YUNG Z, WEI M T, *et al*. An improved feature selection

algorithm based on MAHALANOBIS distance for network intrusion detection[A]. Sensor Network Security Technology and Privacy Communication System (SNS & PCS), 2013 International Conference on[C]. 2013.69-73.

[4] TESFAHUN A, BHASKARI D L. Intrusion detection using random forests classifier with SMOTE and feature reduction[A]. Cloud & Ubiquitous Computing & Emerging Technologies (CUBE), 2013 International Conference on[C]. 2013.127-132.

[5] ARAUJO N V S, OLIVEIRA R, FERREIRA E W T, *et al*. Kappa-fuzzy aRTMAP: a feature selection based methodology to intrusion detection in computer networks[A]. Trust, Security and Privacy in Computing and Communications (TrustCom), 2013 12th IEEE International Conference on[C]. 2013.271-276.

[6] ZHANG F, WANG D. An effective feature selection approach for network intrusion detection[A]. Networking, Architecture and Storage (NAS), 2013 IEEE Eighth International Conference on[C]. 2013. 307-311.

[7] HSU H H, HSIEH C W, LU M D. Hybrid feature selection by combining filters and wrappers[J]. Expert Systems with Applications, 2011, 38(7):8144-8150.

[8] XIE J, WANG C. Using support vector machines with a novel hybrid feature selection method for diagnosis of erythemato-squamous diseases[J]. Expert Systems with Applications, 2011, 38(5):5809-5815.

[9] 汪廷华, 田盛丰, 黄厚宽. 特征加权支持向量机[J]. 电子与信息学报, 2009,31(3):514-518.
WANG T H,TIAN S F,HUANG H K. Feature weighted support vector machine[J]. Journal of Electronics & Information Technology,2009, 31(3): 514-518.

[10] CORTES, VAPNIK V. Support vector networks[J]. Machine Learning, 1995,20:273-297.

[11] 汪廷华, 陈峻婷. 核函数的选择研究综述[J]. 计算机工程与设计,2012,33(3):1181-1186.
WANG T H,CHEN J T.Survey of research on kernel selection[J]. Computer Engineering and Design, 2012,33(3):1181-1186.

[12] KDDCup99KDDdataset[EB/OL]. http://kdd.ics.uci.edu/ databases/ kddcup99/kddcup99.html, 2011.

[13] LI Y, XIA J, ZHANG S, *et al*. An efficient intrusion detection system based on support vector machines and gradually feature removal method[J]. Expert Systems with Applications, 2012, 39(1):424-430.

[14] XUE.Q Z, CHUN.H G, JIA J L. Intrusion detection system based on feature selection and support vector machine[A]. Communications and Networking in China, ChinaCom'06, First International Conference on[C]. 2006.1-5.

[15] XIAN J, PEI Y L, WEI G, *et al*. An algorithm application in intrusion forensics based on improved information gain[A]. Web Society (SWS), 2011 3rd Symposium on[C]. 2011.100-104.

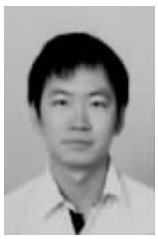
[16] BURGES, C J C. A tutorial on support vector machines for pattern recognition[J]. Data Mining and Knowledge Discovery, 1998, 2(2): 121-167.

作者简介：

武小年（1972-），男，湖北监利人，桂林电子科技大学副教授，主要研究方向为信息安全、分布式计算。

彭小金（1988-），男，江西新余人，桂林电子科技大学硕士生，主要研究方向为信息安全。

杨宇洋（1989-），男，广西柳州人，桂林电子科技大学硕士生，主要研究方向为信息安全。



方堃（1990-），男，湖北武汉人，桂林电子科技大学硕士生，主要研究方向为信息安全。