·安全技术· 文章编号：1000-3428(2019)06-0175-06 文献标志码：A 中图分类号：TP309

基于信息增益与主成分分析的网络入侵检测方法

王旭仁[1,2]，马慧珍[1]，冯安然[1]，许祎娜[1]

（1. 首都师范大学 信息工程学院，北京 100048；
2. 中国科学院信息工程研究所，中国科学院网络测评技术重点实验室，北京 100093）

摘 要： 为提高入侵检测效率，需要对数据进行特征提取以降低数据维度。结合信息增益（IG）和主成分分析（PCA），提出一种网络入侵检测方法。通过 IG 提取分类能力强的属性特征，利用 PCA 对其降维，并采用 Naive Bayes 进行分类检测。对数据集 KDDCUP99 进行测试，结果表明，该方法的检测率为 94.5%，高于 PCA-LDA、FPCA、KPCA 方法。

关键词： 信息增益；主成分分析；入侵检测；特征提取；降维

中文引用格式： 王旭仁，马慧珍，冯安然，等. 基于信息增益与主成分分析的网络入侵检测方法[J]. 计算机工程，2019，45(6)：175-180.

英文引用格式： WANG Xuren，MA Huizhen，FENG Anran，et al. Network intrusion detection method based on information gain and principal components analysis[J]. Computer Engineering，2019，45(6)：175-180.

Network Intrusion Detection Method Based on Information Gain and Principal Components Analysis

WANG Xuren[1,2]，MA Huizhen[1]，FENG Anran[1]，XU Yi'na[1]

（1. College of Information Engineering，Capital Normal University，Beijing 100048，China；
2. Key Laboratory of Network Assessment Technology，Institute of Information Engineering，
Chinese Academy of Sciences，Beijing 100093，China）

【Abstract】 In order to improve the efficiency of intrusion detection，it is necessary to extract the features of data to reduce the data dimensions. This paper proposes a network intrusion detection method by combining Information Gain（IG）and Principal Components Analysis（PCA）. The method uses IG to extract the attribute features with strong classification ability，uses PCA to reduce the dimension of the feature data，and uses Naive Bayes method for classification and detection. The test results of the data set KDDCUP99 show that the detection rate of the method is 94.5%，which is much higher than those of PCA-LDA，FPCA，and KPCA methods.

【Key words】 Information Gain（IG）；Principal Components Analysis（PCA）；intrusion detection；feature extraction；dimension reduction

DOI： 10.19678/j.issn.1000-3428.0050585

0 概述

网络入侵检测是保障网络安全必不可少的一项技术，其通过识别网络中的访问数据是入侵行为还是正常行为来保护网络，本质是一个二分类问题。入侵检测主要分为误用检测和异常检测。误用检测通过与已知的攻击进行对比来识别异常或潜在的行为，对已知的攻击具有较好的检测率，但是检测不到新的或不常见的入侵。异常检测[1-2]根据网络中正常的数据链接或正常的访问行为建立“normal”模型，当数据链接或访问行为不在此前建立的模型范围内，则生成警告。机器学习[3-4]和统计方法[5-6]在网络入侵方面的应用提高了入侵检测率，但是网络数据的冗余属性和不相关属性依然影响入侵检测的效率。提取和选择数据中主要信息或相关性高的信息对提高入侵检测的效率非常重要。

信息增益（Information Gain，IG）[7]作为一种特征选择算法，被用于决策树生成过程的根节点选择。IG 表示得知特征 X 的信息而使得类 Y 的信息的不确定性减少的程度。不同的特征具有不同的 IG，IG 大的特征具有更强的分类能力。文献[8]使用 IG 对 KDDCUP99 数据进行特征的相关性分析，对每个属性

基金项目： 国家自然科学基金（61373161）；中国科学院信息工程研究所中国科学院网络测评技术重点实验室开放课题（201710）。
作者简介： 王旭仁（1972—），女，副教授、博士，主研方向为网络安全、智能技术；马慧珍、冯安然、许祎娜，硕士研究生。
收稿日期： 2018-03-02 **修回日期：** 2018-05-11 **E-mail：** wang-xuren@163.com

计算 IG,并将其大小作为衡量属性对类标签的影响。

主成分分析(Principal Components Analysis,PCA)将对协方差矩阵影响最大的方向上的数据值作为主成分,用这些数据值构成新子空间表示原始数据,从而达到降维的目的。PCA 作为特征提取方法之一被广泛应用于模式识别、图像处理等领域,同样在网络入侵检测中得到了有效的应用[9-11]。近期研究人员将主成分分析与线性判别分析组合(Principal Components Analysis and Linear Discriminant Analysis,PCA-LDA)[12]、模糊主成分分析(Fuzzy Principal Components Analysis,FPCA)[13]、核主成分分析(Kernel Principal Components Analysis,KPCA)[14]应用于网络入侵检测,提高了其检测率。

本文提出一种 IG 与 PCA 相结合的特征提取方法(简称 IG_PCA 方法)。IG 的特征选择可以找到对分类能力强的属性,从而删除冗余的属性,而 PCA 利用协方差矩阵来提取主成分,在特征提取方面具有一定的优势。两者结合可以最大程度地去除冗余信息并提取主要组成信息,充分利用数据中的有用信息,降低数据的维度。

1 属性相关性分析

1.1 信息增益

设 S 为具有类标签的训练样本集。假设有 c 个类,第 i 个类包含 s_i 个训练样本,s 为总的训练样本个数。训练样本集 S 的经验熵计算如下:

$$H(s_1,s_2,\cdots,s_c)=-\sum_{i=1}^{c}\frac{s_i}{s}\text{lb}\frac{s_i}{s} \tag{1}$$

特征 f 有 v 个不同的取值$\{\rho_1,\rho_2,\cdots,\rho_v\}$,根据特征 f 的取值将样本集 S 分为 v 个子集$\{S_1,S_2,\cdots,S_v\}$,S_j是特征 f 取值为 ρ_j的样本集合,其中 s_{ij}是 S_j 第 i 类的样本个数,特征 f 的经验条件熵计算如下:

$$E(f)=\sum_{j=1}^{v}\frac{s_{1j}+s_{2j}+\cdots+s_{cj}}{s}\times H(s_{1j},s_{2j},\cdots,s_{cj}) \tag{2}$$

则特征 f 的 IG 为:

$$Gain(f)=H(s_1,s_2,\cdots,s_c)-E(f) \tag{3}$$

1.2 主成分分析

PCA 被广泛地应用于许多领域的数据分析和预处理,其实际上是一种降维方式。它的主要目的是减少初始变量的维数,同时尽可能地保留这些样本中存在的方差,将许多相关性高的变量转化成彼此相互独立或不相关的变量。

设一个数据集有 m 个对象 $x_1,x_2,\cdots,x_m$,每个对象包含 n 个变量。为了获得 $n'(n'<<n)$个主成分,PCA 过程如下:

步骤 1 原始数据标准化:

$$Z_{ij}=\frac{x_{ij}-\bar{x}_j}{s_j},i=1,2,\cdots,m,j=1,2,\cdots,n \tag{4}$$

其中,$\bar{x}_j=\frac{\sum_{i=1}^{m}x_{ij}}{m}$,$s_j^2=\frac{\sum_{i=1}^{m}(x_{ij}-\bar{x}_j)^2}{m-1}$。

步骤 2 求相关系数矩阵:

$$\boldsymbol{R}=\frac{\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}}{m-1} \tag{5}$$

步骤 3 求 $\boldsymbol{R}$ 的特征根 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_n$ 及相应的单位特征向量:

$$\boldsymbol{a}_i=\begin{bmatrix}\alpha_{1i}\\ \alpha_{2i}\\ \vdots\\ \alpha_{ni}\end{bmatrix},i=1,2,\cdots,n \tag{6}$$

步骤 4 计算主成分:

$$t_i=\alpha_{1i}Z_1+\alpha_{2i}Z_2+\cdots+\alpha_{ni}Z_n,\ i=1,2,\cdots,n' \tag{7}$$

以 n'个主成分作为新的数据向量,取代原始数据。

经过属性特征提取后,可以最大程度地去除不重要或冗余的属性特征。

1.3 IG 与 PCA 的组合

本文提出组合的降维方法。利用 IG 和 PCA 各自的优点最大限度地降低数据维度,同时提高检测率。

定义 1 $F_{\text{init}}=\{f_1,f_2,\cdots,f_n\}$,原始数据的属性集合 F_{init}包含 n 个属性。

定义 2 $F_{\text{IG}}=\{F_{\text{IG1}},F_{\text{IG2}},\cdots,F_{\text{IG}i}\}$,$i=1,2,\cdots,k$,其中,$F_{\text{IG}i}$表示第 i 个属性子集,$|F_{\text{IG}i}|=i$,$F_{\text{IG}(i-1)}\subseteq F_{\text{IG}i}$。$k$ 为属性子集的数目,$k<n$。例如,$F_{\text{IG1}}=\{f_i\}$,f_i为 IG 值最大的属性,i 表示 F_{init}的第 i 个属性。

定义 3 $F_{\text{PCA}}=\{F_{\text{PCA1}},F_{\text{PCA2}},\cdots,F_{\text{PCA}i}\}$,$i=1,2,\cdots,k$,$k<n$。其中,$F_{\text{PCA}i}$表示第 i 个属性子集,$|F_{\text{PCA}i}|=i$,$F_{\text{PCA}(i-1)}\subseteq F_{\text{PCA}i}$。例如,$F_{\text{PCA1}}=\{t_1\}$,$t_1$ 表示最大特征值对应的主成分。

定义 4 $F_{\text{IG_PCA}}=\{F_{\text{IG1_PCA}},F_{\text{IG2_PCA}},\cdots,F_{\text{IG}i_\text{PCA}}\}$,$i=1,2,\cdots,k$,其中,$F_{\text{IG}i_\text{PCA}}$表示由 $F_{\text{IG}i}$属性子集生成得集合,$F_{\text{IG}i_\text{PCA}}=\{F_{\text{IG}i_\text{PCA1}},F_{\text{IG}i_\text{PCA2}},\cdots,F_{\text{IG}i_\text{PCA}l}\}$,$l=1,2,\cdots,L$,其中,$F_{\text{IG}i_\text{PCA}l}$表示第 l 个属性子集,且 $|F_{\text{IG}i_\text{PCA}l}|=l$,$F_{\text{IG}i_\text{PCA}(l-1)}\subseteq F_{\text{IG}i_\text{PCA}l}$。$k$、$l$ 为属性子集的数目,$k<n$,$L\leqslant k$。例如,$F_{\text{IG}i_\text{PCA1}}=\{t_{\text{IG}i_\text{PCA1}}\}$,$t_{\text{IG}i_\text{PCA1}}$表示最大特征值对应的主成分。

本文方法的具体步骤如下:

1)生成属性集 $F_{\text{IG}}=\{F_{\text{IG1}},F_{\text{IG2}},\cdots,F_{\text{IG}i}\}$,$i=1,2,\cdots,k$。

2)对 F_{IG}进行字符属性数值化。

3)对步骤 2 中的每个子集进行 PCA。

4)生成属性集 $F_{\text{IG_PCA}}=\{F_{\text{IG1_PCA}},F_{\text{IG2_PCA}},\cdots,F_{\text{IG}i_\text{PCA}}\}$,$i=1,2,\cdots,k$。

5)对步骤 4 得到的属性子集进行评估,选取最优子集 $F_{\text{IG}i_\text{PCA}l}$。

2　入侵检测

2.1　数据集及评估方法

2.1.1　数据集

本文使用 KDDCUP99 数据集[15]。该数据集由 1998 年 DARPA 入侵检测数据经整理获得,因 10% 的训练数据与 10% 的测试数据被用于 1999 年 KDD CUP 竞赛中,成为著名的 KDDCUP99 数据集。KDDCUP99 数据集包括训练与测试数据集,其包含的 39 种攻击被分为以下 4 种主要类型:

1)拒绝服务攻击(DOS):攻击者设法让目标机器停止提供服务,例如泛洪攻击等。

2)远程攻击(R2L):从远端机器上的非授权读取,例如基于字典的口令猜测。

3)本地用户非法提升权限的攻击(U2R):非授权地获取管理员权限,例如各种"缓冲区溢出"攻击。

4)网络扫描(Probing):监视和其他探测技术,包括端口扫描和漏洞扫描。

KDDCUP99 数据集的每条链接由 41 个属性(记为 $f_i, i=1,2,\cdots,41$)和 1 个类标签组成,其特征可以分为 4 个主要类型:

1)TCP 连接基本特征($f_1 \sim f_9$,其中包括 3 个字符属性):包含了一些连接的基本属性,如连续时间、协议类型、传送的字节数等。

2)TCP 连接的内容特征($f_{10} \sim f_{22}$):用于评估原始 TCP 数据包的有效负载,如登录失败的次数等。

3)基于时间的网络流量统计特征($f_{23} \sim f_{31}$):2 s 时间单元的流量特征。

4)基于主机的网络流量统计特征($f_{32} \sim f_{41}$):按照目标主机进行分类,使用一个具有 100 个连接的时间窗,统计当前连接之前 100 个连接与当前连接具有相同目标主机的统计信息。

本文实验使用的训练集包含 494 021 个链接,其中有 97 278 个正常链接,22 种攻击类型。测试集包含 311 029 个链接,正常链接有 60 593 个,其中14 种在训练数据集中不存在的新的攻击类型。

2.1.2　字符属性数值化

KDDCUP99 数据集有 41 个属性,其中包括 3 个字符属性,然而字符属性在某些分类算法(如 K-means、KNN 等)中无法使用,在其他算法中也可能会影响分类效果。为保证数据的属性为标准格式属性,将字符属性数值化[9]。假设一个字符属性包含 m 个值,将这 m 个值与 m 个坐标一一对应,对应的属性值记为 1,否则为 0。例如,KDDCUP99 数据集的"protocol type"的属性值包括 tcp、udp、icmp,如果一条链接的数据的 "protocol type"的值为 tcp(udp 或 icmp),可转换为(1,0,0)((0,1,0) 或(0,0,1))。由于"protocol type"包含 3 个不同的字符属性值,"service"包含 70 个不同的字符属性值,"flag"包含 11 个不同的字符属性值,所以 KDDCUP99 数据集由原来的 41 属性变为 122 个属性。

文献[16]提出了另一种字符属性数值化方法:协议类型"protocol type"可数值化为 icmp-0、tcp-1、udp-2、others-3;"flag"可数值化为 rej-1、rsto-2、rstr-3、s0-4、s3-5、sh-6、sf-7、others-8;由于"service"的属性值过多,为避免数据维度过大,该属性忽略不计。

由于不能确定被忽略"service"属性对分类精度的影响,因此本文采取文献[9]方法对数据进行字符属性数值化。

2.1.3　评估方法

本文采用检测率(Detection Rate,DR)、误检率(False Positive Rate, FPR)、精度(Precision)和 *F-Score* 作为评估参数衡量分类效果。一个性能好的入侵检测具有高的检测率、精度和 *F-Score* 以及低的误检率。

$$DR = \frac{TP}{TP + FN} \tag{8}$$

$$FPR = \frac{FP}{FP + TN} \tag{9}$$

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$F\text{-}Score = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{11}$$

其中,TP 为将攻击类链接检测为攻击类链接数量,FN 为将攻击类链接检测为正常类链接数量,FP 为将正常类链接检测为攻击类链接数量,TN 为将正常类链接检测为正常类链接数量。

2.2　实验结果

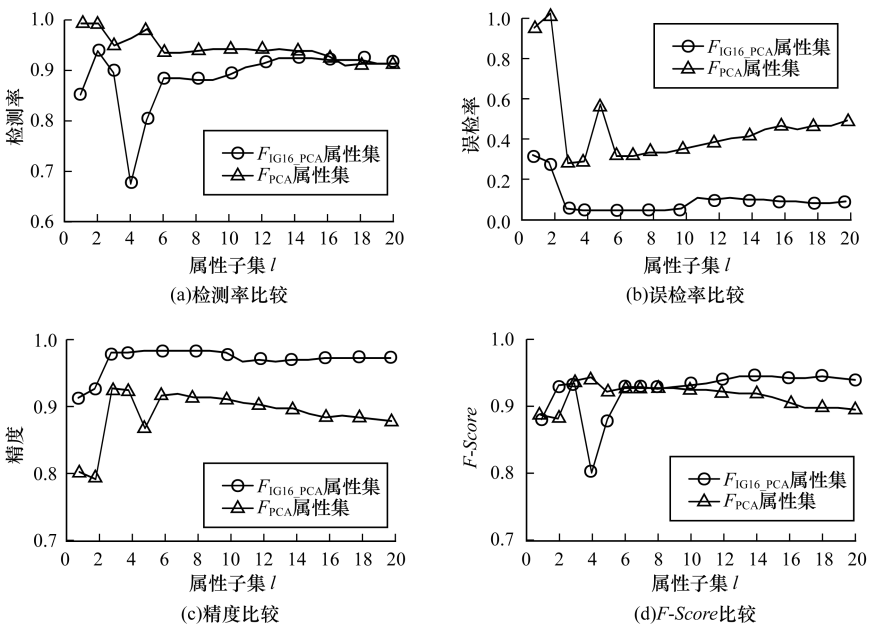
根据式(3)得到训练数据集 41 个属性的 IG 值,如表 1 所示。各个属性根据 IG 值的大小降序排列生成属性集 $F_{IG} = \{F_{IG1}, F_{IG2}, \cdots, F_{IGk}\}$,$k$ 为属性 IG 值大于等于 0.1 的属性数目。根据表 1 可知,$k = 25$,$F_{IG1} = \{f_5\}$,$F_{IG2} = \{f_5, f_{23}\}$,以此类推。

表 1　各属性的 IG 值

IG 值	属性	IG 值	属性	IG 值	属性
1.438 401 1	f_5	0.544 934 80	f_{25}	0.026 930 30	f_8
1.390 205 0	f_{23}	0.543 882 80	f_{39}	0.006 167 50	f_{22}
1.355 236 7	f_3	0.521 428 50	f_{26}	0.00 291 890	f_{16}
1.097 717 5	f_{24}	0.436 411 60	f_{12}	0.002 276 80	f_{19}
1.096 183 0	f_{36}	0.34 709 250	f_{32}	0.001 670 00	f_{17}
1.015 926 5	f_2	0.306 014 20	f_{37}	0.001 543 50	f_{11}
0.905 425 0	f_{33}	0.183 230 20	f_{31}	0.000 892 20	f_{14}
0.899 787 3	f_{35}	0.172 274 90	f_{40}	0.000 671 50	f_7
0.871 382 3	f_{34}	0.161 985 00	f_{41}	0.000 419 50	f_{18}
0.784 463 8	f_{30}	0.122 383 40	f_{27}	0.000 096 10	f_9
0.777 368 6	f_{29}	0.111 119 20	f_{28}	0.000 079 30	f_{15}
0.764 567 9	f_4	0.063 142 10	f_1	0.000 000 00	f_{21}
0.582 020 3	f_6	0.050 266 00	f_{10}	0.000 000 00	f_{20}
0.566 283 6	f_{38}	0.039 662 00	f_{13}		

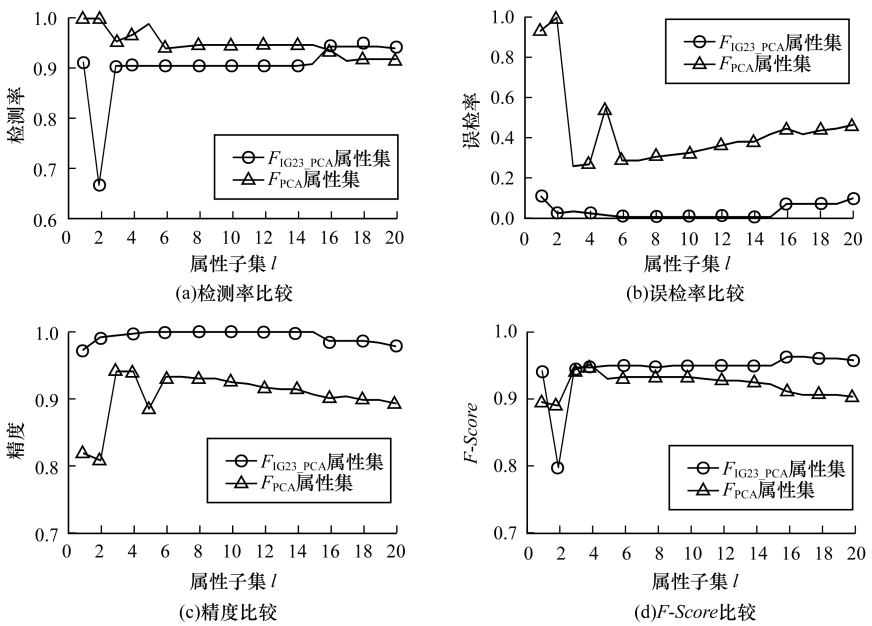
对 F_{IG} 属性集中的属性子集 F_{IGi} 进行属性数值化与 PCA 生成属性集 $F_{IG_PCA}=\{F_{IG1_PCA},F_{IG2_PCA},\cdots,F_{IG25_PCA}\}$。例如：$F_{IG25_PCA}=\{F_{IG25_PCA1},F_{IG25_PCA2},\cdots,F_{IG25_PCAl}\}$，$F_{IG25_PCA1}=\{t_{IG25_PCA1}\}$，$F_{IG25_PCA2}=\{t_{IG25_PCA1},t_{IG25_PCA2}\}$ 等，t_{IG25_PCA1}、t_{IG25_PCA2} 分别为 1.2 节中 λ_1、λ_2 对应的主成分。

本文采用 Naive Bayes 分类器进行分类检测，从中选取最优子集。同时对 F_{init} 进行字符属性数值化与 PCA 生成 F_{PCA}。图 1～图 3 为 IG_PCA 算法与 PCA 算法生成的属性子集在 Naive Bayes 分类器下的评估结果。经过实验发现，由 F_{IG16}、F_{IG23}、F_{IG25} 生成的属性集 F_{IG16_PCA}、F_{IG23_PCA}、F_{IG25_PCA} 的检测效率较好，且当 $l\geqslant 17$ 时，其检测率、精度、*F-Score* 的值呈下降趋势，而误检率的值呈上升趋势。因此，本文实验结果仅展示 $l\leqslant 20$ 的数据。从图 1～图 3 可以看到，虽然 F_{PCA} 的检测率高于 F_{IG16_PCA}、F_{IG23_PCA}、F_{IG25_PCA}，但其误检率也明显高于 F_{IG16_PCA}、F_{IG23_PCA}、F_{IG25_PCA}，且 F_{IG16_PCA}、F_{IG23_PCA}、F_{IG25_PCA} 的精度、*F-Score* 值均优于 F_{PCA} 的，所以 IG 和 PCA 组合的特征提取方法的总体检测率要优于 PCA。而 PCA 的最优属性子集为 F_{PCA4}，其 $DR=0.969$、$FPR=0.266$、$Precision=0.938$、$F\text{-}Score=0.953$。



(a)检测率比较 (b)误检率比较 (c)精度比较 (d)*F-Score*比较

图 1 F_{IG16_PCA} 与 F_{PCA} 的检测率、误检率、精度、*F-Score* 对比



(a)检测率比较 (b)误检率比较 (c)精度比较 (d)*F-Score*比较

图 2 F_{IG23_PCA} 与 F_{PCA} 的检测率、误检率、精度、*F-Score* 对比

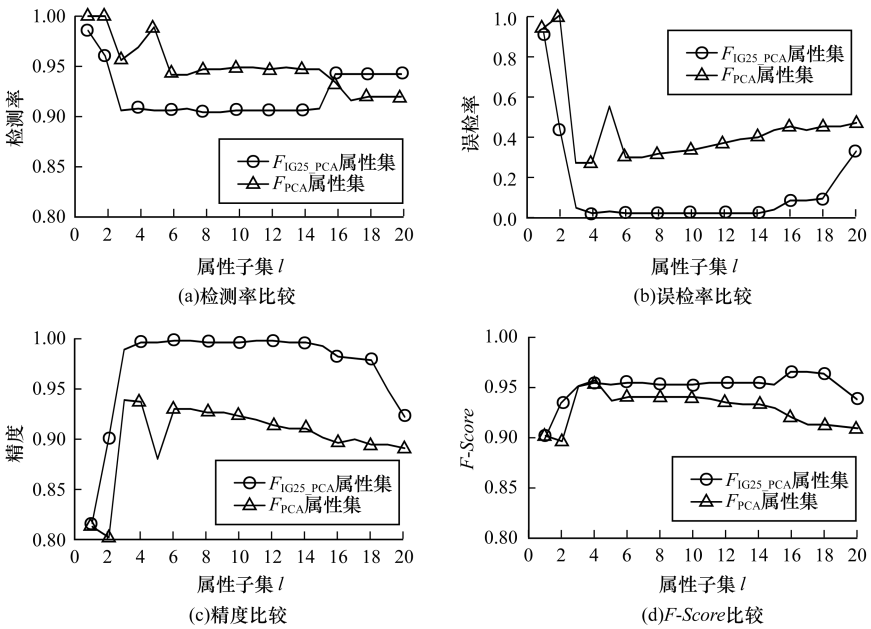


图 3　F_{IG25_PCA} 与 F_{PCA} 的检测率、误检率、精度、*F-Score* 对比

图 4 为 IG_PCA 算法生成的属性集 F_{IG16_PCA}、F_{IG23_PCA}、F_{IG25_PCA} 的检测率、误检率、精度、*F-Score* 对比。从图 4 可以看出,三者的精度、*F-Score* 的值没有太大的差距。当 $l=16$ 时,F_{IG16_PCA16} 的 $DR=0.929$、$FPR=0.06$、$Precision=0.984$、$F\text{-}Score=0.956$,F_{IG23_PCA16} 的 $DR=0.945$、$FPR=0.07$、$Precision=0.982$、$F\text{-}score=0.963$,F_{IG25_PCA16} 的 $DR=0.942$、$FPR=0.069$、$Precision=0.982$、$F\text{-}Score=0.962$。显然,F_{IG23_PCA16} 的检测率最高,比 F_{IG16_PCA16} 与 F_{IG25_PCA16} 分别高 0.016、0.003,而 F_{IG23_PCA16} 的误检率比 F_{IG16_PCA16} 与 F_{IG25_PCA16} 分别高 0.001、0.000 1。F_{IG23_PCA16} 的分类检测效率更好,则 F_{IG23_PCA16} 为生成属性集中的最优属性子集,维度为 16。F_{init} 在 Naive Bayes 分类检测算法下的 $DR=0.915$、$FPR=0.056$、$Precision=0.985$、$F\text{-}Score=0.949$,比 F_{IG23_PCA16} 的检测率低 0.03、误检率低 0.014、精度高 0.003、*F-Score* 低 0.014,总体来看 F_{IG23_PCA16} 的检测效率高于 F_{init}。

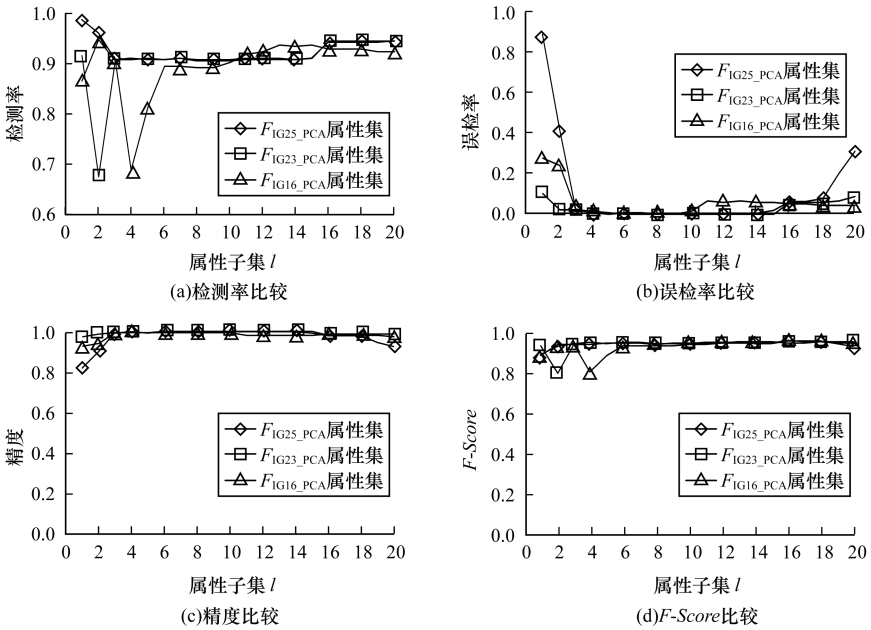


图 4　F_{IG16_PCA}、F_{IG23_PCA}、F_{IG25_PCA} 的检测率、误检率、精度、*F-Score* 对比

经上述分析可知,采用 IG_PCA 方法将数据维度降为 16 时检测率最高,在提高检测准确率的同时也极大地降低了数据维度。本文方法与其他方法的对比结果如表 2、表 3 所示。由表 2 可知,虽然 IG_PCA 的误检率高于 PCA-LDA、FPCA、KPCA,但是比其他 3 种方法的检测率高出 0.3,所以本文方法

的检测率更好。由表 3 可知,PCA-LDA 的 DOS、U2R、Probing 的检测率比 IG_PCA 高 0.016 4、0.142 9、0.051 3,FPCA 的 U2R、R2L、Probing 的检测率比 IG_PCA 的高 0.134 6、0.007 6、0.073 4,KPCA 的 DOS、U2R、R2L 的检测率比 IG_PCA 的高 0.008 8、0.099 3、0.001 1,而 IG_PCA 的 DOS 的检测率比 FPCA 高 0.221 4,比 PCA-LDA 的 R2L 的检测率高 0.005 5,比 KPCA 的 Probing 的检测率高 0.107 3,但是 IG_PCA 的 U2R 的检测率为 0。总体来说,虽然各攻击类在 PCA-LDA、FPCA、KPCA 的检测率优于 IG_PCA,但是由表 2 可知 IG_PCA 的异常检测率远远高于其他 3 种方法。

表 2 各种方法性能对比

方法	检测率	误检率	精度	F-Score
PCA-LDA	0.587	0.019	0.986	0.736
FPCA	0.499	0.044	0.919	0.642
KPCA	0.635	0.015	0.990	0.774
IG_PCA	0.945	0.070	0.982	0.963

表 3 各种方法攻击检测率对比

方法	DOS	U2R	R2L	Probing
PCA-LDA	0.967 7	0.142 9	0.028 0	0.891 2
FPCA	0.731 1	0.134 6	0.041 1	0.913 3
KPCA	0.961 3	0.099 3	0.034 6	0.732 6
IG_PCA	0.952 5	0.000 0	0.033 5	0.839 9

3 结束语

提高入侵检测率一直是安全类软件的目标,而网络数据中的冗余信息严重影响检测的准确性。本文提出 IG_PCA 方法,结合 IG 特征选择和 PCA 特征提取的优点对数据属性进行提取。采用 Naive Bayes 对 KDDCUP99 数据集进行分类检测,结果表明,本文方法将数据维度降为 16 时检测率比 PCA-LDA、FPCA、KPCA 方法高出 0.3。但是该方法没有提高对各攻击类的检测率,下一步将对此进行改进。

参考文献

[1] DENNING D E. An intrusion-detection model[J]. IEEE Transactions on Software Engineering, 1987, 13(2): 222-232.

[2] ANDERSON J P. Computer security threat monitoring and surveillance[R]. Ford, USA: James P. Anderson Company, 1980.

[3] WANG Ke, STOLFO S J. Anomalous payload-based network intrusion detection [C]//Proceedings of International Workshop on Recent Advances in Intrusion Detection. Berlin, Germany: Springer, 2004: 203-222.

[4] GADDAM S R, PHOHA V V, BALAGANI K S. K-means + ID3: a novel method for supervised anomaly detection by cascading k-means clustering and ID3 decision tree learning methods[J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(3): 345-354.

[5] PANDA M, PATRA M R. Network intrusion detection using Naive Bayes[J]. International Journal of Computer Science and Network Security, 2007, 7(12): 258-263.

[6] AMOR N B, BENFERHAT S, ELOUEDI Z. Naive Bayes vs decision trees in intrusion detection systems [J]. ACM Symposium on Applied Computing, 2004, 18(6): 420-424.

[7] QUINLAN J R. Induction of decision trees[J]. Machine Learning, 1986, 1(1): 81-106.

[8] KAYACIK H G, ZINCIR-HEYWOOD A N, HEYWOOD M I. Selecting features for intrusion detection: a feature relevance analysis on KDD 99 intrusion detection datasets [C]//Proceedings of the 3rd Annual Conference on Privacy, Security and Trust. Washington D. C., USA: IEEE Press, 2005: 1-6.

[9] BOUZIDA Y, CUPPENS F, CUPPENS-BOULAHIA N, et al. Efficient intrusion detection using principal component analysis[C]//Proceedings of the 3rd Conference on Security and Network Architectures. La Londe, France: [s. n.], 2004: 1-12.

[10] THASEEN I S, KUMAR C A. Intrusion detection model using fusion of PCA and optimized SVM[C]//Proceedings of International Conference on Contemporary Computing and Informatics. Washington D. C., USA: IEEE Press, 2015: 879-884.

[11] AHMAD I, HUSSAIN M, ALGHAMDI A, et al. Enhancing SVM performance in intrusion detection using optimal feature subset selection based on genetic principal components[J]. Neural Computing and Applications, 2014, 24(7/8): 1671-1682.

[12] ABUROMMAN A A, REAZ M B I. Ensemble of binary SVM classifiers based on PCA and LDA feature extraction for intrusion detection [C]//Proceedings of Advanced Information Management, Communicates, Electronic and Automation Control Conference. Washington D. C, USA: IEEE Press, 2017: 636-640.

[13] HADRI A, CHOUGDALI K, TOUAHNI R. Intrusion detection system using PCA and fuzzy PCA techniques [C]// Proceedings of International Conference on Advanced Communication Systems and Information Security. Washington D. C., USA: IEEE Press, 2016: 1-7.

[14] ELKHADIR Z, CHOUGDALI K, BENATTOU M. Intrusion detection system using PCA and kernel PCA methods[J]. International Journal of Computer Science, 2016, 43: 72-79.

[15] HETTICH S, BAY S D. KDD cup 1999 data[EB/OL]. [2018-02-08]. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

[16] 李丛, 闫仁武, 朱长水, 等. 融合 FAST 特征选择 ABQGSA-SVM 的网络入侵检测[J]. 计算机应用研究, 2017, 34(7): 2172-2179.

编辑 刘盛龄