基于 KNN 算法及禁忌搜索算法的特征选择方法在入侵检测中的应用研究

张 昊,陶 然,李志勇,蔡镇河
(北京理工大学信息科学技术学院,北京 100081)

摘 要: 在入侵检测中应用特征选择能够在保持原有信息完整性的基础上,去除其中的冗余特征,有效地提高入侵检测系统的检测速度. 本文提出了一种新的特征选择方法,即基于 KNN 算法及禁忌搜索算法的特征选择方法. 实验结果表明该特征选择方法能够有效去除网络数据信息中的冗余特征,减少特征选择时间;并且能够在保证检测准确率的前提下,有效提高系统的检测速度.

关键词: 入侵检测; 特征选择; 特征关联性; 禁忌搜索
中图分类号: TN915.08 **文献标识码**: A **文章编号**: 0372-2112 (2009) 07-1628-05

A Research and Application of Feature Selection Based on KNN and Tabu Search Algorithm in the Intrusion Detection

ZHANG Hao, TAO Ran, LI Zhi-yong, CAI Zhen-he
(School of Information Technology, Beijing Institute of Technology, Beijing 100081, China)

Abstract: Utilizing the feature selection in the intrusion detection can delete the redundant features on the base of protecting the integrity of original data and improve the detection speed of the system efficiently. This paper proposes a new feature selection method that is based on KNN and Tabu search algorithm. The experiment result shows that this method can remove the redundant features, and reduce the time of feature selection. This method not only can guarantee the correct rate of detection but also improve the detection speed efficiently.

Key words: intrusion detection; feature selection; feature relevance; Tabu search

1 引言

入侵检测技术是网络安全的一个重要研究领域. 网络级的入侵检测可以分为数据包的捕获、数据包的预处理以及对数据包进行攻击检测的过程[1]. 系统通过分析网络流量或系统审计纪录等,来发现网络或系统中是否有违反安全策略的攻击行为,以便系统管理员采取有效的措施,弥补系统漏洞和填补系统功能[2,3],因而系统要对网络中的海量数据进行分析,这样必然会影响系统处理的实时性,降低系统的检测速度,严重影响系统的检测效果. 因此,在保证检测正确率的前提下,如何提高系统的检测速度已经成为当前的一个研究热点. 研究者发现特征选择可以在保持原有网络数据信息完整性的基础上,去除其中的冗余特征,从而达到提高系统检测速度的目的. 从现有特征选择算法来看,文献[4]中 Jain 等人提出了诸如正向搜索(forward searching)、反向搜索(backward searching)、顺序搜索(sequential searching)等启发式搜索策略(heuristic searching strategies),文献[5]中 Kudo 等人提出了比启发式搜索更有优势的的随机搜索策略(stochastic searching strategy),如遗传算法(genetic algorithm). 在大规模数据集上的特征选择,这些搜索策略的计算资源耗用大,收敛速度慢,时间复杂度较高. 而在文献[6]中提出的非搜索性策略,虽然其时间复杂度相对较低,但是所得到的特征子集中有较多冗余特征,严重影响了分类准确率.

我们针对现有这些算法中所存在的缺点,提出了一种新的特征选择方法,即基于 KNN 算法及禁忌搜索算法的特征选择方法. 禁忌搜索算法是一种具有很强搜索能力的全局逐步寻(global stepwise optimization)优算法,是局部搜索(local search)的一种扩展. 但是禁忌搜索算法自身也存在着不足,它对初始解(initial solution)有很强的依赖性,单独使用禁忌搜索算法通常难以达到令人

收稿日期:2008-03-10;修回日期:2009-03-16
基金项目:国防基础科研项目(No. C1120060497-06)

满意的结果. 因此,本文对禁忌搜索算法进行了改进,首先利用基于特征关联(feature correlation)性的 KNN 算法来消除网络数据中的冗余特征,并将所得到的特征子集提供给禁忌搜索算法作为其初始解. 然后使用禁忌搜索算法对其邻域进行搜索,并得到最优特征子集. 经过改进的禁忌搜索算法兼具了搜索性特征选择算法与非搜索性特征选择算法的优点,能够较好的解决搜索性算法时间复杂度较高,以及非搜索性算法特征子集冗余特征较多,分类准确率低的问题.

2　基于 KNN 算法及禁忌搜索算法的特征选择

2.1　禁忌搜索算法

禁忌搜索的思想最早是由 Glover 等人在 1985 年提出的,并由 Glover 在 1986 年、1989 年和 1990 年对该方法做出了进一步的定义和发展. 禁忌搜索算法是对局部邻域搜索的一种扩展,是一种全局逐步寻优算法,是对人类智力过程的一种模拟. 它是在邻域搜索(neighborhood search)的基础上,通过设置禁忌表(tabu list)来禁忌一些已经历的操作,并利用藐视准则(aspiration criterion)来奖励一些优良状态,其中初始解、邻域结构、候选解(candidate solution)、禁忌长度、禁忌对象、藐视准则、终止条件(stop criterion)等是影响禁忌搜索算法性能的关键.

禁忌搜索算法的基本思想是[7]:给定一个初始解 X,并令其为暂定最优解 TempBest,选取合适的邻域,并在邻域中确定一定数量的候选解,组成候选解集合 $N(X)$. 然后在 $N(X)$ 中寻找最优解 Y,若 Y 对应的目标值优于当前最优解,且其不在禁忌表中,且不满足藐视准则,则在 $N(X)$ 中寻找次优解,进行上面的判断,直至在候选解集中找到这样的解,并进行相应的入表、修改表中各项的步骤. 如 Y 的目标值不优于当前最优解 TempBest 的目标值,则重新在 X 的邻域中寻找合适数量的解,组成新的候选解集合,继续后续的寻找 Y、与 TempBest 比较、修改禁忌表的步骤,直到满足终止条件,并输出 TempBest.

禁忌搜索算法的具体步骤描述如下:

(1) 给定算法参数,随机产生初始解 X,设置禁忌表为空;

(2) 判断算法终止条件是否满足. 如果满足,则结束算法并输出优化结果;如果不满足,继续以下步骤;

(3) 利用当前解的邻域函数产生其所有邻域解,并从中确定若干候选解;

(4) 对候选解判断藐视准则是否满足. 若满足,则用满足藐视准则的最佳状态 Y 代替 X 成为新的当前解,即 $X = Y$,并用与 Y 对应的禁忌对象替换最早进入禁忌表的禁忌对象,同时用 Y 替换 TempBest,然后转步骤(6);若不满足,继续以下步骤;

(5) 判断候选解对应的各对象的禁忌属性,选择候选解集中非禁忌对象所对应的最佳状态为新的当前解,同时用与之对应的禁忌对象替换最早进入禁忌表的禁忌对象元素;

(6) 转步骤(2).

我们可以很明显地看到,邻域函数、禁忌对象、禁忌表和藐视准则,构成了禁忌搜索算法的关键. 其中,邻域函数沿用局部邻域搜索的思想,用于实现邻域搜索;禁忌表和禁忌对象的设置,体现了算法避免迂回搜索的特点;藐视准则,则是对优良状态的奖励,它是对禁忌策略的一种放松.

需要指出的是,由于禁忌搜索算法具有灵活的记忆功能和藐视准则,并且在搜索过程中可以接受劣解,所以具有较强的爬山能力,搜索时能够跳出局部最优解,转向解空间的其他区域,从而增加获得更好的全局最优解的概率,所以禁忌搜索算法是一种局部搜索能力很强的全局迭代寻优算法. 但是禁忌搜索也有明显不足,即它对初始解的依赖性较强,好的初始解有助于搜索很快的达到最优解,而较坏的初始解往往会使搜索很难或不能达到最优解. 因此为了提高搜索效率,我们需要为禁忌搜索算法找到较好的初始解.

2.2　使用 KNN 算法生成禁忌搜索算法所需初始解

根据入侵检测数据的高维特性,我们采用 KNN 算法来获得禁忌搜索算法所需要的初始解. KNN 算法利用特征间的关联性来消除高维网络数据中的冗余特征,并将得到的特征子集作为禁忌搜索算法的初始解.

首先我们需要对特征关联性进行度量. 特征关联性的度量方法可以分为两类,一类是线性关联,包括:线性关联系数、Pearn 积矩相关、最小衰减误差平方和最大信息压缩等. 另一类建立在信息理论上,如熵等. 在本文中我们采用最大信息压缩指数作为特征关联性度量[6]:

$$2\lambda_{XY} = (D(X) + D(Y) - \sqrt{(D(X) + D(Y))^2 - 4D(X)D(Y)(1 - \rho_{XY}^2)}) \quad (1)$$

其中,$\rho_{XY} = \frac{COV(X, Y)}{\sqrt{D(X)D(Y)}}$,$COV()$,$D()$ 分别为协方差和方差.

该算法首先根据公式(1)计算特征集中每个特征及其第 K 个相近特征之间的 λ 值,选中 λ 最小的特征,保留该特征,删除其 K 个相近特征. 在该过程中设置一个门限值 ε,另其值为上一次循环中被删除的第 K 个近邻的 λ 值. 在下面的循环中,比较第 K 个近邻的 λ 值是否大于 ε 值,如果大于,则减小 K 的值. 因此,K 值在循环过程中是减小的.

设原始特征数目为 D;原始特征集为 $T=\{F_i, i=1,\cdots,D\}$;R 为消除冗余的特征子集;$\lambda(F_i,F_j)$ 表示特征 F_i,F_j 之间的不相关性,λ 值越大,特征越不相关;r_i^k 表示特征 F_i 与 R 中其第 K 个相近特征的不相关性.

初始解生成算法(KNN 算法)的具体步骤如下[4]:

(1) 选择初始 K 值,$K\leq D$,初始化 R,使 $R=T$;

(2) 对于每个特征 $F_i\in R$,计算 r_i^k;

(3) 找出令 r_i^k 最小的特征 F_i',保留 F_i 并删除 F_i 的 K 个最相近特征,$\varepsilon=r_i^k$';

(4) 如果 $K>cardinality(R)-1$,则 $K=cardinality(R)-1$;

(5) 如果 $K=1$,跳至步骤(8);

(6) 如果 r_i^k'$>\varepsilon$,则 $K=K-1$,r_i^k'$=\inf_{F_i\in R} r_i^k$(K 会不断减 1,直到有至少一个第 K 个近邻的 λ 值小于 ε);如果 $K=1$,则转至步骤(8)(如果 R 中没有特征和其近邻的 λ 值比 ε 小,就选择 R 中所有特征);

(7) 如果 r_i^k'$\leq\varepsilon$,跳至步骤(2);

(8) 输出 R.

初始解生成算法(KNN 算法)流程如图 1 所示.

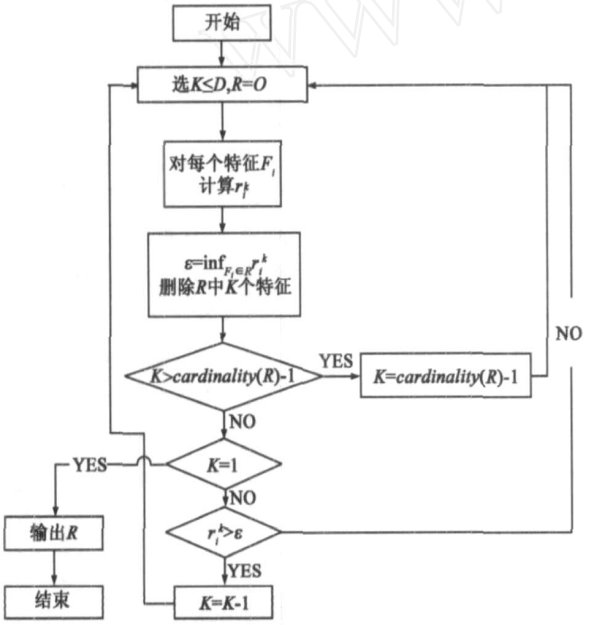

图1 初始解生成算法(KNN算法)流程

2.3 禁忌搜索算法参数设置

邻域结构:确定邻域结构,首先要定义算法的移动方式.在这里移动是指从一个解转向另一个解.对于任一可行解 X,只要定义了其移动方式,就定义了从该解经过一步移动所能到达的所有其他解,这个所有一步可达的解的集合就称为 X 的邻域.本文解决入侵检测中的特征选择问题时,将初始解 X 的邻域定义为,每次改变一个特征的状态,即加入一个特征或去掉一个特征.

候选解选择:候选解为初始解 X 的邻域中的所有 n 个特征.

禁忌表和禁忌长度:禁忌表采用队列结构实现,每次迭代将禁忌对象添加到队首,队列溢出时排在队尾的被删除.禁忌长度是指被禁对象在不考虑藐视准则的情况下允许选取的最大次数.在本文中我们设置禁忌长度为 n.

藐视准则:在禁忌搜索的过程中,可能会出现当前解的邻域解全部被禁止,或是某些解如果能解除其禁忌状态将能使目标函数值明显改善的情况.我们设置的藐视准则是:第一,位于禁忌表队首的解不能被解禁,避免陷入死循环;第二,要解禁的解必须优于当前最优解.

终止条件:设定最大迭代步数为 n,当禁忌搜索迭代到最大步数时,搜索结束.

2.4 基于 KNN 算法及禁忌搜索算法的特征选择方法描述

采用改进的禁忌搜索算法进行特征选择的具体实现步骤如下:

(1) 将特征向量用一个 0/1 串位表示,并使用 KNN 算法利用特征间的关联性来消除其中的冗余特征,将所得到的特征子集作为初始解 X,并将初始解作为迭代的起点;

(2) 根据所定义的邻域结构得到初始解 X 的邻域解集;

(3) 在邻域解集中寻找分类错误率最小的解.为了能够在适应不同的分类器类型的情况下,模拟实际工作中分类器错误率,本文采用了模拟错误率函数[8]:

$$e=e_0+\sum_{i_1,i_2} w_{i_1,i_2}\bar{\alpha}_{i_1}\bar{\alpha}_{i_2}+\varepsilon+\sum_{j=1}^{S} v_j\hat{\alpha}_{i_1}\cdots\hat{\alpha}_{i_k} \quad (2)$$

式(2)中 $\alpha=(\alpha_1,\alpha_2,\cdots,\alpha_n)$ 表示 n 维特征空间中的一个特征选择向量;$\alpha_i=0$ 表示第 i 个特征不在特征子集中,而 $\alpha_i=1$ 则表示第 i 个特征在特征子集中;$\bar{\alpha}_i=1-\alpha_i$;$e_0$ 为一个常数,表示基本错误率;w_{i_1,i_2} 表示删除第 i_1,i_2 个特征后所产生的错误率;ε 模拟了训练分类器时的误差;$\sum_{j=1}^{S} v_j\hat{\alpha}_{i_1}\cdots\hat{\alpha}_{i_k}$ 表示特定的特征组合对错误率的影响,其中 $\hat{\alpha}_i$ 可以是 α_i 或 $\bar{\alpha}_i$ 中的任意一个;S 表示所有特征进行组合后得到的总的组合数目;v_j 表示第 j 个特征组合对错误率的影响[8];

(4) 寻找满足藐视准则的解.若能够找到,则执行步骤(5);若未能找到,则直接将步骤(3)的结果作为最优解;

(5) 对步骤(3)和步骤(4)的结果进行比较,从中找出最优解;

(6) 对禁忌表进行相应修改;

(7) 判断是否满足终止条件,如果满足终止条件,则输出最优解;如果不满足终止条件,则返回步骤(2),重复搜索过程.

特征选择流程如图 2 所示.

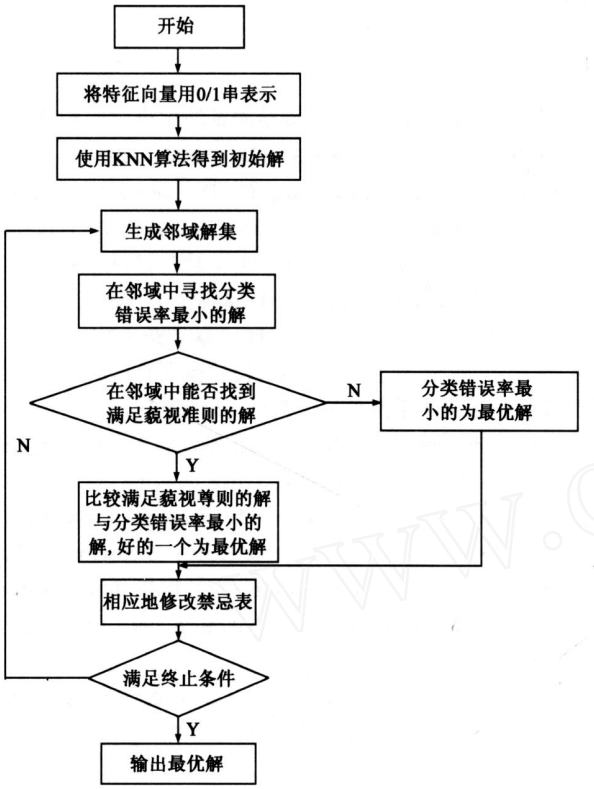

图2　基于KNN算法及禁忌搜索算法的特征选择流程图

3　实验分析

3.1　实验数据集描述

为了能够准确地对本文所提出的特征选择算法进行评价,本文选用了 KDD99 数据集作为实验数据集. KDD99 数据集是在入侵检测领域中广泛使用的数据集,它主要分为训练数据集以及测试数据集两部分. 该数据集提供了从一个模拟美国空军局域网上采集的 9 个星期的网络连接数据,其中训练数据集包含了 7 个星期的大约五百万条连接记录,测试数据集包含了 2 个星期的大约两百万条连接记录,其中每个记录包含 41 个特征. 因为 KDD99 数据集中的数据量非常庞大,为了能够便于验证本算法,我们需要对数据集进行取样,使数据量减少. 我们分别对训练数据集与测试数据集中的数据采取随机抽样,然后将随机抽样出的数据进行组合,形成实验用的训练数据子集和测试数据子集,其中,训练子集共有 11927 条,测试子集共有 24015 条.

3.2　实验方案设计

为了可以更为清晰、准确地验证本文所提出的特征选择方法性能,我们设计了以下三个实验方案.

方案一:首先在随机抽样后的数据集上应用本文提出的特征选择方法,得到对应的特征子集,然后分别在训练集上基于所有 41 个特征和特征选择后的特征子集建立入侵检测模型. 比较基于所有 41 个特征的入侵检测模型和基于特征选择后的特征子集的模型在检测时间以及检测准确率方面的性能.

方案二:首先在数据集上应用本文所提出的特征选择算法,得到对应的特征子集,并基于该特征子集建立入侵检测模型,然后采用遗传算法(GA 算法)和 Relief 算法对同一训练数据集进行特征选择,最后分别对采用本算法的入侵检测模型与应用遗传算法和 Relief 算法的入侵检测模型在检测时间和检测准确率方面的性能进行比较.

方案三:使用 SVM 作为分类器,比较方案二在特征选择后得到的三个特征子集用于样本分类时所产生的分类错误率.

3.3　实验结果分析

首先在随机抽样后的实验用训练数据集上应用本文提出的特征选择算法,所产生的特征子集见表 1.

表 1　应用特征选择算法后得到的特征子集

攻击类型	得到的特征子集
DOS	dos ,protocol _ type ,src _ bytes ,count ,dst _ host _ same _ srv _ rate
PROBE	probe ,duration ,service ,src _ bytes ,dst _ bytes ,count ,dst _ host _ diff _ srv _ rate
R2L	R2L ,duration ,service ,src _ bytes
U2R	U2R ,duration ,service ,src _ bytes ,root _ shell ,dst _ host _ count
NORMAL	protocol _ type ,service ,src - bytes ,count ,dst _ host _ count

从表 1 中我们可以看出经过特征选择后得到的特征子集中的特征个数只有 5 个左右. 接下来我们针对每一种攻击类型,分别在未经过特征选择的数据集以及表 1 所示的特征选择后的特征子集上建立入侵检测模型,并对每个入侵检测模型在检测时间以及检测准确率上的表现进行综合评价. 综合评价结果见表 2.

表 2　特征选择前后检测时间及检测准确率比较结果

入侵检测模型	检测时间(sec)		检测准确率	
	所有特征	特征子集	所有特征	特征子集
DOS	1.09	0.22	84.9 %	98.2 %
PROBE	1.16	0.28	88.0 %	97.8 %
R2L	0.98	0.19	87.5 %	99.0 %
U2R	1.23	0.31	85.4 %	97.1 %
NORMAL	1.13	0.29	87.8 %	98.0 %

从表 2 中我们可以看出,基于本文所提出的特征选择方法的入侵检测模型在检测时间以及检测准确率方面的表现都要优于未经特征选择的入侵检测模型.

在随机抽样后的实验用训练数据子集上分别应用本文提出的特征选择方法,以及基本遗传算法和 Relief 算法,实验结果如表 3 所示.

表 3 本文算法以及遗传算法和 Relief 算法性能比较

入侵检测模型	检测时间(sec)			分类准确率		
	本文算法	GA 算法	Relief 算法	本文算法	GA 算法	Relief 算法
DOS	0.22	0.36	0.42	98.2%	98.9%	95.6%
PROBE	0.28	0.41	0.45	97.8%	97.4%	96.0%
R2L	0.19	0.32	0.37	99.0%	98.5%	94.2%
U2R	0.31	0.49	0.43	97.1%	96.4%	95.9%
NORMAL	0.29	0.43	0.46	98.0%	97.8%	94.5%

最后,我们使用 SVM 作为分类器,将本文提出的特征选择方法以及基本遗传算法和 Relief 算法对实验训练数据集进行特征选择后得到的特征子集作为分类样本,实验结果如表 4 所示.

表 4 本文算法同 GA 及 Relief 算法在特征选择时间和分类准确率上的比较结果

数据集	特征选择时间(sec)			分类准确率		
	本文算法	GA 算法	Relief 算法	本文算法	GA 算法	Relief 算法
训练子集	0.22	0.36	0.42	98.2%	98.9%	95.6%
测试子集	0.48	0.57	0.54	97.8%	97.4%	94.7%

从表 4 中我们可以看出,由本文提出的特征选择方法和基本遗传算法所产生的特征子集作为分类样本的分类准确率差别不大,并且准确率明显高于 Relief 算法;而在特征选择时间方面,本文提出的方法则是表现最好的.

结合实验方案一、实验方案二以及实验方案三的结果我们可以看出,本文提出的特征选择方法,非常有效地减少了网络数据信息的特征维数;在此基础上所建立的入侵检测模型在检测时间及检测准确率方面的表现明显优于未经特征选择的入侵检测模型;在同 GA 算法及 Relief 算法进行比较后,可以看出本方法能够在保证较高的分类准确率的情况下,使特征选择的时间复杂度降低,较为有效地缩短特征选择的时间.

4 结论

本文提出了一种基于 KNN 算法及禁忌搜索算法的特征选择方法,本方法首先利用基于特征关联性的 KNN 算法来生成禁忌搜索算法所需的初始解.然后使用禁忌搜索算法对其邻域进行搜索,从而得到最优特征子集.通过建立入侵检测模型对 KDD99 数据集进行实验验证表明,本文提出的特征选择方法应用于入侵检测系统后,能够在保证检测准确率的前提下,有效地提高系统的检测性能,并且在检测时间和分类准确率方面的表现都要优于现有特征选择方法,证明了本文方法的有效性和可行性.

参考文献:

[1] 牟永敏,李美贵,梁琦. 入侵检测系统中模式匹配算法的研究[J]. 电子学报,2006,34(12A):2488 - 2490.
Mu Yongmin, Li Meigui, Liang Qi. The survey of the pattern matching algorithm in intrusion detection system[J]. Acta Electronica Sinica, 2006, 34(12A):2488 - 2490. (in Chinese)

[2] Endorf C, Schultz E, Mellander J. Intrusion Detection & Prevention[M]. New York, McGraw-Hill, 2004. 386 - 391.

[3] 周鸣争,楚宁,强俊. 基于构造性核覆盖算法的异常入侵检测[J]. 电子学报,2007,35(5):862 - 867.
Zhou Mingzheng, Chu Ning, Qiang Jun. An anomaly intrusion detection based on constructive kernel covering algorithm[J]. Acta Electronica Sinica, 2007, 35(5):862 - 867. (in Chinese)

[4] Jain A K, Zongker D. Feature selection: Evaluation, application, and small sample performance[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, 19(2):153 - 158.

[5] Kudo M, Sklansky J. Comparison of algorithms that select features for pattern classifiers[J]. Pattern Recognition, 2000, 33(1):25 - 41.

[6] Mitra P, Murthy C A, Pal S K. Unsupervised feature selection using feature similarity[J]. IEEE Trans Pattern Recognition and Machine Intelligence. 2002, 3(24):301 - 312.

[7] Glover F. Tabu search[J]. ORSA J. Computing, 1990, 2II(1):4 - 32.

[8] Siedlecki W, Sklansky J. A note on genetic algorithm for large - scale feature selection[J]. Pattern Recognition Letters, 1989, 10(11):335 - 347.

作者简介:

张 昊 男,1981 年 10 月生.北京理工大学信息科学技术学院电子工程系博士研究生,研究方向信息安全与对抗.
E-mail:kuobai@126.com

陶 然 男,1964 年生.北京理工大学电子工程系信息安全与对抗学科首席教授,博士生导师.

李志勇 男,1977 年生.北京理工大学信息科学技术学院电子工程系博士研究生,研究方向信息安全与对抗.

蔡镇河 男,1985 年出生.北京理工大学信息科学技术学院电子工程系硕士研究生,研究方向信息安全与对抗.