文章编号:1009-3087(2015)01-0036-06 DOI:10.15961/j.jsuese.2015.01.005

基于选择性集成分类器的通用隐写分析

张敏情[1,2],狄富强[1*],刘 佳[1,2]
(1.武警工程大学 电子技术系 网络与信息安全武警部队重点实验室,陕西 西安 710086;
2.武警工程大学 网络与信息安全研究所,陕西 西安 710086)

摘 要:面对高维度的特征集和大规模的样本集,隐写分析技术对分类器的要求越来越高。在集成分类器的基础上提出了一种面向通用隐写分析的选择性集成分类器。首先基于随机森林生成若干个基分类器,然后利用基于遗传算法的选择性集成算法剔除掉个别影响整体性能的基分类器,最后根据遗传优化得到的最优权值向量赋予剩余的基分类器不同权值以用来加权投票集成。实验表明,提出的选择性集成分类器测试性能优于现有分类器,特别在基分类器数量较大、特征维数较高时与现有集成分类器相比,有效降低了检测错误率。

关键词:隐写分析,集成分类器,选择性集成,遗传算法,加权投票
中图分类号:TP309 **文献标志码**:A

Universal Steganalysis Based on Selective Ensemble Classifier

ZHANG Minqing[1,2], *DI Fuqiang*[1*], *LIU Jia*[1,2]
(1. Key Lab. of Network & Info. Security, Electronic Dept., Eng. Univ. of the Armed Police Force, Xi'an 710086, China;
2. Inst. of Network & Info. Security, Eng. Univ. of the Armed Police Force, Xi'an 710086, China)

Abstract: With massive feature set and high-dimensional sample set, steganalysis has a increasingly demanding for classifiers. Based on ensemble classifier, a kind of selective ensemble classifier for universal steganalysis was proposed. At first, some base learners were generated based on the random forest and then some of them were wept out using GASEN (genetic algorithm based selective ensemble) algorithm. At last, remaining base classifiers were given different weights according to the optimal weight vector from genetic optimization to get used to the weighted vote integration. Experiments showed that the elective ensemble classifier performed better than existing single classifier. Compared with the existing ensemble classifier, especially in the case of larger base classifiers or higher number of features, the computational complexity was slightly increased, but the error rate reduced effectively.

Key words: steganalysis; ensemble classifier; selective ensemble; genetic algorithm; weighted vote

当前基于统计的隐写分析技术主要通过有监督的分类器来实现,其基本思路是先提取对嵌入信息敏感的统计特征,然后利用训练好的分类器进行分类。对于类似 HUGO 等新的隐写方法[1-3],为了寻求能更好表征图像的模型,隐写分析者所提取的特征越来越高[4-6]。当面对过高的特征维数和大规模的样本集时,大多数现有分类器因训练时间过长已不再适用。设计出更为准确高效的分类器和分类方法是相关领域亟须解决的问题。Kodovsky 等[7]把集成分类的思想引入到数字媒体隐写分析领域,通过对若干基分类器的检测结果进行投票得到最终结果,虽然降低了训练复杂度但分类准确率和预测速度有待提高。该方法未考虑不同基分类器预测性能的差异,集成投票阶段每一个基分类器的权值相同,而且随着基分类器数目增多不仅预测速度明显下降,其所需的存储空间也会迅速增加。作者首先利

收稿日期:2014-06-24
基金项目:国家自然科学基金资助项目(61379152);陕西省自然科学基金资助项目(2014JQ8301)
作者简介:张敏情(1967—),女,教授,博士.研究方向:密码学;信息安全. E-mail:api_zmq@126.com
* 通信联系人 E-mail:18710752607@163.com
网络出版时间:2014-9-10 11:11:32 **网络出版地址**:http://www.cnki.net/kcms/detail/51.1596.T.20140910.1111.006.html

用基于遗传算法的选择性集成算法(genetic algorithm based selective ensemble,GASEN)剔除掉影响集成分类整体性能的部分基分类器,然后利用最优权值向量对所有基分类器的结果进行加权融合得到最终结果。实验表明该方法有效地提高了分类准确率和预测速度。

1　基于遗传优化的选择性集成方法

集成学习能显著提高一个学习系统的泛化能力[8]并被专家列为机器学习领域 4 大研究方向之首。随着基学习器数目增多,集成学习器不仅预测速度会明显下降,其所需的存储空间也将迅速增加[9]。周志华等[10]首次提出的选择性集成(selective ensemble)思想,由于其在提高泛化能力和降低预测阶段的开销方面存在优势,近年来在国内外集成学习领域引起了强烈反响并成为研究的热点。选择性集成假定已经产生多个基分类器,并基于某种策略来优中选优,可以在使用更少基学习器的情况下取得比全部集成更好的性能。

1.1　选择性集成的理论基础

假设用于隐写分析的训练集大小为 N,基分类器个数为 m,y_i 和 f_{ji} 分别代表样本 i 的预期输出和样本 i 在第 j 个基分类器中的实际输出。损失函数 $L(a,b)$ 定义为:当 $a=b$ 时值为 0,否则值为 1,则第 j 个基分类器在数据集上的泛化误差为:

$$E_j = \frac{1}{N}\sum_{i=1}^{N} L(y_i, f_{ji}) \tag{1}$$

若第 i 个分类器的权重为 w_i,则整个集成分类器在该数据集上的泛化误差为:

$$E = \frac{1}{N}\sum_{i=1}^{N} L(y_i, E_{ji}) \tag{2}$$

其中,$E_{ji} = \arg\max_a(W_a)$,$a \in \{1,0\}$,$a=1$ 代表图像经过隐写,否则没有隐写。$W_a = \sum_{j=1}^{m}\{\amalg(f_{ji}=a)w_i\}$,$\amalg(\cdot)$ 仅当括号内条件成立时为 1,否则为 0。现在从集成分类器中删除第 t 个基分类器,则剩余基分类器构成的集成分类器在该数据集上的泛化误差为:

$$E = \frac{1}{N}\sum_{i=1}^{N} L(y_i, E_{ji}) \tag{3}$$

其中,$E_{ji} = \arg\max_a(W_a)$,$W_a = \sum_{j=1}^{m}\{\amalg(f_{ji}=a)w_i - \amalg(f_{ti}=a)w_t\}$,若

$$E - E = \frac{1}{N}\sum_{i=1}^{N}\{L(y_i, E_{ji}) - L(y_i, E_{ji})\} > 0 \tag{4}$$

则代表删除第 t 个基分类器后不仅可以降低预测开销还能提高泛化能力。

1.2　GASEN 算法

遗传算法(genetic algorithm,GA)是一种基于进化论和遗传学的现代优选算法,具有极高的鲁棒性和广泛的适用性。选择性集成的基本思路就是从众多可行的方案中优选出最佳的解决方案。GASEN 算法采用遗传算法选出个体学习器,基本思想是给每个学习器都指定一个能够刻画出学习器在集成中重要性的权重从而决定学习器的去留,如果权重小于预设的阈值就将其对应的分类器去掉。具体做法如下:

首先用自助法产生基学习器,然后赋予可以反映在集成时重要性的权值 w_i,把所有学习器权值组成的权值向量 $\boldsymbol{W}=[w_1,w_2,\cdots,w_m]$ 作为遗传算法中的种群个体(染色体)。通过遗传算法求出最优权值向量并归一化处理,借助预设的阈值对该向量进行选择。定义 $\hat{E}_w^V$ 为权值向量 $\boldsymbol{W}$ 对应的集成在验证集 V 上的误差。伪代码为:

算法 1　GASEN 算法

输入:训练集 S,验证集 V,学习算法 L,迭代次数 T,阈值 λ。

实现过程:

1. For　$t=1$ to T {
2.　　$S_t = \text{Bootstrap}(S)$;
3.　　$N_t = L(S_t)$;
4. }
5. 产生权值向量 $\boldsymbol{W}=[w_1,w_2,\cdots,w_m]$;
6. 以 $f(w)=1/\hat{E}_w^V$ 为适应度函数对 $\boldsymbol{W}$ 利用遗传算子进行进化;
7. 得到最优权值向量 $\boldsymbol{W}^*$。

输出:

$$N^*(x) = \arg\max_{y\in Y}\sum_{w_t^*>\lambda; N_t(x)=y} 1。$$

2　基于选择性集成分类的隐写分析

2.1　特征子空间构造

近年来隐写分析的发展趋势表明更复杂和高维的图像模型能本质上带来更好的检测效果。作者利用基于空域的富模型[4],挑选出 SRMQ1 特征[4]中表现较为理想的部分特征用于集成分类。为了降低训练复杂度,每个基分类器在一个特征子集而不是整个特征空间上训练,并且特征子集维数远远小于特征集 F 的总维数。为了增加基分类器之间的差异性,对训练集 M 的采样方式基于随机森林中有放

回重采样的 bootstrap 技术(自助法),因此约 37% 的原样本数据属于未被抽到的袋外数据(out-of-bag,OOB)。基于随即森林生成的 L 个基分类器,用验证集 V 经遗传优化后仅保留了 S 个性能较好的基分类器,最后对测试集 T 加权投票得到预测结果 R,整个选择性集成分类的流程如图 1 所示。

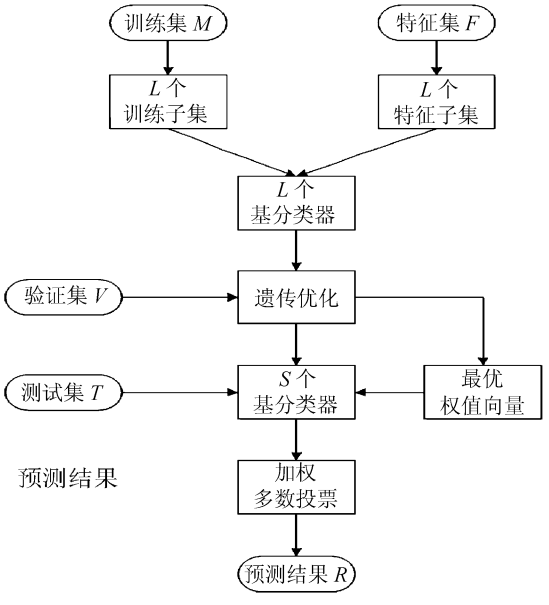

图 1 算法流程图

Fig. 1 Flow chart of algorithm proposed

定义第 m 个训练样本在隐写前后的特征分别为 X_m 和 X_m,N^{trn} 和 N^{tst} 分别代表训练集和测试集大小,$B^{(n)}(X)$ 代表第 n 个基分类器对特征 X 的测试结果,则第 n 个基分类器对应的袋外数据测试误差 OOB_n 定义如下:

$$OOB_n = \frac{1}{2N^{\text{trn}}}\sum_{m=1}^{N^{\text{trn}}}\left(B^{(n)}(X_m) + 1 - B^{(n)}(X_m)\right) \tag{5}$$

2.2 基分类器的生成

设计的集成分类器是由若干个结构简单,运算速度较快的费舍尔线性判别分类器(Fisher linear discriminate,FLD)构成的。假设特征总维数为 d,每一个基分类器所用的特征子空间维数为 d_{sub},以下是基于随机森林的基分类器生成过程:

算法 2 基分类器生成算法

输入:特征子空间维数 d_{sub},训练集中样本对的特征 $X^{\text{trn}} = \{x_m, x_m\}_{m=1}^{N^{\text{trn}}}$。

实现过程:

1. For $l = 1$ to L {

2. 产生随机子空间 $D_l = \{1,2,\cdots,d\}$,$|D_l| = d_{\text{sub}} < d$;

3. 产生随机样本 $N_l = \text{Bootstrap}(N^{\text{trn}})$,$|N_l| = N^{\text{trn}}$;

4. 取 N_l 和 D_l 对应的特征集合 $X_l = \{x_m^{(D_l)}, x_m^{(D_l)}\}_{m\in N_l^b}$ 训练基分类器 B_l}。

输出:$OOB = \{OOB_1, OOB_2, \cdots, OOB_l\}$,基分类器参数 V_l 和阈值 T_l。

2.3 基分类器的选择性集成

选择性集成的实质是在众多基分类器中通过优化选择一个精选的基分类器集合。首先对 2.2 节产生的基分类器编号,使用权值向量 $\boldsymbol{W} = [w_1, w_2, \cdots, w_l]$ 对分类器组合进行实值编码后得到一个染色体 h,其中,权重 w_i 作为一个基因对应第 i 个分类器,每一种 h 对应着一种基分类器的集成方式。在遗传进化过程中,由多个染色体组成一个世代。在每一个世代处理过程中,经以下 2 种操作实现集成方式优选:

1)通过单浮点交叉来交换2个染色体的基因获得更优秀的染色体。

2)通过单个基因变异获得更优秀的基因。

遗传算法在寻优过程中,对初始种群的选择具有一定的依赖性,而且初始种群的选择会影响优化的速度和性能。实验发现由于GASEN算法产生初始种群的方法是随机产生,存在时间复杂度较大和容易陷入局部最优解这2个缺点。对此,本文算法充分利用了 2.2 节产生基分类器时得到能较好反映基分类器性能的 OOB 数据,并加入到初始种群中,得到了较好的效果。令

$$OOB_i^* = OOB_i / \sum_{i=1}^{L} OOB_i \tag{6}$$

OOB_i^* 作为归一化操作后的初始权重 w_i,并用初始权重向量 $\boldsymbol{W} = [w_1, w_2, \cdots, w_l]$ 编码成染色体 h^*。用自助法采样构成大小与训练集相同的验证集 V。以下是提出的选择性集成算法:

算法 3 基分类器选择性集成算法

输入:$OOB = \{OOB_1, OOB_2, \cdots, OOB_l\}$,验证集 V,测试集 D,阈值 λ,迭代次数 T。

输出:预报精度 Acc_{final}。

过程:

1. 初始化种群 $host = \{h^*, h_1, h_2, \cdots, h_m\}$,其中,$h_i$ 为随机生成的权值向量 $\boldsymbol{W} = [w_1, w_2, \cdots, w_l]$ 对应的实值编码 h_i;

2. for $t = 1, \cdots, T$

1)计算当前所有染色体的适应度:用染色体 $\boldsymbol{W}$ 作为基分类器权重计算出在验证集 V 上集成分类器

的加权预测错误率 $\hat{E}_w^V$,该染色体的适应度定义为 $f(w) = 1/\hat{E}_w^V$;

2) 对种群执行选择、交叉、变异操作

end;

3. 找出适应度最高的染色体权重向量 $\boldsymbol{w}^*$,若其中 $w_i \geq \lambda$,则对应的基分类器被保留。得到 S 个基分类器的组合 $S = \{C_1^*, C_2^*, \cdots, C_s^*\}$ 和对应的权值向量 $\boldsymbol{W}^* = [w_1^*, w_2^*, \cdots, w_s^*]$;

4. for $t = 1, \cdots, S$

用测试集 D 计算预报结果 $result = \{result_1^*, result_2^*, \cdots, result_s^*\}$;

end

5. $\boldsymbol{W}^*$ 作为 S 个基分类器的投票权重,对 $result$ 进行简单多数投票得最终结果 $result_{\text{final}}$ 和预报精度 Acc_{final}。

2.4　基于权值融合的集成策略

对于一个待测试样本 $y \in y_{\text{tst}}$,第 s 个基分类器用相对应的映射 $\boldsymbol{v}_s^{\mathrm{T}}\boldsymbol{y}^{(D_s)}$ 和对应的临界值 T_s 相比较,得到它的预测结果 $B_s(\boldsymbol{y}^{(D_s)})$。

$$S = \{C_1^*, C_2^*, \cdots, C_s^*\}$$

和

$$\boldsymbol{W}^* = [w_1^*, w_2^*, \cdots, w_s^*]$$

分别为 2.3 节得到的选择后的 S 个基分类器组合和对应的权值向量,令:

$$W_i = w_i^* / \sum_{j=1}^{S} w_j^* \tag{7}$$

被保留的基分类器经上式归一化后的权重作为其投票权重,经简单多数投票得到最终集成分类的结果如下:

$$B(y) = \begin{cases} 1, & \text{当} \sum_{s=1}^{S} W_s B_s(\boldsymbol{y}^{(D_s)}) > S/2; \\ 0, & \text{当} \sum_{s=1}^{S} W_s B_s(\boldsymbol{y}^{(Ds)}) < S/2; \\ \text{随机}, & \text{其他} \end{cases} \tag{8}$$

T_s 为第 s 个基分类器的判断阈值,则预测结果为:

$$B_s(\boldsymbol{y}^{(D_s)}) = \begin{cases} 1, & \text{当 } \boldsymbol{v}_s^{\mathrm{T}}\boldsymbol{y}^{(D_s)} > T_s; \\ 0, & \text{其他} \end{cases} \tag{9}$$

提出的基于选择性集成的隐写分析方法中基分类器优化过程总结为:

1) 增加基分类器:基于随机森林,根据 OOB 结果以及能接受的最大错误率决定是否增加一个基分类器,确定选择性集成前最终的基分类器个数。

2) 减少基分类器:基于 GASEN 算法,通过整体优化找出最佳的基分类器权值向量,根据某一权值是否小于预设阈值决定是否剔除该分类器。

3) 融合分类器:基于加权融合,根据遗传优化过程中得到的最优权值向量赋予剩余的基分类器权重,作为最终集成分类过程中的投票权重。

3　实验结果及分析

3.1　实验准备

由于目前对载体图像所建模型的复杂性和差异性越来越大,未来对隐写检测的实践验证不可避免地需要更大数量的样本集。为了验证本文方法的有效性,从 BOWS2、BossRank 以及自采集的图像库选取 1 000 幅载体图像、15 000 幅隐写图像和 4 000 幅混杂图像进行实验。其中,图像大小均为 512×512,质量因子为 80;隐写图像分别采用 MBS、nsF5、Outguess、YASS 和 HUGO。实验工具为 Matlab 7.12,遗传算法采用 Houck 开发的 GAOT(genetic algorithm for optimization toolbox)工具箱,试验中取每一世代的染色体数量为 20,阈值设为 0.05,最大训练世代为 100。

3.2　检测错误率比较

为了衡量算法的检测性能,传统的做法是使用 ROC 曲线,但它只能进行定性的描述。为了定量地描述算法的检测效果,参考文献[11]的做法,把漏报率(false negative rate, FNR)和虚警率(false positive rate, FPR)的平均值的最小值作为检测错误率(error rate, ER),并用来衡量算法的检测性能。其计算公式为:

$$ER = \min_{FNR} \frac{1}{2}(FNR + FPR) \tag{10}$$

首先选取经不同嵌入率的 nsF5 隐写图像,然后对 G-SVM(基于高斯核函数的支持向量机)、L-SVM(线性支持向量机)、文献[7]的集成分类器和本文分类器进行实验,结果如图 2 所示。

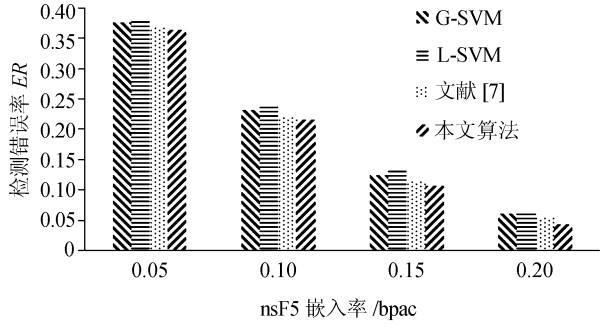

图 2　4 种分类器的检测错误率

Fig. 2　Error rate of four classifiers

结果表明,提出的隐写分析方法与现有分类器相比检测错误率有一定降低。同时发现,选择性集成分类器更能降低对高嵌入率隐写图像的检测错误率。这是因为高嵌入率隐写图像的特征具有较大的数值,其与载体图像的特征差异性更大,集成分类器以及选择性集成分类器更能够发挥集成优势,在一定程度上降低了检测错误率。

3.3 复杂度比较

分类器的训练复杂度和预测阶段开销也是衡量分类器性能的重要指标,其中分类器总体的运算复杂度和时间开销主要来自训练阶段。选取3.1节准备的一定数量(训练样本集大小分别为1 000、2 000、3 000、4 000、5 000)的不同图像对4种分类器的训练时间进行测试,结果见图3。

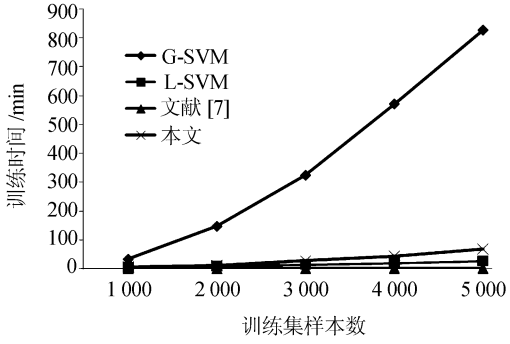

图3 4种分类器的训练时间

Fig. 3 Training time of four classifiers

实验表明,随着训练样本集的增大,以G-SVM分类器为代表的结构较为复杂的分类器因训练复杂度过大其可操作性会迅速下降。本研究以及文献[7]的集成分类器因为在训练阶段首先生成了若干结构简单、训练复杂度极低的FLD分类器,在面对同等规模的训练样本集时极大地减少了训练时间。提出的的分类器与文献[7]中的分类器比较,增加了剔除部分基分类器的遗传优化阶段,所以训练时间有一定增加。但该方法由于减少了基分类器的个数,在测试阶段以及预测阶段的时间开销会有所降低。

3.4 特征数较大以及基分类器较多时检测性能的比较

由于面对cf^*(7 850维)等高维特征和较大训练样本时,运用L-SVM、G-SVM等分类器进行检测的时间过长,仅选取文献[7]提出的分类器和提出的分类器对所有图像的检测错误率进行比较。表1给出经改进前后基分类器个数和平均错误率的比较结果。

表1 改进前后的性能比较

Tab. 1 Performance before and after improvement

文献[7]		本文算法	
基分类器个数	平均错误率	基分类器个数	平均错误率
100	0.271	12	0.269
300	0.196	35	0.176
500	0.152	52	0.127
700	0.132	78	0.095
1 000	0.076	113	0.026

实验结果表明,在特征维数较大和基分类器个数较多的情况下,该方法用于隐写分析的检测错误率有明显改善。分析原因主要是:

1)未经选择性集成的集成分类器生成若干基分类器虽然均为结构简单的FLD分类器,但因每个基分类器输入的训练样本集和特征子空间均来自随机选取,经过训练后分类器的检测精度存在差异。当基分类器较多时可能存在部分影响集成分类器整体性能的基分类器。提出的基于遗传优化的选择性集成在若干基分类器的可能权值组合中选择最优权值组合,之后只选取权值大于预设阈值的基分类器,有效提高了检测精度。

2)选择性集成分类器在选出需要保留的基分类器后利用遗传优化得到的最优权值向量对剩余基分类器进行基于权值的多数投票,与没有权值的多数投票方法相比,降低了整体的泛化误差。

4 总结及下一步工作

当前基于富模型和集成分类器的通用隐写分析方法因其较高的准确性和广泛的适用性已经成为当前隐写分析领域的研究热点。基于GASEN算法对生成的FLD分类器进行选择性集成,并赋予剩余的基分类器一定权重用来加权投票,生成了性能良好的分类器,取得了较好的实验效果。提出的选择性集成分类器具有良好的可扩展性和广泛的适用范围。下一步将继续改进遗传优化阶段构造的适应度函数,进一步提高选择性集成分类器的检测精度。

参考文献:

[1] Pevny T, Filler T, Fridrich J. Using high-dimensional image models to perform highly undetectable steganography [C]//Proceedings of Information Hiding 12th International Workshop. Heidelberg: Springer, 2010: 161 - 177.

[2] FridrichJ, Holub V. Challenging the doctrines of JPEG steganography [C]//Proceedings of SPIE, Electronic Ima-

ging, Media Watermarking, Security, and Forensics. San Francisco:SPIE,2014:32 – 39.

[3] Fridrich J, Denemark T, Holub V. Further study on the security of S-UNIWARD[C]//Proceedings of SPIE, Electronic Imaging, Media Watermarking, Security, and Forensics. San Francisco:SPIE,2014:24 – 30.

[4] FridrichJ, Kodovsky J. Rich models for steganalysis of digital images[J]. IEEE Transactions on Information Forensics and Security,2012,7(3):868 – 882.

[5] Kodovsky J, Fridrich J. Steganalysis of JPEG images using rich models[C]//Proceedings of SPIE, Electronic Imaging, Media Watermarking, Security, and Forensics of Multimedia XIV. San Francisco:SPIE,2012:10 – 17.

[6] Shi Y, Sutthiwan P, Chen Licong. Textural features for steganalysis[C]//Proceedings of 13th Information Hiding Conference. Prague:LNCS,2012:24 – 32.

[7] Kodovsky J, Fridrich J, Holub V. Ensemble classifiers for steganalysis of digital media[J]. IEEE Transactions on Information Forensics and Security, 2012, 7(2): 432 – 444.

[8] Zhou Zhihua. Ensemblemethods: Foundations and algorithms[M]. Boca Raton:CRC Press,2012:15 – 16.

[9] Zhang Chunxia. Research on some algorithms in ensemble learning[D]. Xi'an:Xi'an Jiaotong University,2010. [张春霞. 集成学习中有关算法的研究[D]. 西安:西安交通大学,2010.]

[10] Zhou Zhihua, Wu Jianxin, Tang Wei. Ensembling neural networks:Many could be better than all[J]. Artificial Intelligence,2002,137(1/2):239 – 263.

[11] Kodovsky J, Pevny T, Fridrich J. Steganalysis in high dimensions: Fusing classifiers built on random subspaces [C]//Proceedings of Electronic Imaging, Watermarking, Security, and Forensics of Multimedia XIII. San Francisco: SPIE,2011:1 – 13.

(编辑　杨　蓓)