ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 31,No. 1,Jan. 2016,pp. 197－204
DOI:10.16337/j.1004-9037.2016.01.022

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

一种限制输出模型规模的集成进化分类算法

宋文展　元昌安　覃　晓　周　凯　郑　彦

(广西大学计算机与电子信息学院,南宁,530004)

摘　要: AdaBoost 算法是一种典型的集成学习框架,通过线性组合若干个弱分类器来构造成强学习器,其分类精度远高于单个弱分类器,具有很好的泛化误差和训练误差。然而 AdaBoost 算法不能精简输出模型的弱分类器,因而不具备良好的可解释性。本文将遗传算法引入 AdaBoost 算法模型,提出了一种限制输出模型规模的集成进化分类算法(Ensemble evolve classification algorithm for controlling the size of final model,ECSM)。通过基因操作和评价函数能够在 AdaBoost 迭代框架下强制保留物种样本的多样性,并留下更好的分类器。实验结果表明,本文提出的算法与经典的 AdaBoost 算法相比,在基本保持分类精度的前提下,大大减少了分类器数量。

关键词: 集成学习;AdaBoost 算法;遗传算法;弱分类器

中图分类号: TP301.6　　**文献标志码**:A

Ensemble Evolve Classification Algorithm for Controlling Size of Final Model

Song Wenzhan, Yuan Changan, Qin Xiao, Zhou Kai, Zheng Yan

(Computer and Electronic Information College, Guangxi University, Nanning, 530004, China)

Abstract: AdaBoost algorithm is a typical ensemble learning framework. It linearly combines a set of weak classifiers to construct a strong one, whose classification accuracy, generalization error and training error are all improved. However, the AdaBoost algorithm is weak interpretability since it cannot simplify weak classifiers from output model. Hence, one presents a new algorithm, ensemble evolve classification algorithm for controlling the size of final model (ECSM), by introducing the genetic algorithm into the AdaBoost algorithm model. Gene evolution and fitness function can mandatory reserve the species diversity of samples in the AdaBoost iteration framework, and leave better classifiers. With keeping the classification accuracy, experimental results show that the proposed algorithm greatly reduce the number of classifiers compared with the classical AdaBoost algorithm.

Key words: ensemble learning; AdaBoost algorithm; genetic algorithm; weak classifier

引　　言

分类一直是数据挖掘的一个研究热点。与单一的分类模型相比较,集成学习[1]分类器通过联合一系列的弱分类器来组建分类模型,从而在分类精度和预测性能上都优于前者。集成学习是一种新的机

基金项目:国家自然科学基金(61363037)资助项目。
收稿日期:2014-09-15;**修订日期**:2014-10-18

器学习技术，由于集成学习具有较好的泛化误差与训练误差，因此它成为了国际机器学习界的关注热点。领域内已有一大批优秀的集成学习算法，比如随机森林、Bagging 和 Boosting[2,3]。AdaBoost 算法原理简单，但却是很有效的机器学习监督方法。作为一种迭代算法，AdaBoost 在每一次迭代中都对样本进行赋值，经过对样本权重的改变，使得下一轮中可以重点关注那些较难分类的样本，并且在每一轮迭代后都产生一个新的弱分类器。正是因为这种机制导致最后一次输出模型的弱分类器个数过多，难以清晰表达。本文拟改变对样本的训练机制，不进行累计赋值。此外，结合遗传进化算法，对分类器权重进行优化以达到控制模型输出规模。实验表明，本文提出的集成进化分类算法在保持分类精度的情况下能够很好地控制模型的输出规模。

1 相关工作

1.1 集成学习 AdaBoost 算法

AdaBoost 是由 Freund 和 Schapire 在 1995 年提出的一种 Boosting 改进算法[4]，一经提出就得到了国际学术界的广泛关注，并被评为数据挖掘十大算法之一。Boosting 也称为提升法，它是通过线性组合多个分类精度只比随机猜测略好的弱分类器来构成一个强学习器，与单个弱分类器相比它有着更好的分类精度[5]。AdaBoost 算法起源于 Valiant 在 1984 年提出的计算机学习模型可能近似正确理论(Probably approximately correct, PAC)[6]，理论中证明了弱学习算法与强学习算法等价性问题，即给定一个分类精度比较差的分类器，如 Decision stumps，能否经过一系列在不改变弱分类器分类原理的基础之上提升成为一个强学习器。

在数据分类领域，Schapire 对上述问题给出了肯定的回答，并提出了 Boosting 算法。Boosting 给出一批训练样本集 $S\{(x_1,y_1),\cdots,(x_m,y_m)\}$ 以及类标 $Y\{1,-1\}$，令 $h_1,h_2,\cdots,h_m,\cdots$ 为模型经过多次迭代后得出的一组弱分类器，则 Boosting 算法所得到的强学习器为

$$g(x)=\sum_{j=1}^{m}\alpha_j h_j(x) \tag{1}$$

式中：h_j 为弱分类器的预测值且 $h_j\in\{1,-1\}$；α_j 为弱分类器 h_j 的权重，α_j 与 h_j 都是在 Boosting 算法迭代过程当中得出。对于任意的一个样本 x，经过线性组合弱分类器所得出的预测值 y 为

$$y=\mathrm{sign}(g(x)) \tag{2}$$

式中：$g(x)>0$ 时，$y=1$；$g(x)<0$ 时，$y=-1$。

Boosting 存在的问题是如何调整训练样本，使得训练的弱分类器具有差异性。AdaBoost 算法是通过改变数据分布来改进的[7]：依据每一次迭代中样本是否被正确分类来决定样本在下一轮中的权值大小。每一次迭代都对训练样本进行一次赋值，本轮中被正确分类的样本将会降低其权值，相对而言，权值没有发生改变的误分类样本在下一轮中将会具有更高的权值，也就得到对误分类样本的重点关注。

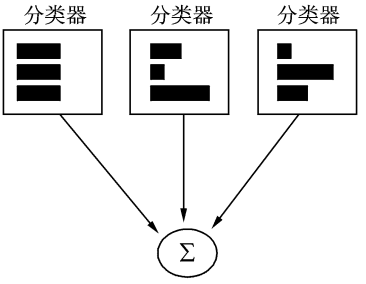

图 1 AdaBoost 原理示意图

Fig. 1 Outline of AdaBoost

Boosting 另一个需要解决的问题就是如何合理地联合多个弱分类器。AdaBoost 算法则是采用加权投票技术代替了平均投票机制，让分类精度较高的弱分类器具有更高的权值。如图 1 所示。

AdaBoost 的优良特性也包括较好的泛化能力，Freund 和 Schapire 在文献[8]中用弱学习假设空间的 VC 维来衡量泛化误差，并给出了算法的泛化误差上界[9]

$$P[H(x)\neq y]+O(\sqrt{Td/M}) \tag{3}$$

式中:$P[H(x)\neq y]$为训练样本集上的经验概率。由上式得出泛化误差与训练集的样本个数 M、训练的迭代次数 T 以及弱学习算法的 VC 维 d 相关。

除了 AdaBoost 算法外,还有其他一些非常优秀的 Boosting 家族算法,比如 Friedman 提出的 GentleBoost,Freund 提出的 BrownBoost 等。

传统的 AdaBoost 算法每次都对误分类的样本赋予相对更高的权值,如此迭代,误分类样本的权值将会持续增加。另外每一次迭代都会产生一个弱分类器,经过成百上千次的训练之后,将会产生非常多弱分类器,模型规模过于庞大且不利于表达。为有效解决以上两个问题,本文对 AdaBoost 进行如下改进:(1)改变 AdaBoost 的训练方式,不对样本进行累计赋值;(2)结合遗传算法对分类器权重进行合理划分,以控制模型弱分类器数量。

1.2 遗传算法

1975 年,Holland 提出了一种基于种群并行搜索的遗传算法(Genetic algorithm, GA),该算法依据生物进化理论和遗传变异理论,模拟了进化学中的一些基本现象:遗传、突变、自然选择以及杂交等,通过适者生存的竞争选择,使种群不断向前进化,最终得到最优解。

遗传算法的主要步骤如下:

(1)初始化。设置最大进化代数 T,随机产生 M 个个体作为初始种群 $P(0)$。

(2)个体评价。用适应度函数 f 去评价 $p(0)$中的每个个体。

(3)遗传操作。(a)交叉。对选中的个体进行基因互换(有多种交叉算子);(b)变异。对选中的个体上的基因值作变动;(c)选择。将选择算子用于种群(目的是把较优的个体遗传到下一代),种群 $P(t)$经过以上的遗传操作后得到下一代种群 $p(t+1)$。

(4)终止。迭代次数达到 T 时算法终止。

遗传算法特点:(1)算法处理的是种群而非单个个体,即对空间中多个解进行评估,防止进入局部最优解;(2)不需要搜索空间的辅助信息,通过遗传操作并结合适应度函数来指导搜索,具有较好的普适性和扩展性;(3)算法处理的对象是编码而非参数本身,搜索不受优化函数连续性的约束。

2 基于 GA 约束的集成进化分类算法

本文提出的算法是在 Boosting 算法框架上结合进化算法来选择分类精度较高的弱分类器,并通过进化算法特有的遗传操作来生成新的弱分类器,这样就能很好地控制最终模型输出的弱分类器个数,在分类精度不降的情况下比 AdaBoost 算法具有更好的可读性。

2.1 初始化种群

下面将给出集成进化分类算法(ECSM)中的一些基本定义。

定义 1 基因 g 一个基因 g 对应于一个 Decision stumps。Stumps 是一种只有 3 个节点的两层决策树,其分类精度仅比随机猜测略好,常用于二分类问题。

定义 2 个体 s 一个个体 s 由多个基因 g 组成,$s=\{g_1,g_2,\cdots,g_n\}$,$n\in(1,5)$。

定义 3 种群 p 一个种群 p 由多个个体 s 组成,$p=\{i_1,i_2,\cdots,i_t\}$,$t\in(1,20)$。

定义 4 类标 针对本文的二分类问题,类标设置为 1 和 -1,当属性值大于分类值时,类标为 1,反之为 -1。

算法选择 Decision stumps 作为弱分类器,在初始化阶段,随机生成 N 个个体,每个个体有 T 个基因,并给每个 stumps 赋予一个随机的权值 α_t($\sum_{t=1}^{T}\alpha_t=1$)。stumps 中的分类属性、分类值以及类标签也

是随机生成。

为了使收敛速度提升,在初始化阶段也会对个体进行筛选,即对随机个体基因的类标正确性进行判定。例如图 2 中个体 s,对属性 5,4,8,7 根据其分类值 21,1,4,3.1 所得到的正确类标与随机个体中生成的随机类标 1,−1,1,1 进行比较,当初始个体中基因的类标正确率达到或超过 50%时才保留该个体。

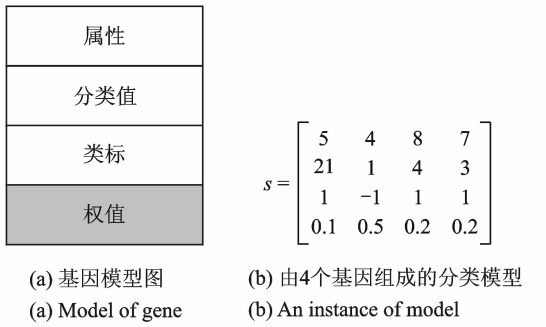

图 2 基因模型图和由 4 个基因组成的分类模型

Fig. 2 Model of gene and an instance of model

2.2 进化操作

在初始化阶段会随机生成 N 个个体,每个个体中含有 T 个基因。通过对当代的个体进行操作可得到下一代种群,主要的操作有以下几种:交叉、变异和选择。在最后一代中,具有最大适应度的个体会被选为最佳方案并输出用于分类操作[10]。

交叉操作是在同一代种群中,对不同的两个个体方案进行随机的交换基因,这样就可以在下一代种群中产生新的个体,实现种群的多样性,并有可能得到更好的个体方案。本文采用最简单的交叉操作—单点交叉。实现的过程是在个体中随机选择一个交叉点,并互换两个个体交叉点前后的基因,以此得到新的个体。如图 3 所示。

$$\text{交叉前}\quad s_1=\left[\begin{array}{cc|cc}5&4&8&7\\21&1&4&3.1\\1&-1&1&1\\0.1&0.5&0.2&0.2\end{array}\right]\qquad s_2=\left[\begin{array}{cc|cc}3&2&6&5\\27&1&4&15\\1&-1&-1&1\\0.2&0.45&0.2&0.15\end{array}\right]$$

$$\text{交叉后}\quad s_{1\text{new}}=\begin{bmatrix}5&4&6&5\\21&1&4&15\\1&-1&-1&1\\0.1&0.5&0.2&0.15\end{bmatrix}\qquad s_{2\text{new}}=\begin{bmatrix}3&2&8&7\\27&1&4&3.1\\1&-1&1&1\\0.2&0.45&0.2&0.2\end{bmatrix}$$

图 3 交叉操作

Fig. 3 Crossover evolution

变异操作是对种群中随机选择的个体进行操作,具体到本文中可有 4 种变异操作:属性变异、分类值变异、类标变异以及权值变异。在每一代的种群中,都会进行以上 4 种变异操作。其中属性变异是指随机选择一个属性代替原属性。分类值变异是指从当前属性的取值范围中随机选择一个属性值代替原有的分类值。类标的变异可由 1 突变为−1,或相反。权值的变异则是在(0,1)范围内生成一个新的权值代替原有的权值,最后所有权值还需标准化到总和为 1($\sum_{t=1}^{T}\alpha_t=1$)。标准化过程实际上是对其他基因增加或减少相应比例的权值。例如图 4 中的权值从 0.5 变异成 0.3,其中整体减少的权值是 0.2,为了归一化到总和为 1,此时需要把减少的 0.2 按照 1∶3∶2∶2 的比例分发到每一个基因。同理,变异后权值总体大于 1 的情况就是相应的按照一定比例去减少每个基因的权值。以上变异操作所选择的个体都具有随机性,这样才能给下一代种群提供更丰富的个体。

选择操作是对经过交叉跟变异操作的种群进行筛选,适应度高的个体可进入到下一代种群中[11],然后再随机生成剩余的个体以补足群体数量,这样的操作可以给下一代种群提供最好的基分类器。实

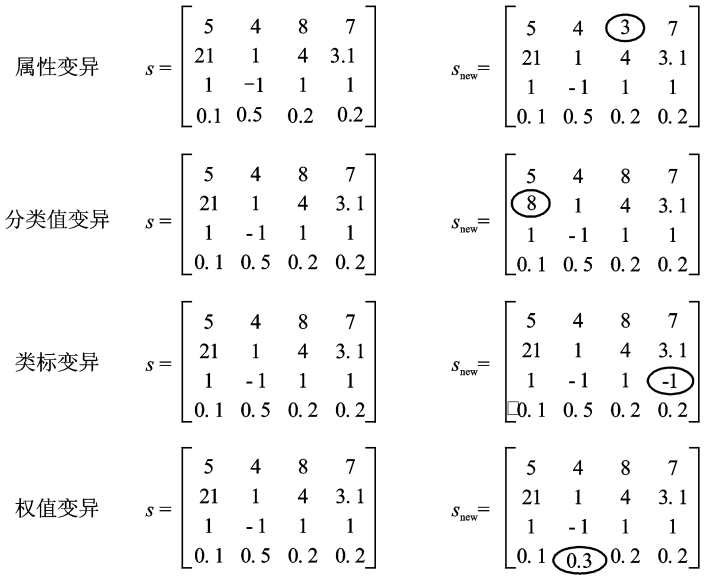

图 4 变异操作

Fig. 4 Mutation evolution

验中尝试过选择原始种群的 1/4,1/3,1/2 个个体遗传至下一代,发现选择 1/4 时能取得最好的效果,其原因在于保留的个体过多,算法搜索的空间就大,容易出现过度拟合的现象,陷于局部最优解。

2.3 适应度函数

对于种群中个体的适应度评价,本文用分类精准度。通过最简单的分类准确率直接评判一个分类器的性能,确保在算法迭代过程中,给下一代种群提供最好的基分类器。

$$f(c)=\frac{\sum_{i=1}^{n} m}{n} \tag{4}$$

式中:n 为数据集中的样本总数,如果样本 i 分类正确,则 $m=1$,否则 $m=0$。$f(c)$的值越大,说明分类器的分类效果越好。

2.4 算法描述

输入:输入样本集 $C\{(x_1,y_1),\cdots,(x_m,y_m)\}$,类标 $Y\{1,-1\}$,迭代次数。

输出:分类结果。

步骤:(1)初始化:随机生成 $N\times T$ 个基因;(2)遗传操作:交叉、变异和选择;(3)个体评价:用适应度函数去评价每一个个体;(4)保留 5 个优越个体至下一代;(5)新一代种群:随机生成 $N\times T-5$ 个基因;(6)终止:迭代次数达到 K 时算法终止;(7)最终分类模型。

3 实验分析

实验数据将采用 UCI 机器学习社区的 5 个真实数据集,它们分别是:Kyphosis,Blood Transfusion Service Center,Wisconsin Diagnostic Breast Cancer,Haberman's Survival 以及 Statlog (Shuttle)。如表 1 所示。

表 1 实验中的 5 个真实数据集

Tab. 1 Five real datasets in the proposed experiments

类别	Kyphosis	BTSC	WDBC	HS	Shuttle
数据集大小	81	748	569	306	14 500
数据属性个数	3	4	30	3	9
数据类标签	2	2	2	2	2

实验采用 WEKA 平台中的 AdaBoost 算法作为对比，WEKA 是由新西兰怀卡托大学推出的基于 JAVA 环境的机器学习软件，全称是 Waikato environment for knowledge analysis。此平台集成了大量数据挖掘算法，包括聚类、分类和回归等。

设置其迭代次数 I 分别为 500，1 000，1 500 以及 3 000(即最终模型弱分类器个数)。集成进化分类算法采用 Decision stumps 作为弱分类器，设置遗传进化次数为 500，1 000，1 500 以及 3 000，初始种群最大的个体数为 20，每个个体最大的基因数 max$T$ 为 5(即最终模型弱分类器个数的最大值)。为了得到更加准确的分类精度，实验采用十倍交叉法，能避免复杂数据分类时出现大的精度波动问题。表 2～6 是实验结果。

表 2 2 种算法分别在 Kyphosis 上经过 10 倍交叉检验的结果

Tab. 2 Result of tenfold cross validation in Kyphosis

迭代次数 I	ECSM		AdaBoost	
	正确率	弱分类器个数	正确率	弱分类器个数
500	0.804 2	3	0.777 7	500
1 000	0.785 8	3	0.802 4	1 000
1 500	0.816 8	3	0.814 8	150 0
3 000	0.811 4	3	0.790 1	3000

表 3 在 BTSC 上经过 10 倍交叉检验的结果

Tab. 3 Result of tenfold cross validation in BTSC

迭代次数 I	ECSM		AdaBoost	
	正确率	弱分类器个数	正确率	弱分类器个数
500	0.761 0	4	0.790 1	500
1 000	0.776 8	4	0.791 4	1 000
1 500	0.784 1	4	0.790 8	1 500
3 000	0.770 5	4	0.775 8	3 000

表 4 在 WDBC 上经过 10 倍交叉检验的结果

Tab. 4 Result of tenfold cross validation in WDBC

迭代次数 I	ECSM		AdaBoost	
	正确率	弱分类器个数	正确率	弱分类器个数
500	0.942 5	5	0.952 4	500
1 000	0.956 2	5	0.973 6	1 000
1 500	0.963 4	5	0.970 1	1 500
3 000	0.947 5	5	0.962 1	3 000

表 5 在 HS 上经过 10 倍交叉检验的结果

Tab. 5 Result of tenfold cross validation in HS

迭代次数 I	ECSM		AdaBoost	
	正确率	弱分类器个数	正确率	弱分类器个数
500	0.842 6	3	0.862 5	500
1 000	0.892 0	3	0.883 4	1 000
1 500	0.881 6	3	0.867 5	150 0
3 000	0.862 4	3	0.872 2	300 0

表 6 在 shuttle 上经过 10 倍交叉检验的结果

Tab. 6 Result of tenfold cross validation in shuttle

迭代次数 I	ECSM		AdaBoost	
	正确率	弱分类器个数	正确率	弱分类器个数
500	0.883 2	5	0.893 5	500
1 000	0.892 2	5	0.908 9	1 000
1 500	0.878 5	5	0.912 0	1 500
3 000	0.861 1	5	0.907 7	3 000

从表 2 中可以看出,ECSM 在 Kyphosis 数据集上分类精度在 78.58%～81.68%之间波动,AdaBoost 的精度在 77.77%～81.48%之间,在迭代次数是 1 500 时 ECSM 的分类精度最好,且其最终模型的分类器个数仅为 3 个。表 3 中 ECSM 在 BTSC 数据集上的分类精度为 76.1%～78.41%,AdaBoost 的精度在77.58%～79.14%之间,在迭代次数是 1 000 时 AdaBoost 的分类效果最好,但 ECSM 的弱分类器个数仅为 4 个。表 4 中 ECSM 在 WDBC 数据集上的分类精度为 94.25%～96.34%,而 AdaBoost 的分类效果为95.24%～97.36%,并且在迭代 1 000 次时 AdaBoost 分类效果最好,但 ECMS 的 5 个弱分类器却远远少于 AdaBoost。表 5 中 ECSM 在 1 000 次迭代时取得了比 AdaBoost 还要好的分类精度为 89.20%,且其波动范围为 84.26%～89.20%,而 AdaBoost 的波动范围是 86.25%～88.34%,在取得更好分类效果的同时仅仅使用了 3 个弱分类器。表 6 中使用的数据集实例个数超过了 10 000,ECSM 算法使用了 $\max T=5$ 个弱分类器个数,分类精度保持在 86.11%～89.22%之间,但传统的 AdaBoost 算法却能够取得更好的效果为 89.35%～91.20%,新算法的优势在于大大地缩减了模型的规模。

在 BTSC 和 WDBC 数据集上 ECSM 的分类精度只能接近 AdaBoost 算法,不能取得更好的效果,原因在于这两个数据集样本较多,数据离散化且波动幅度较大,说明 ECSM 算法对离散度大的数据集处理能力稍显不足,不过在模型的输出规模方面仍有很好的效果。从这 5 个实验结果还可以看出,算法并不会随着弱分类器个数的增加而取得更好的分类效果。

4 结束语

Boosting 作为最流行的分类算法框架之一,在很多不同的领域都有不错的表现。尤其是其不受限于分类器的类型,总能有效地把弱学习器提升为强学习器,所以具有很强的通用性。本文通过结合进化算法,在 AdaBoost 的迭代过程当中保留下物种的多样性,并合理地给分类器赋权值,得出了性能不错的集成进化分类算法。通过大量的实验表明,本文提出的集成进化算法既能有不错的分类效果,又能很好的控制模型规模。不过该算法对离散化且波动范围较大的数据处理能力仍需进一步的提高。对于算法的评价准则是下一步研究的重点,选择一个更为合理的评价函数,更有利于算法中分类器的选择。当然,对于算法效率方面也是应该重点关注的内容之一。

参考文献：

[1] Liu X, Matwin S. An ensemble method based on AdaBoost and meta-learning[J]. Advances in Artificial Intelligence, 2013, 78(4):278-285.

[2] Kim T K, Cipolla R. Multiple classifier boosting and tree-structured classifiers[J]. Machine Learning for Computer Vision, 2013,41(1):163-196.

[3] Zhu M. Kernels and ensembles: Perspectives on statistical learning[J]. Am Stat, 2008,62:97-109.

[4] Freund Y, Schapire R E. Experiments with a new boosting algorithm[J]. Machine Learning: Proceedings of the 13th conference, 1996,33(1):148-156.

[5] Cao Yingmiao, Qi Guang. Advance and prospects of AdaBoost algorithm[J]. Acta Automatica Sinica, 2013,39(6):745-758.

[6] Valiant L G. A theory of the learnable [M]. Communications of the ACM, 1984, 27(11):1134-1142.

[7] 廖红文,周德龙. AdaBoost 算法改进综述[J]. 计算机系统应用,2012,21(5):240-244.
Liao Hongwen, Zhou Denglong. Review of AdaBoost and its improvement[J]. Computer Systems & Applications,2012,21(5):240-244.

[8] Freund Y, Schapire R E. A decision-theoretic generalization of on-line learning an application to boosting [J]. Computer and System Sciences, 1997,55(1):119-139.

[9] She C, Li H. Boosting through optimization of margin distributions[J]. IEEE Trans on Neural Networks, 2010,21(4):659-666.

[10] 万宝吉,张涛,李文祥. 基于多分类器融合的未知嵌入率图像隐写分析方法[J]. 数据采集与处理,2014,29(5):749-756.
Wan Baoji, Zhang Tao, Li Wenxiang, et al. Multi classifer fusion based unknown image steganalysis[J]. Journal of Data Acquisition and Processing,2014,29(5):749-756.

[11] 刘奕晨,王毅,牛奕龙. 基于标准差的自适应激素调节遗传算法[J]. 数据采集与处理,2012,27(3):33-339.
Liu Yichen, Wang Yi, Niu Yilong, et al. An adaptive genetic algorithm based on hormone regulation[J]. Journal of Data Acquisition and Processing,2012,27(3):333-339.

作者简介：

宋文展(1988-)，男，硕士研究生，研究方向：数据挖掘，E-mail: songwenzhan @ foxmail. com。

元昌安(1964-)，男，教授，研究方向：智能计算、图像处理。

覃晓(1973-)，女，副教授，研究方向：智能计算、图像处理。

周凯(1991-)，男，硕士研究生，研究方向：图像处理。

郑彦(1993-)女，硕士研究生，研究方向：图像处理。