Journal of Computer Applications
计算机应用,2013,33(3):789 -792
ISSN 1001-9081
CODEN JYIIDU
2013-03-01
http://www.joca.cn

文章编号:1001 -9081(2013)03 -0789 -04
doi:10.3724/SP.J.1087.2013.00789

基于粒子群优化的不均衡数据学习

曹　鹏[1,2*],李　博[1,2],栗　伟[1,2],赵大哲[1,2]
(1.东北大学 信息科学与工程学院,沈阳 110004;　2.医学影像计算教育部重点实验室(东北大学),沈阳 110179)
(*通信作者电子邮箱 neusoftcp@gmail.com)

摘　要:为了提高重采样算法在不均衡数据学习的性能,提出一种基于粒子群优化的不均衡数据学习方法。通过粒子群优化,以不均衡数据分类评价准则作为目标函数,来优化重采样算法中最佳的采样率,同时对特征进行选择,从而达到最佳的数据分布。该算法在大量UCI数据集上进行了测试,与其他不均衡学习算法进行比较,结果表明该算法具有更高的分类性能;并验证了同时优化采样率和特征集合,可有效地改进不均衡数据分类效果。

关键词:粒子群优化;群体智能;不均衡数据分类;重采样;特征选择
中图分类号:TP391　**文献标志码**:A

Imbalanced data learning based on particle swarm optimization

CAO Peng[1,2*], LI Bo[1,2], LI Wei[1,2], ZHAO Dazhe[1,2]
(1. *College of Information Science and Engineering, Northeastern University, Shenyang Liaoning* 110004, *China*;
2. *Key Laboratory of Medical Image Computing of Ministry of Education* (*Northeastern University*), *Shenyang Liaoning* 110179, *China*)

Abstract: In order to improve the classification performance on the imbalanced data, a new Particle Swarm Optimization (PSO) based method was introduced. It optimized the re-sampling rate and selected the feature set simultaneously, with the imbalanced data evaluation metric as objective function through particle swarm optimization, so as to achieve the best data distribution. The proposed method was tested on a large number of UCI datasets and compared with the state-of-the-art methods. The experimental results show that the proposed method has substantial advantages over other methods; moreover, it proves that it can effectively improve the performance on the imbalanced data by optimizing the re-sampling rate and feature set simultaneously.

Key words: Particle Swarm Optimization (PSO); swarm intelligence; imbalanced data classification; re-sampling; feature selection

0　引言

不均衡数据分类的研究受到越来越多的重视[1-2]。现实生活中很多数据分布都是不均衡的,例如欺骗信用卡检测、网络入侵检测、医疗诊断等应用[3]。在不均衡数据分类时,各个类别的样本数目存在较大的差异,导致不同类别的样本对于训练算法提供的信息不对称;因为传统分类器都是基于准确率最大化来进行训练,所以常常忽略了数量小的类别信息,而在很多领域中往往少数类的识别率显得更为重要,最终影响传统分类器的分类结果。Weiss 等[4]提到不均衡数据的分布不是最优的,在分类器训练学习之前,需要根据不同的评价标准,来对样本空间进行修改。数据重采样是一种非常有效的解决不均衡数据分类的方法,它直接对训练集进行操作,改变训练集样本分布,降低不平衡程度,处理后的样本用来构建分类器。本文使用随机降采样与 SMOTE 升采样相结合的组合采样算法提升数据的均衡性,既能去除训练样本中噪声样本和重复信息,又可以扩大少数类潜在的样本空间。

虽然重采样算法[5-7]在一些数据集上取得了不错的效果,但是这类方法在解决不均衡数据分类上仍然存在一些缺陷:1)不能确定采样的最优比例。很多重采样算法都是使两类数量达到均衡,而这并不能保证最佳的数据分布,最佳的采样率会最大限度提高不均衡学习的准确度;2)采样算法只考虑了空间分布中样本的选择,没有考虑特征的选择,而特征选择对于不平衡分类问题同样具有重要意义[8-9]。只有解决以上两点不足,才可以达到最佳的不均衡数据分类效果。针对以上两点不足,本文亟待需要解决两个关键问题:1)如何对采样比例和特征选择进行评估;2)对于采样比例(连续值)和特征(离散值)这种混合类型变量,采样何种优化算法。

鉴于此,本文提出一种以最大化不均衡数据评价标准(*G-mean*,*AUC*)作为目标函数,使用粒子群优化算法,对不均衡数据采样的采样率和不均衡数据集中的特征集合同时进行评估和优化,选择出最佳的采样比例且同时获得最佳的特征子集提高不均衡数据学习能力。本文算法是一种封装(Wrapper)算法,可以使用多种分类算法作为基分类器,本文使用 C4.5 决策树算法进行验证。

1　算法描述

在本章中,首先介绍随机降采样和 SMOTE 升采样算法,以及粒子群优化算法的原理。之后,提出一种基于粒子群优化的算法,以 *AUC* 和 *G-mean* 分别作为目标函数,不断优化更新降采样和升采样的采样率以及特征子集,使训练集达到最优的数据分布。

收稿日期:2012-09-03;**修回日期**:2012-10-08。
基金项目:国家自然科学基金资助项目(61001047);中央高校基本科研业务费专项资金资助项目(N110618001)。
作者简介:曹鹏(1982 -),男,辽宁沈阳人,博士研究生,主要研究方向:机器学习、影像挖掘;　李博(1985 -),男,辽宁沈阳人,博士研究生,主要研究方向:影像检索与挖掘;　栗伟(1980 -),男,辽宁沈阳人,博士研究生,主要研究方向:文本挖掘;　赵大哲(1960 -),女,辽宁沈阳人,教授,主要研究方向:软件工程、数据挖掘、医学影像处理。

1.1 基于随机降采样与 SMOTE 升采样的组合采样算法

基于数据采样的技术是处理不平衡数据的重要且常用方法,目前主要有两种数据重采样技术,通过增加少数类(Minority Class)训练样本数的升采样和减少多数类(Majority Class)样本数的降采样,使不平衡的样本分布变得比较平衡,从而提高分类器对少数类的识别率。

由于在多数类样本中存在着噪声样本和大量的重复信息,这些冗余信息将会严重影响分类器的准确率,因此需要对这些冗余样本进行剔除并保留有效信息。随机降采样算法随机地选取一些多数类样本,将这些样本从多数类中移除,来调节原始数据集的平衡度。

传统的随机升采样是基于样本复制来增加样本数量,由于决策区间过小往往会引起过拟合,Chawla 等[10] 提出的 SMOTE 算法是一种简单有效的升采样方法,该方法首先为每个少数类样本随机选出几个邻近样本,并且在该样本与这些邻近的样本的连线上随机取点,生成无重复的新的少数类样本。

本文使用随机降采样和 SMOTE 升采样结合的采样方法,不仅可以降低多数类的数量,同时可以扩展少数类样本的潜在分布,从而达到训练集中数据的均衡。

1.2 粒子群优化算法

粒子群优化(Particle Swarm Optimization, PSO)算法是一种基于群体的演化算法[11],其优势在于算法的简洁性,易于实现,没有很多参数需要调整。很多实验已经证明粒子群的优化算法要优于遗传算法[12]。

在 PSO 中,每个优化问题的解都是搜索空间中的一只鸟。称之为“粒子(Particle)”。所有的粒子都有一个被优化的函数决定的适应值,每个粒子还有一个速度决定它们飞翔的方向和距离。然后粒子们就追随当前的最优粒子在解空间中搜索。PSO 初始化为一群随机粒子。然后通过迭代找到最优解。在每一次迭代中,粒子通过跟踪两个“极值”来更新自己。第一个就是粒子本身所找到的最优解。这个解叫作个体极值 *pbest*;另一个极值是整个种群目前找到的最优解,这个极值是全局极值 *gbest*。粒子的速度与位置更新公式如下:

$$v_i^{t+1} = w \times v_i^t + c_1 \times r_1 \times (pbest_i^t - x_i^t) + c_2 \times r_2 \times (gbest^t - x_i^t) \tag{1}$$

$$x_i^{t+1} = x_i^t + v_i^{t+1} \tag{2}$$

其中:$i = \{1,2,\cdots, SN\}$, SN 是粒子群的规模;t 是迭代次数;w 是惯性因子,代表粒子先前的速度或惯性,起着权衡局部最优能力和全局最优能力的作用;r_1 和 r_2 为 [0,1] 的随机数,用来保证粒子群的多样性;c_1 和 c_2 是学习因子,表示粒子本身的思考能力和社会行为能力。

1.3 基于粒子群优化的不均衡数据学习

采样算法通过改变两类样本的数量,使样本的分布更加均衡,更加利于分类器学习;但是在不均衡数据分类时,特征也对最终的分类有至关重要的作用,特征以及样本共同决定着数据分布。根据不平衡分类问题的特点,选取最具有区分能力的特征,有利于提高少数类的识别率,所以特征选择方法对于不平衡分类问题同样具有重要意义。而且特征的选择和采样也有密切的关系。特征选择不仅可以增加 SMOTE 升采样算法的效率,并且可以在新的特征子空间下使 SMOTE 升采样产生更准确的样本;而采样后数据样本的不同,导致选择的特征也不同。所以需要对样本和特征同时进行选择,才能达到最佳的数据分布。本文提出一种基于粒子群算法,对不均衡数据的采样率和特征同时优化,使 *AUC*/*G-mean* 最大化的优化算法(PSO-based Random under-sampling & SMOTE and Feature Selection algorithm, PSO-RSFS),来提升最终不均衡数据的分类性能。

粒子群中解的形式如图 1 所示,包括升采样和降采样的采样比例 *OsLevel* 和 *UsLevel*,以及特征向量 ***feature_vector***(f_1, f_2, …, f_n)。其中特征向量是二进制的形式,1 代表此特征选中,0 代表未选中。

OsLevel	*UsLevel*	f_1	f_2	…	f_{n-1}	f_n

图 1 PSO 中解的表达形式

由于粒子群算法是对连续值进行优化,为了使其同样可以对特征向量的二进制离散值进行优化,使用 sigmoid 函数,对式(1)中生成的速度 v 利用式(3)和(4),使连续值 v 转换为离散值 x,即 0 和 1,从而适用于特征集合的选择。

$$v_i^{t\prime} = sig(v_i^t) = \frac{1}{1 + e^{-v_i^t}} \tag{3}$$

$$x_i^{t+1} = \begin{cases} 1, & r_i < v_i^{t\prime} \\ 0, & \text{其他} \end{cases} \tag{4}$$

在粒子群优化过程中,需要设置适应度函数,这取决于分类器性能的评价标准。准确率是作为分类的主要评价标准,表示分类结果中所有分类正确的数量在总类别中所占的比率,不能正确反映不平衡数据集的分类性能。为此,针对不均衡数据,需选择更为合理的评价标准。

ROC(Receiver Operating Characteristic)曲线[13],能够全面地描述分类器在不同判决阈值时的性能,已成为不平衡数据分类器性能评价的准则。采用 ROC 曲线下面积 *AUC* 来对分类器的性能进行评估。

G-mean 是一种衡量数据整体分类性能的评价指标[14],定义如下:

$$\begin{cases} Sensitivity = \dfrac{TP}{TP + FN} \\ Specificity = \dfrac{TN}{TN + FP} \end{cases} \tag{5}$$

$$G\text{-}mean = \sqrt{Sensitivity * Specificity} \tag{6}$$

在粒子群优化中,分别采用 *AUC* 和 *G-mean* 作为适应度函数(即目标函数),来进行优化。为了防止过拟合,在优化过程中验证分类的结果时,把训练集划分成训练子集 *Trt*(70%)和验证子集 *Trv*(30%)。对训练子集 *Trt* 进行采样和特征选择,构造新的数据集合 *BalTrt*,并训练分类器,之后在验证子集 *Trv* 进行测试得到分类结果,作为当前解的适应度值。PSO-RSFS 算法描述如下:

输入 数据集 *Dataset*, *MCN*, *SN*, *metric* *M*(*AUC* 或 *G-mean*),采样步长 *step*。

输出 *AUC* 分类结果。

1) 划分数据集 *Dataset* 成训练集 *TrDataset* 和测试集 *TeDataset*;
2) 初始化每个解 $x_i(i = 1,2,\cdots, SN)$;
3) 全局最优解 $BestM = 0$;
   for $i = 1$ to *MCN*
      for $j = 1$ to *SN*
4) 获得当前解 x_j 中的升采样比例 *OsLevel*,降采样比例 *UsLevel* 和特征子集 ***feature_vector***
      for $k = 1$ to *NumFolds*
5) *TrDataset* 划分成训练子集 Trt_k 和验证子集 Trv_k
6) 根据 *OsLevel*, *UsLevel* 和 ***feature_vector*** 对 Trt_k 采样和特征选择操作,生成新的训练子集 $BalTrt_k$
7) 在 $BalTrt_k$ 构建分类器 C_k,并在验证集 Trv_k 进行测试,得到 M_k

end for

8) $M_j = average(M_k)$,即取 $NumFolds$ 个 M_k 的平均值作为对应当前解 x_j 在第 i 轮的适应度值

end for

9) 根据式(1)和(2)及 $step$ 更新粒子群中每个解的位置,及更新全局最优解 $BestM$

end for

10) 得到 $BestM$ 对应的 $OsLevel$,$UsLevel$ 和 $\boldsymbol{feature\_vector}$。对训练集 $TrDataset$ 进行组合采样和特征选择操作,生成新的训练子集 $BalTrDataset$,构建分类器 C

11) 用分类器 C 在测试集 $TeDataset$ 进行测试,计算最终的 AUC 分类结果

2　实验评估

为评估算法的性能,本文选择 10 组具有不同实际应用背景的 UCI 数据集。数据集详细信息见表 1。

表 1　实验数据集描述

数据集	样本数	特征维数	类比例/%
Abalone(18)	731	7	6
Glass(5,6,7)	214	9	24
German(1)	1 000	20	30
Pima(1)	768	8	35
Spambase(2)	4 601	57	40
Waveform(0)	5 000	40	33
Vehicle(1)	940	18	23
Letter(4)	20 000	16	4
Page(2,3,4,5)	5 473	10	10
Yeast(1)	1 484	9	16

注:数据集名称后的(n)中的数字 n 代表少数类的类别。

在本文算法中,需要设置 4 个参数,升采样步长设置为 5%,降采样步长设为 2%。本文根据文献[15],对于粒子群算法的参数设置为 $C_1 = 2.8$, $C_2 = 1.3$, $w = 0.5$。为了提高搜索效率,对于粒子群中的迭代次数 MCN 和粒子的个数 SN,不采用固定值,而根据优化变量的个数进行动态设置:$SN = 1.5 \times$ 需要优化的参数个数,$MCN = 10 \times SN$,优化参数包括两个采样比例和特征集合。

本实验为了客观对比多个不均衡数据分类方法,以下所有实验都采用 10 折交叉验证。对应粒子群优化中的每一个解(采样率和特征子集),在计算其适应度时,同样为了保证基于采样率和特征子集进行分类的客观准确性,在每次验证时使用 3 折交叉验证($NumFolds = 3$),取 3 次的均值作为当前解的适应度。

表 2 列出了 PSO-RSFS 算法根据不同指标(AUC 和 $G\text{-}mean$)进行优化的结果:两种采样的采样率和特征子集个数。由于目标函数不同,所以优化的侧重方面也不一样,得到的采样率和特征子集均不相同。该算法和在不均衡数据分类中常用的方法进行比较:随机降采样(RUS)、SMOTE 升采样、SMOTEBoost 组合升采样算法[16]和代价敏感组合分类方法 MetaCost[17]。其中三种采样算法的采样标准都是使两类的数量均等,MetaCost 算法的错分代价参数根据两类样本数量比例进行设置。从表 3 中的比较结果可知,PSO-RSFS 算法无论采样哪种目标函数,都提高了传统方法的性能。除了数据集 Vehicle 之外,PSO-RSFS 在其余 9 个数据集都达到了最高值,说明 PSO-RSFS 算法在不均衡数据分类上比其他常用的算法更优越。同时可以观察到,以 AUC 作为目标函数进行优化可获得更高的 AUC 值,这是因为首先 AUC 具有更强的评价能力,以 AUC 作为目标函数可以提升分类的泛化能力,其次是由于优化和验证的指标一致性,所以获得了更高的分类性能。通过 PSO-RSFS 优化方法在不均衡数据集下训练分类算法,得到了最优的特征子集和正负样本比例,改善了数据分布,从而提高了分类识别能力,也验证了同时优化采样比例和特征集合在不均衡分类中的重要性。

表 2　PSO-RSFS 算法的优化结果

数据集	评测指标	$UsLevel$	$OsLevel$	$Feature\ subset\ size$
Abalone	AUC	20	315	5
	$G\text{-}mean$	32	285	6
Glass	AUC	8	260	5
	$G\text{-}mean$	10	245	7
German	AUC	12	295	9
	$G\text{-}mean$	10	280	9
Pima	AUC	8	220	7
	$G\text{-}mean$	8	210	6
Spambase	AUC	14	170	32
	$G\text{-}mean$	12	195	29
Waveform	AUC	6	320	16
	$G\text{-}mean$	12	275	21
Vehicle	AUC	12	200	11
	$G\text{-}mean$	16	225	11
Letter	AUC	24	645	11
	$G\text{-}mean$	30	585	10
Page	AUC	18	445	7
	$G\text{-}mean$	26	400	7
Yeast	AUC	8	210	4
	$G\text{-}mean$	8	225	5

表 3　分类方法的对比结果(AUC)

数据集	RUS	SMOTE	SMOTE Boost	Meta Cost	PSO-RSFS	
					$G\text{-}mean$	AUC
Abalone	0.758	0.784	0.799	0.795	0.809	0.822
Glass	0.885	0.903	0.912	0.893	0.927	0.945
German	0.843	0.856	0.863	0.864	0.881	0.889
Pima	0.804	0.826	0.826	0.834	0.855	0.861
Spambase	0.814	0.825	0.818	0.823	0.876	0.855
Waveform	0.778	0.784	0.785	0.788	0.795	0.828
Vehicle	0.691	0.692	0.691	0.741	0.695	0.735
Letter	0.922	0.933	0.933	0.928	0.933	0.935
Page	0.915	0.944	0.989	0.950	0.976	0.989
Yeast	0.821	0.849	0.849	0.855	0.874	0.882
平均	0.823	0.840	0.846	0.847	0.862	0.874

3　结语

不均衡数据在实际应用中广泛存在,如何有效处理不均衡数据也成为目前的一个新的研究热点。本文提出了一种结合采样和特征选择的不均衡数据分类算法,同时使分别用 AUC 和 $G\text{-}mean$ 评价指标作为不均衡学习的目标,不仅可以自动获得最佳的采样比例和特征子集,而且提升了不均衡数据的分类性能。由于本文算法是一种封装(Wrapper)算法,可以根据需要使用不同的目标函数,如更侧重于少数类识别性能的 F-measure。也可以根据性能要求选择不同的分类算法。最后使用 UCI 数据集对方法进行了验证,并与其他不均衡学习算法进行比较,结果证明了本文方法提升了分类器在不均衡数据的性能,也验证了采样率的优化和特征的选择在不均衡分类中的重要性,且需要同时对两者进行优化。由于本文

算法只针对二类类别的分类，下一步会对算法扩展到对多类类别的分类。

参考文献：

[1] 叶志飞，文益民，吕宝粮. 不平衡分类问题研究综述[J]. 智能系统学报，2009，4(2)：148－156.

[2] YANG Q, WU X. 10 challenging problems in data mining research [J]. International Journal of Information Technology & Decision Making, 2006, 5(4): 597－604.

[3] HE H B, GARCIA E A. Learning from imbalanced data [J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9): 1263－1284.

[4] WEISS G M, PROVOST F. Learning when training data are costly: the effect of class distribution on tree induction [J]. Journal of Artificial Intelligence Research, 2003, 19(1): 315－354.

[5] CHEN S, HE H B, GARCIA E A. RAMOboost: ranked minority oversampling in boosting [J]. IEEE Transactions on Neural Networks, 2010, 21(10): 1624－1642.

[6] RAMENTOL E, CABALLERO Y, BELLO R, *et al*. SMOTE－RSB*: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory [J]. Knowledge and Information Systems, 2012, 33(2): 245－265.

[7] 许丹丹，王勇，蔡立军. 面向不均衡数据集的 ISMOTE 算法 [J]. 计算机应用，2011，31(9)：2399－2401.

[8] WASIKOWSKI M, CHEN X W. Combating the small sample class imbalance problem using feature selection [J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10): 1388－1400.

[9] ZHENG Z H, WU X Y, SRIHARI R. Feature selection for text categorization on imbalanced data [J]. ACM SIGKDD Explorations Newsletter — Special Issue on Learning from Imbalanced Datasets, 2004, 6(1): 80－89.

[10] CHAWLA N V, BOWYER K W, HALL L O, *et al*. SMOTE: synthetic minority over-sampling technique [J]. Journal of Artificial Intelligence Research, 2002, 16: 321－357.

[11] KENNEDY J, EBERHART R C. Particle swarm optimization [C]// Proceedings of IEEE International Conference on Neural Networks. Piscataway, NJ: IEEE Press, 1995, 4: 1942－1948.

[12] HASSAN R, COHANIM R, de WECK O. A comparison of particle swarm optimization and the genetic algorithm [C]// Proceedings of the 46th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference. [S. l.]: AIAA, 2005: 1－13.

[13] FAWCETT T. An introduction to ROC analysis [J]. Pattern Recognition Letters, 2006, 27(8): 861－874.

[14] THAI-NGHE N, GANTNER Z, SCHMIDT-THIEME L. Cost-sensitive learning methods for imbalanced data [C]// Proceedings of 2010 International Joint Conference on Neural Networks. Piscataway, NJ: IEEE Press, 2010: 1－8.

[15] CARLISLE A, DOZIER G. An off-the-shelf PSO [C]// Proceedings of the Particle Swarm Optimization Workshop. Indianapolis: [s. n.], 2001: 1－6.

[16] CHAWLA N V, LAZAREVIC A, HALL L O, *et al*. SMOTEBoost: improving prediction of the minority class in boosting [C]// PKDD 2003: Proceedings of the Seventh European Conference on Principles and Practice of Knowledge Discovery in Databases, LNCS 2838. Berlin: Springer-Verlag, 2003: 107－119.

[17] DOMINGOS P. MetaCost: a general method for making classifiers cost-sensitive [C]// KDD '99: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 1999: 155－164.

（上接第 747 页）

[7] CHANG E C, XU J. Remote integrity check with dishonest storage server[C]// ESORICS '08: Proceedings of the 13th European Symposium on Research in Computer Security. Berlin: Springer, 2008: 223－237.

[8] SHACHAM H, WATERS B. Compact proofs of retrievability[C]// ASIACRYPT '08: Proceedings of the 14th International Conference on the Theory and Application of Cryptology and Information Security. Berlin: Springer, 2008: 90－107.

[9] WANG C, WANG Q, REN K, *et al*. Privacy-preserving public auditing for data storage security in cloud computing[C]// INFOCOM '10: Proceedings of the 29th Conference on Computer Communications. Piscataway, NJ: IEEE Press, 2010: 1－9.

[10] GOHEL M, GOHIL B. A new data integrity checking protocol with public verifiability in cloud storage[C]// IFIPTM '2012: Trust Management VI, IFIP Advances in Information and Communication Technology, LNCS 374. Berlin: Springer, 2012: 240－246.

[11] HAO Z, ZHONG S, YU N H. A privacy-preserving remote data integrity checking protocol with data dynamics and public verifiability[J]. IEEE Transactions on Knowledge and Data Engineering, 2011, 23(9): 1432－1437.

[12] WANG Q, WANG C, LI J, *et al*. Enabling public verifiability and data dynamics for storage security in cloud computing[C]// ESORICS '09: Proceedings of the 14th European Symposium on Research in Computer Security. Berlin: Springer, 2009: 355－370.

[13] QIAN W, CONG W, KUI R, *et al*. Enabling public auditability and data dynamics for storage security in cloud computing[J]. IEEE Transactions on Parallel and Distributed Systems, 2011, 22(5): 847－859.

[14] WANG Q, WANG C, KUI R, *et al*. Toward secure and dependable storage services in cloud computing[J]. IEEE Transactions on Services Computing, 2012, 5(2): 220－232.

[15] 曹天杰，张永平，汪楚娇. 安全协议[M]. 北京：北京邮电大学出版社，2009. 129.

[16] BAO F, DENG R, ZHU H. Variations of Diffie-Hellman problem [C]// ICICS '03: Proceedings of the 5th International Conference on Information and Communications Security. Berlin: Springer, 2003: 301－312.

[17] ERWAY C, ALPTEKIN K, PAPAMANTHOU C, *et al*. Dynamic provable data possession [C]// CCS '09: Proceedings of the 16th ACM Conference on Computer and Communications Security. New York: ACM Press, 2009: 213－222.

[18] BOWERS K D, JUELS A, OPREA A. Proofs of retrievability: theory and implementation[C]// CCSW '09: Proceedings of 2009 ACM Workshop on Cloud Computing Security. New York: ACM Press, 2009: 43－53.

[19] BONEH D, LYNN B, SHACHAM H. Short signatures from the Weil pairing[C]// ASIACRYPT '01: Proceedings of the 7th International Conference on the Theory and Application of Cryptology and Information Security. Berlin: Springer, 2001: 514－532.

[20] BONEH D, GENTRY C. Aggregate and verifiably encrypted signatures from bilinear maps[C]// Eurocrypt '03: Proceedings of the International Conference on the Theory and Applications of Cryptographic Techniques. Berlin: Springer, 2003: 416－432.