决策树C4.5算法改进与应用

陈 杰,邬春学

(上海理工大学 光电信息与计算机工程学院,上海 200093)

摘 要:针对决策树算法C4.5在处理数据挖掘分类问题中出现的算法低效以及过拟合问题,提出一种改进的TM-C4.5算法。该算法主要改进了C4.5算法的分支和剪枝策略。首先,将升序排序后的属性按照边界定理,得出分割类别可能分布的切点,比较各点的信息增益和通过贝叶斯分类器得到的概率,使用条件判断确定最佳分割阈值;其次,使用简化的CCP(Cost-Complexity Pruning)方法和评价标准,对已生成决策树的子树根节点计算其表面误差率增益值和S值,从而判断是否删除决策树节点和分支。实验结果表明,用该算法生成的决策树进行分类更为精确、合理,表明TM-C4.5算法有效。

关键词:C4.5;TM-C4.5算法;CCP;贝叶斯分类器;剪枝策略;评价标准

DOI:10.11907/rjdk.181302

中图分类号:TP312 **文献标识码**:A **文章编号**:1672-7800(2018)010-0088-05

Improvement and Application of Decision Tree C4.5 Algorithm

CHEN Jie, WU Chun-xue

(*School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai* 200093, *China*)

Abstract: Aiming at the inefficiency and over-fitting problem of decision tree algorithm C4.5 in the classification of data mining problems, an improved TM-C4.5 algorithm is proposed. The algorithm mainly improves the branching and pruning strategy of C4.5 algorithm. First, the ascending ordered attribute values are combined with the boundary theorem to get the cut points of the possible segmentation classifications. The information gain rate of each point and the probability obtained by the Bayesian classifier are compared, and the optimal segmentation threshold is determined according to the rules. Secondly, the simplified algorithm of CCP (Cost-Complexity Pruning) and evaluation criteria were used to calculate the surface error rate gain and S value of the subtree root node of the generated decision tree to judge whether to delete the decision tree node and branch. The analysis of the experimental results shows that the classification of the decision tree made by this algorithm is more accurate and reasonable, indicating the validity of TM-C4.5 algorithm.

Key Words: C4.5; TM-C4.5 algorithm; CCP; Bayesian classifier; pruning strategy; evaluation standard

0 引言

分类技术是数据挖掘领域中一种非常重要的研究方法[1]。目前解决分类问题应用较广泛的算法有决策树算法、贝叶斯分类、人工神经网络算法、K邻近算法、支持向量机算法等,而决策树学习是以实例为基础的归纳学习算法[2],在数据挖掘领域属于一种常用、简单有效的快速分类算法[3]。

本文对决策树C4.5算法的分支和剪枝策略进行改进,旨在获得更加高效、明确以及合理的分类效果。为提高算法的计算效率,文献[4]提出利用等价无穷小性质改变C4.5算法中的熵、信息增益和信息增益率[4],虽然计算过程中减少了对数运算函数的调用次数,但由于忽略了常量值的计算,使得误差值变大,导致分类结果准确率下降。针对缺失属性值导致分类准确率下降的问题,文献[5]提

收稿日期:2018-03-09

基金项目:上海市科学计划项目(16111107502, 17511107203)

作者简介:陈杰(1992-),男,上海理工大学光电信息与计算机工程学院硕士研究生,研究方向为数据分析与挖掘;邬春学(1964-),男,博士,上海理工大学光电信息与计算机工程学院教授、硕士生导师,研究方向为嵌入式系统及应用、计算机控制技术及工程。

出在决策树生成过程中，当分支的子集中属性值未知时，返回叶子节点，标记为 unknown，并在之后的剪枝中将比例超过 1/3（unknown 节点与叶子节点之比）的 unknown 节点删除[5]。与传统的 C4.5 算法相比，该算法的时间复杂度在属性缺失率较高时能提高很多，但在数据集缺失率较低甚至没有缺失率情况下，该算法的时间复杂度相比传统的 C4.5 算法没有明显提升。另外，该算法对于属性缺失率阈值的设置缺乏合理的计算方法。文献[6]提出在实验中加入风险评估机制，并增添了覆盖率和高风险率作为该机制的评价标准，通过将其与朴素贝叶斯和决策树的分类结果对比，得出决策树的覆盖率优于另外两种分类器的结论[6]。

综合上述学者的研究成果，为了获取更加高效和精确的决策树模型，本文结合边界定理和贝叶斯分类器，获得更为精确的分割切点，使用简化的 CCP 方法和评价标准对决策树进行修剪以提高分类效率。

1　决策树 C4.5 算法

ID3 算法和 C4.5 都是由 QUINLAN 提出的，后者是基于前者的改进算法。ID3 算法虽然可以有效处理离散型属性，但对于连续型属性却难以处理。连续型属性的处理是数据挖掘领域热门问题之一，影响着数据挖掘算法的准确性、复杂性和可理解性[7]。同时，ID3 算法容易产生过度拟合现象[8]。C4.5 算法解决了 ID3 算法产生过度拟合的缺点，并增加了一些功能，如未知属性的处理、连续属性的离散化和规则的产生等[9]，具体改进如下：①在 ID3 的基础上，为避免用信息增益 Gain 选择属性时偏向选择取值多的属性之不足，C4.5 选择用信息增益率 GainRatio 选择属性作为树的节点；②对构造的树进行剪枝，防止过度拟合；③完成对连续属性的离散化处理；④对不完整数据进行处理。

C4.5 算法流程见图 1。

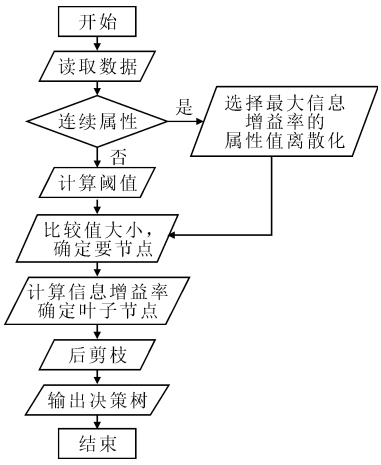

图 1　算法流程

C4.5 算法步骤如下[10]：①创建树的根节点 R；②假设数据的训练集 S 中的全部训练样本皆属于同一类别 C，则将 R 设为叶子节点，并且为叶子节点添加类别标记 C；③假设训练集 S 中的样本数量少于给定阈值，或者样本的全部属性均完成测试，则将节点 R 作为叶子节点，将 R 所属的类标记为训练集中频数最高的类别；④计算属性列表中每个属性的信息增益率 GainRatio；⑤假设属性 N 为属性列表中信息增益率最高的属性；⑥假设属性为连续性属性，将该属性进行离散化处理；⑦对每个由节点 R 产生新的叶子节点，假设该叶子节点的样本子集为空集，则将此节点标记为叶子节点，并且该节点的类别为训练集频数最高的类别，反之该叶子节点继续进行分裂；⑧采取悲观剪枝策略，对生成的决策树进行剪枝。

相关公式如下：

(1)信息增益率。

$$GainRatio(S,A) = Gain(S,A)/SplitInfo(S,A) \quad (1)$$

其中 A 表示样例的属性值集合。

(2)分裂信息。

$$SplitInfo(S,A) = -\sum_{i\in Value(A)}(|S_i|/|S|)*\log_2(|S_i|/|S|) \quad (2)$$

其中，i 为属性 A 所有可能的不同取值，$|S_i|$ 为分割的样本子集中取值为 i 时的样例数量，$|S|$ 为分割的样本子集总样例数量。

(3)信息增益。

$$Gain(S,A) = Entropy(S) - \sum_{v\in Values(A)}(|S_v|/|S|)*Entropy(S_v) \quad (3)$$

其中，$Values(A)$ 表示属性 A 可取值的集合，$Entropy(S_v)$ 是属性 A 值为 v 的熵值大小，$|S|$ 和 $|S_v|$ 分别表示对应样例集合中样例的数目。

(4)熵。

$$Entropy(S) = -p_+\log_2 p_+ - p_-\log_2 p_- \quad (4)$$

其中，p_+、p_- 表示样例集中类别为正、反例的比例，定义 $0\log0=0$。

后剪枝方法 CCP，又叫代价复杂度剪枝，是选择节点表面误差率增益值最小的非叶子节点。删除该节点的左右子树，在有多个节点表面误差率增益值相同的情况下，选择子节点数多的进行剪枝。

(5)表面误差率增益值公式。

$$\alpha = (R(t) - R(T))/(N(T) - 1) \quad (5)$$

其中 $R(t)=r(t)*p(t)$，$R(t)$ 表示叶子节点的误差代价，$r(t)$ 为节点的错误率，$p(t)$ 为节点数据量的占比。

$$R(T) = \sum_{i}^{m} r_i(t)*p_i(t) \quad (6)$$

其中，$R(T)$ 表示子树的误差代价，$r_i(t)$ 为子节点 i 的错误率，$p_i(t)$ 表示节点 i 的数据节点占比，$N(T)$ 表示子树节点个数。

2　TM－C4.5 算法

C4.5 在对连续属性进行处理时，假设训练集 D 有 N

个属性，第 $i(1\leqslant i\leqslant N)$ 个属性的值有 M 个，则该算法是将 M 个值都作为候选分割点，然后分别计算它们的信息增益，选择最大的作为最佳分割点。在这个过程中，计算每个候选分割点时都需要遍历整个数据集，找出大于和小于该点的记录数[11]。本文结合医学上糖尿病患者识别研究问题，利用基于 C4.5 的改进算法 TM－C4.5 对二型糖尿病进行判别。

二型糖尿病训练集属性大多是连续属性，使用原算法会极大降低算法效率。因此，本文提出结合边界定理和贝叶斯分类器的改进 C4.5 算法——TM-C4.5，对连续属性值的分割点选择进行改进。

边界定理：不管用于训练的数据集有多大，维度有多高，也不管数据集中有多少类，其类别的分布如何，连续型属性的最佳划分点总是在边界点处。

贝叶斯分类器：是各种分类器中分类错误概率最小的分类器，其设计方法是一种最基本的统计分类方法。贝叶斯公式如下：

$$P(C_i \mid A)=\frac{P(A \mid C_i)P(C_i)}{\sum_{j=1}^{n}P(A \mid C_j)P(C_j)} \tag{7}$$

其中，$P(C_i|A)$ 是属性为 A 时类 C_i 发生的概率，$P(A|C_i)$ 是类为 C_i 时 A 发生的概率。

连续属性离散化改进步骤如下：①将训练集每个连续属性 $A_i(i=1,2\cdots)$ 的值分别按升序排序；②获取类为 C_1 时 A_i 的最小属性值；③获取类为 C_2 时 A_i 的最大属性值；④基于每个类 $C_j(j=1,2)$ 的最大、最小值计算出切点值；⑤计算切点值的信息增益，选择最大值作为候选分割点；⑥计算切点值概率，选择最大候选作为分割点；⑦若信息增益和概率都最大则为同一个分割点，重复步骤⑤、步骤⑥、步骤⑦。

C4.5 算法生成的决策树模型复杂度过大，容易产生过拟合现象，决策树生成规则难以理解，算法效率低，因此多种决策树剪枝方式应运而生[12]。常用剪枝方法有预剪枝和后剪枝两种。预剪枝指在树到达一定高度停止生长，而此时节点成为叶子节点或到达某节点的实例数目小于设定阈值时停止生长。预剪枝带来的问题是过早地使节点成为叶子节点，导致数据集的某些属性可能会丢失。而后剪枝则不同，它是等决策树完全长成后剪去部分子树并使用树叶代替。因此，本文采用后剪枝方法中的 CCP (Cost－Complexity Pruning)方法，先对其进行简化，而后通过增加评价标准对单一的 CCP 方法进行补充，以免修剪子树过于粗糙。

对 CCP 计算公式进行简化：

$$\beta=\frac{R_j-L_j}{D}/(M_i-1) \tag{8}$$

其中，R_j 为子树根节点的错误样例数，L_j 为各叶子节点的错误样例数之和，D 为总数据量，M_i 为子树叶子节点个数。

引入评价标准：

$$S=w_1a(t)+w_2c(t) \tag{9}$$

其中，w_1、w_2 为权重，满足 $w_1+w_2=1$，通常情况下，定义两者的值都为 1/2；$a(t)$ 为待剪切子树各叶子节点的分类准确率，$c(t)$ 为待剪切子树各叶子节点的实例数量占实例总数之比。S 值越大，表示待剪切的子树越优。在剪切过程中，挑选待剪切子树时，选择孩子节点为叶子节点的节点进行剪切，以避免过度剪切。计算每个待剪切子树的 S 值，将选择同时满足最小表面误差率增益值和最小 S 值的节点替代成叶子节点，并赋值该节点的属性值为该类别的最高频数。

TM-C4.5 模型流程如图 2 所示。

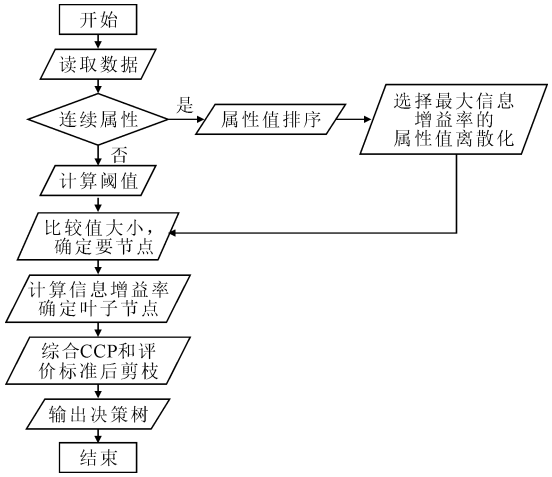

图 2 TM-C4.5 算法流程

3 实验结果

3.1 数据预处理

实验系统环境为：Windows7；编程工具：PyCharm；编程语言：python。实验数据来自 UCI 数据集，一共 768 条记录，数据集划分为 11 组，将其中 10 组 600 条记录用作训练集。采用随机抽取方法抽取训练集；抽取训练集之后，将实验室数据集剩余的 168 条记录作为测试集数据。实验数据集样本包含 9 种属性，分别为：①number of time pregnant(怀孕次数)；②Plasma glucose concentration a 2 hours in an oral glucose tolerance test(口服葡萄糖耐量试验中血浆葡萄糖浓度为 2 小时)；③Diastolic blood pressure (mm Hg)(舒张压)；④Triceps skin fold thickness (mm) 三头肌皮褶皱厚度；⑤2－Hour serum insulin (mu U/ml) (2 小时血清胰岛素)；⑥Body mass index(BMI)；⑦Diabetes pedigree function(糖尿病谱系功能)；⑧Age (年龄)；⑨类 Class(0 or 1)。

其中，将怀孕次数在 0～1 之间标记为 Low，将 2～5 之间标记为 Medium，大于等于 6 则标记为 High[13]；属性 BMI(体质指数)＝体重(kg)/身高(m)。根据 WHO 标准划分等级为：BMI 数值小于 18.5 设为偏瘦，18.5～24.9 之间设为正常，大于等于 25 设为超重。由于实验数据采

样范围过于宽泛，并不能代表某一个国家或地区的典型划分依据，因此在本实验数据集基础上，通过计算属性 BMI 的信息增益率获取一个节点阈值，将小于该值的属性值标记为 Light，大于该值的则为 Overweight。属性 AGE 的处理与属性 BMI 的计算方式相似，其值分别标记为 Young、Elderly。为使生成的决策树简明，方便后续算法进行，将数据集属性名分别缩写为 NTP、PG、DBP、TSF、HSI、BMI、DPF、AGE。根据统计得出各属性缺失值数据及所占比例，将数据以图表展示，分别如图 3 和图 4 所示。

从图 3 和图 4 可以看出，属性 NTP 的缺失值记录数比例虽然达到了 14.5%，但由于属性的现实参考意义很大，因此不能作为缺失值处理；属性 TSF 和 HSI 的缺失值分别达到了 227 个和 374 个，占据了总记录数的 29.6%和 48.7%，将该属性删除而不是删除包含这些缺失值的记录，以得到更多记录值。

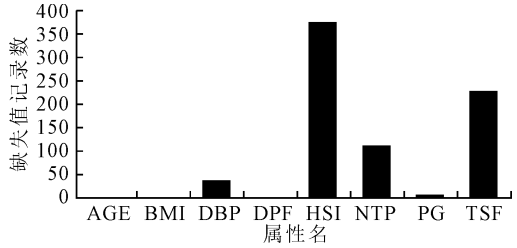

图 3　各属性缺失值记录数

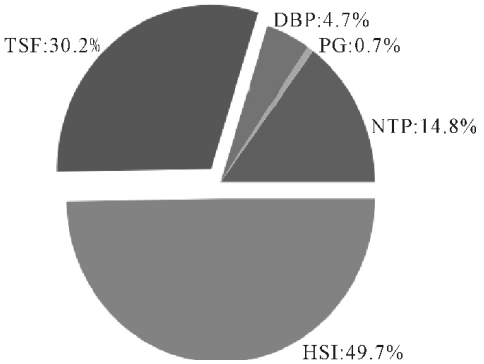

图 4　各属性缺失值所占比例

通过以上分析，获得被处理分析的数据集属性取值如表 1 所示。

表 1　数据集属性 ID 及取值

属性名	属性 ID	取值
Number of time pregnant	NTP	Low，Medium，High
Plasma glucose	PG	continuous
Diastolic blood pressure	DBP	continuous
Body mass index	BMI	continuous(Light，Overweight)
Diabetes pedigree function	DPF	continuous
Age	AGE	continuous(Young，Elderly)
Class		0，1

从表 1 可以看出，除了 Class 之外的 6 个属性中，有 5 个都是连续属性。在数据量较大以及连续属性过多时，若不对离散化方法进行改进，则算法效率必受到较大影响。

3.2　实验结果

通过结合边界定理和贝叶斯分类器得到的属性各切点的信息增益和概率值如表 2 所示。

表 2　属性各切点的信息增益值和概率值

属性名	切点值	信息增益	概率
PG	70.5	0.009 26	0.176
	97	0.072 56	0.231
	145.5	0.118 25	0.403
	194	0.015 61	0.228
	196	0.006 12	0.144
DBP	27	0.000 83	0.221
	30	$1.349\ 65e^{-6}$	0.362
	72	0.012 64	0.437
	114	0.000 83	0.413
	118	0.000 83	0.257
DPF	0.083	0.000 83	0.095
	0.088	0.000 94	0.216
	1.209	0.007 35	0.394
	2.329	0.002 12	0.197
	2.375	0.002 12	0.132

从表 2 中选择具有最大信息增益和最大概率值的切点值作为最佳分割阈值，由此得到最佳分割阈值，如表 3 所示。

表 3　各连续属性最佳分割阈值

属性 ID	最佳分割阈值	单位
PG	8.08mmol/l	mmol/L
DBP	72	mmHg
BMI	27.5	Kg/m^2
DPF	1.209	
AGE	37	

从表 3 可以看出，PG 的最佳分割阈值为 145.5mg/dl (8.08mmol/l)，这与目前医学上口服葡萄糖耐量实验中 2 小时的正常值水平 7.8mmol/l 很接近，反映了对连续属性离散化的改进是有效的。其余属性如 DBP、BMI、DPF、AGE 的最佳分割阈值分别为 72mmHg、27.5kg/m^2、1.208 5和 37。

此外，改进后的方法与原方法的执行时间分别为 855ms 和 937ms，时间缩短了 8.75%。由于数据集的数据量有限，因而效率提高不大，但相比原算法已经有了明显提升。

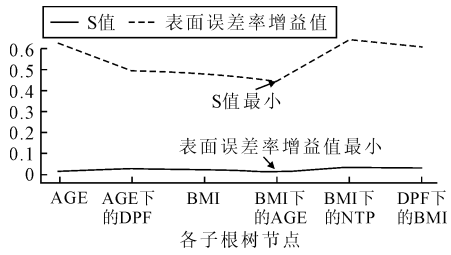

图 5　各子树根节点的表面误差率增益值和 S 值

图 5 是属性 AGE 和 BMI 下的各节点表面误差率增益值和 S 值情况，通过改进方法选择删除两者值都较小的节点。

由图 5 可以看出，BMI 下的子树 AGE 的表面误差率

增益值为 0.014 3，为最小值，其 S 值 0.442 也最小，表明该节点的分类精度和覆盖率都比较差，因此删除该节点改为用叶子节点替代。若没有引入评价标准 S 作为 CCP 方法的补充，在表面误差率增益值相差千分之几的情况下就很难抉择应该删除哪个节点。

表 4 准确率实验结果

测试集	实例数目/个	C4.5 准确率(%)	TM－C4.5 准确率(%)
T_1	100	75.83	79.36
T_2	200	81.63	83.12
T_3	300	83.57	87.94
T_4	400	85.41	90.25
T_5	500	84.79	91.34

表 4 所示为当测试集 $T_i(i=1,2,3,4,5)$的实例数目分别为 100、200、300、400、500 时，C4.5 算法和改进后的 TM-C4.5 算法的分类精度比较结果。

从表 4 可以看出，TM-C4.5 算法的分类精度在不同的测试集下普遍高于原 C4.5 算法的分类精度。测试集实例数目递增时，改进后的算法准确率也逐步上升，证明了 TM-C4.5 算法的有效性。

4 结语

本文引用边界定理作为基础，进行节点贝叶斯概率计算，在决策树生成过程中提高了对子树根节点为连续属性时的离散化效率；在对决策树后剪枝时，在原来的 CCP 方法上添加了评价标准，使得在删除子树根节点时，确保该节点相比于其它节点，在分类过程中对整个决策树影响较小。但是在处理数据集的缺失属性值时，对于占比较大的属性直接删除，虽然省去了处理这些缺失值的麻烦，但是对最后生成的决策树会有一定影响。目前对缺失值的处理方法有删除法、插补法(均值插补、回归插补)等。因此，如何处理那些有较大数量缺失值的属性是下一步研究方向。

参考文献：

[1] 徐洪伟.数据挖掘中决策树分类算法的研究与改进[D].哈尔滨：哈尔滨工程大学，2010.

[2] 王源，王甜甜.改进决策树算法的应用研究[J].电子科技，2010，23(9)：89-91.

[3] 张琳，陈燕，李桃迎，等.决策树分类算法研究[J].计算机工程，2011，37(13)：66-67.

[4] 黄爱辉.决策树 C4.5 算法的改进及应用[J].科学技术与工程，2009(6)：34-36，42.

[5] 邱磊.基于决策树 C4.5 算法剪枝策略的改进研究[D].武汉：华中师范大学，2016.

[6] SONGTHUNG D，SRIPANIDKULHAI K. Improving type 2 diabetes mellitus risk prediction[C]. International Joint Conferenceon Computer Science and Software Engineering (JCSSE)，2016.

[7] DEWAN MD，FARID. Improve the quality of supervised discretization of continuous valued attributes in dataMining[C]. Proceedings of 14th International Conference on Computer and Information Technology (ICCIT 2011)，2011.

[8] 朱琳，杨杨. ID3 算法的优化 [J]. 计算机工程与软件，2016，37(12)：155-161.

[9] 谭俊璐，武建华.基于决策树规则的分类算法研究[J].计算机工程与设计，2010，31 (5)：354-358.

[10] BOVIC KILUNDU，CHRISTOPHE LETOT，PIERRE DEHOMBREUX，et al. Early detection of bearing damage by means of decisiontrees[C]. IFAC Proceedings Volumes，2008.

[11] HAN J C，RODRIGUEZ，BEHESHTI J C M. Diabetes data analysis and prediction model discovery using rapid miner [J]. International Conference on Future GenerationCommunication and Networking，2008(3)：9-99.

[12] 苗煜飞，张霄宏.决策树 C4.5 算法的优化与应用[J]. 计算机工程与应用，2015，51(13)：255-258.

[13] 李瑞，程亚楠.一种改进的 C4.5 算法[J].科学技术与工程，2010，10(27)：6670-6674.

(责任编辑：杜能钢)

(上接第 87 页)

[11] The apache hadoop project [EB/OL]. https://www. hanspub. org/reference/ReferencePapers. aspx? ReferenceID=23066

[12] EKANAYKE J，LI H，ZHANG B，et al. Twister：a runtime for interative mapreduce[C]. Proceeding of the 19th ACM International Symposium on High Performance Distributed Computing. ACM，2010：810-818.

[13] NORSTAD J. A MapReduce algorithm for matrix multiplication [EB/OL]. http://www. norstad. org/matrix-multiply/index. html，2009.

[14] DEAN J，GHEMAWAT S. MapReduce：simplified data processing on large clusters[C]. Proceedings of the 6th Symposium on Operating Systems Design and Implementation，2014：137-150.

[15] SUN Z G，LI T，RISHE N. Large-scale matrix factorization using MapReduce[C]. Proceedings of the 2010 IEEE International Conference on Data Mining Workshops. Washington，DC：IEEE Computer Society，2010：1242-1248.

[16] CHEN K，ZHENG WM. Cloud computing：system instances and current research[J]. Ruanjian Xuebao/Journal of Software，2009，20(5)：1337-1348.

(责任编辑：杜能钢)