文章编号:1001 - 9081(2009)03 - 0839 - 04

基于性价比的分裂属性选择方法

刘星毅
(钦州学院 电大部,广西 钦州 535000)
(qznc@163.com)

摘 要:代价敏感决策树通常讨论测试代价和误分类代价,在其分类过程中,最关键的是节点分裂属性的选择。分析了代价敏感决策树分类问题目前常见的选择分裂属性方法的优、缺点,提出了综合信息量和测试代价并且最大程度降低误分类代价的分裂属性选择方法,UCI 数据集实验结果显示该方法在各个方面好于已有的方法。

关键词:代价敏感;决策树;分裂属性
中图分类号: TP311.13 **文献标志码**:A

Splitting attribute selection method based on cost performance

LIU Xing-yi
(*Department of Radio and TV, Qinzhou University, Qinzhou Guangxi* 535000, *China*)

Abstract: Cost-sensitive decision trees usually concern the discussion of the test cost and misclassification cost. During the classification process, splitting attribute selection is the most important. The paper analyzed the disadvantages and the advantages of the existing methods and proposed a novel method that combined the information ratio in information theory with the cost including the test cost and the misclassification cost to select the split attributes. The experimental results show that this method outperforms significantly the existing methods.

Key words: cost sensitive; decision tree; splitting attribute

1 决策树介绍

数据挖掘要面对的三大挑战性问题是挖掘方法、挖掘对象和挖掘约束,其中分类问题属于三大挑战中的挖掘方法问题。分类算法中最知名的是决策树算法。决策树[1]是以实例为基础的归纳学习方法,它着眼于从一组无次序、无规则的事例中推断出决策树表示形式的分类规则;采用自顶向下的递归方式,在决策树的内部节点进行属性值的比较并根据不同的属性值判断,从该节点向下生长分支,在决策树的叶节点得到结论。决策树方法是一种贪心算法,在操作过程中,每一步都是以获得当前局部最优的结果作为追求的目标。

使用决策树进行分类通常有以下两个步骤:1)使用数据集的部分数据,例如三分之二,建立决策树来训练决策规则;2)根据训练出的规则测试剩下的数据(通常称为测试集),并且根据测试结果对已经建立的决策树进行总结和修正。其中第一步尤为关键,也是广大研究者的兴趣所在。此关键就是如何建立决策树,建树过程又可以分为两个步骤:1)决定分类属性选取规则;2)决定建树过程中停止建树的规则,大部分停止建树的规则类似于文献[1],即出现以下情况的一种就可以停止建树:1)该节点下的训练例子都属于同类;2)该节点下已经没有训练例子可用或者训练例子的数量少于用户规定的数目;3)所有训练例子已经用完。从上面分析可以看出,用决策树进行分类最关键的步骤是分裂属性的选取方法设计问题,本文拟提出一种能同时使分裂信息、分类代价和误分类代价最小三者优化的决策树属性分类算法。

早期的决策树分类的效果一般以准确率为标准,但近来专家认为不同的误分类所带来的损失是不相同的,他们分类的目标都是以误分类代价最小为衡量标准。这样就出现了基于代价敏感的学习方法(Cost-Sensitive Learning ,CSL)[2-4]。代价敏感的学习是当获得每一条件属性值以及发生分类错误时都要付出代价时,进一步优化实用的算法用来学习(从训练数据集中),目标是最小化总的期望代价。例如,参考文献[4,5]分别给出了基于贝叶斯以及决策树的方法,在同时考虑误分类代价和测试代价的情况下,以求得总的代价最小。

但文献[4-5]中所给出的代价的衡量尺度是相同的,也就是说,误分类代价和测试代价具有相同的单位(如医疗诊断中的美元)。在实际应用中,代价的单位尺度往往是很不相同的,有时候我们也很难将多个不同的单位转化进入同一个单位尺度。比如,在医疗诊断中,一个误分类到底该付出多少代价?有时候,一个误分类代价也许会是一个人的生命,那么一个人的生命又值多少钱?本文使用两种不同的代价尺度:有形的代价(测试代价)和无形的代价(误分类代价)。有形的代价可以用货币来准确地衡量,无形的货币则只能用相对值来衡量,正如现实生活中的伤残鉴定。我们假设不同的误分类会产生不同的损失,但这种损失的大小不是简单地用货币来定位的,而是由专家根据具体的实际情况来决定不同的等级,在文中用数值来表示。简言之,在本文中,我们的目标就是要在测试一个新实例时,在一定的有形代价的约束下,尽最大努力付出最小的无形代价,即得到最优的结果。

2 分裂属性的选择

前面分析过决策树的核心问题是如何选取在树的每个节点要测试的属性,称为测试属性或者分裂属性。在建立决策树时减少测试后产生的新子节点的凌乱度是选择分裂属性的基本精神。文献[6]认为选择分类属性的方法可以分为直觉上的方法和信息论方法两大类,文献[6]虽然使用卡方检验方法选择分裂属性,由于原理是利用属性的信息量来分裂属性。因此,可以归为信息论方法。

收稿日期:2008 - 09 - 24;**修回日期**:2008 - 10 - 27。
基金项目:广西自然科学基金资助项目(桂科自 0899018);广西教育厅科研项目(200808MS062)。
作者简介:刘星毅(1972 -),男,广西钦州人,副教授,硕士,CCF 会员,主要研究方向:计算机网络、数据库技术。

代价敏感决策树(CSL)学习中,由于引入了代价概念,很容易就想到使用性价比(Benefit/Cost)的经济概念去考虑分裂属性的选取,即属性分裂时同时考虑信息和代价两个指标,例如EG2[2]和CS-ID3[3]等所谓的性价比的经济方法选择分裂属性,它们已经被证明优于信息论方法,例如文献[6],但是这些算法都没有考虑误分类代价。而文献[4,7-8]在CSL中引入了误分类代价,并且与测试代价相结合提出了使得这两种代价最小化的代价敏感决策树;但这些算法只重视和强调代价问题,而置各个属性的分类能力于不顾。这显然不合理,因为每个属性的分类能力是不同的,并且分类能力的大小或者每个属性所含分类信息的多少跟测试代价的大小通常没有什么比例可言。所以在考虑分裂属性的时候,必须折中考虑代价和属性分类能力的影响。在本文中,我们首先考虑性价比来选取分类属性,然后同时考虑分裂属性必须提供很小的误分类代价。也就是说,本文认为选择分裂属性的选择应该是使误分类代价的减小量较大,且要求属性的性价比比较好。下面将详细介绍本文的算法设计过程。

2.1 性价比(Benefit/Cost)

以性价比(Benefit/Cost)为宗旨的分裂属性选择标准是综合测试代价(Cost)和属性的区分能力(Discrimination Efficiency),就是通常我们所说的每个代价单元能提供的属性区分或者分类信息能力。它在本文中被定义为:

$$f(Benefit/Cost) = \frac{f(\text{cost of attribute } i)}{g(\text{discrimination efficiency of attribute } i)} \tag{1}$$

为了得到 $g(DE)$,假设存在一个事件 E(Evidence Assertion)和一个对此事件的假设推断 H(Hypothesis Assertion),即在 H 的前提下事件 E 发生的概率为 $P(E/H)$,而记在假设 H 的前提下事件的对立假设是 $\sim H$,即 $\sim H$ 的前提下事件 E 发生的概率为 $P(E/\sim H)$,根据概率统计假设理论[9],定义关于此事件 E 的似然比(Likelihood Ratio,LR)为:

$$LR = \frac{P(E/H)}{P(E/\sim H)} \tag{2}$$

LR 的值越大,事件 E 发生的可能性就越大,反之亦然。根据此理论,用 $g(DE)$ 定义分裂属性 A_i 区分能力:

$$g(DE) = \frac{Useful\_Information}{Non\_useful\_Information} = \frac{UI}{NI} \tag{3}$$

定义总的信息量 $TI = UI + NI$,所以把式(3)转化成如下:

$$g(DE) = \left[\frac{UI}{NI}\right] = \left[\frac{TI}{NI}\right] - 1 \tag{4}$$

通过 Shannon 信息熵的定义知道 $TI = 2^{H(TI)}$;$NI = 2^{NI}$ 带入式(4)得到:

$$g(DE) + 1 = 2^{(H(TI) - H(NI))} \tag{5}$$

再由 Shannon 信息熵的定义知:$Gain(A_i) = H(TI) - H(NI)$,因此,式(5)转化成式(6):

$$Gain(A_i) = \text{lb}\left[\frac{TI}{NI}\right] \tag{6}$$

且有

$$2^{Gain(A_i)} = \left[\frac{TI}{NI}\right] + 1 \tag{7}$$

综合式(4)、(7),有:

$$g(DE) = 2^{Gain(A_i)} - 1 \tag{8}$$

测试代价函数 $f(cost)$ 被表示为 $cost + \omega$,其中 ω 是 $cost$ 的无穷小量,目的是为了使得当 $cost$ 为0时保证分母有意义。这样属性 A_i 性价比(Benefit/Cost)式(1)就可以成为式(9):

$$f(Benefit/Cost) = \frac{2^{Gain(A_i)} - 1}{Cost_i + \omega} \tag{9}$$

2.2 误分类代价减少量

本文的模型使用一个启发式搜索的方法来计算每个实行的误分类代价的减少量,即计算该属性的净减少量作为衡量标准。由属性 A_i 所带来的误分类代价的减少量 $Redu\_Mc(A_i)$ 是,

$$Redu\_Mc(A_i) = Mc(i) - \sum_{j=0}^{n} p(i,j)Mc(A_j) \tag{10}$$

其中,$Mc(i)$ 为在未选取属性 A_i 作为分裂属性对当前节点进行分裂时节点 N 的误分类代价。$P(i,j)$ 是属性 A_i 取第 j 个值的概率。如果某个属性的 $Redu\_Mc(A_i)$ 最大并且值是正数,说明这个属性具有最大的区分能力:最大说明属性的区分能力大,是正数说明用这个属性作分裂属性后误分类错误能降低,这是本文选择分裂属性的最低标准,如果在求出的 $Redu\_Mc(A_i)$ 中最大的值不是正数就停止建树。

设当前节点 Node 中包含 p 个正例和 n 个反例,且假设 $p * FN > n * FP$,即当前节点为正例节点,则当前节点的误分类代价:$Mc = n * FP$。若选择 A_i 作为分裂属性,A_i 有 n 个属性值,则当前节点 $Node$ 有 n 个子节点($Node_1$, $Node_2$, …, $Node_n$)。若子节点 N_i 有 p_i 个正例和 n_i 个反例,设前 r 个子节点为正例节点,后 $(n-r)$ 个子节点为反例节点。则用属性 A_i 分裂节点 N 后的误分类代价如下:

$$\sum_{i=0}^{n} Mc(A_i) = n * FP - \left(FP * \sum_{i=0}^{r} P_{FP}(i,j) \cdot n_i - FN * \sum_{i=r+1}^{n} P_{FN}(i,j) \cdot p_i\right) \tag{11}$$

以一个含有332个实例来说明误分类代价的减少量的计算。此节点作为当前节点,用属性 c_2 将其分裂,具体数据见图1。

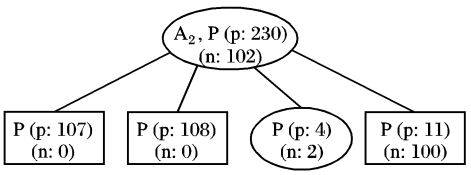

图1 一棵决策树的一部分

图1中,每个节点中括号外的若为P表示该节点为正例节点,若为N则表示反例节点。括号内的"p:230"表示正例数为230;"n:102"表示反例数为102。

该例中,假设 $FP = 4000$,$FN = 8000$,则根据定义,在用属性 c_2 进行分裂前的误分类代价为:

$$Mc = n * FP = 102 * 4000 = 408000$$

$$\sum_{i=1}^{4}(A_2) = 0 * 4000 * \frac{117}{332} + 0 * 4000 * \frac{108}{337} + 2 * 4000 * \frac{6}{332} + 11 * 8000 * \frac{111}{332} = 29566$$

$$Redu\_Mc(A_2) = 408000 - 96000 = 478434$$

算法所期望的是选择一分裂属性使误分类代价的减小量较大,且要求性价比较高。所以综合式(9)和(11)使用 $Performance(A_i)$ 来表示要选取的分裂属性:

$$Performance(A_i) = \frac{2^{Gain(A_i)} - 1}{(Cost_i + \omega)} . Redu\_Mc(A_i) \tag{12}$$

这样根据式(12),当 $f(Benefit/Cost)$ 和 $Redu\_Mc(A_i)$ 最大时 $Performance(A_i)$ 得到最大值。如果两个属性的 $Performance$ 相同,算法趋向于选择 $Redu\_Mc(A_i)$ 大的属性作当前的分裂属性,因为CSL的目标就是使得误分类代价最小。但是仔细观察式(12)可以发现:

在函数 $f(Benefit/Cost)$ 中,分子和分母数量级不一样大

时,结果可能偏向数量级大的(如果 *Cost* 数量级大就偏向 *Cost*),这样就会出现常见的数量级偏置(bias)问题。

同样的道理,如果 $f(Benefit/Cost)$ 和 $Redu\_Mc(A_i)$ 数量级不一样,结果可能偏向数量级大的,也会出现 bias。

对于第一个问题,本文在 $f(Benefit/Cost)$ 中分母加个指数 C(C 为实数)来避免这种 bias,实验结果表明 C 在范围 $\frac{Mean(2^{Gain(A_i)}-1)}{Mean(Cost_i)}$(其中 $Mean(.)$ 表示 . 的均值)效果最好,其实理论上也是这样,因为在这个范围,分子和分母基本上就处于同一个数量级了。

为了解决第二个 bias,因为 $Redu\_Mc(A_i)$ 是个无单位量纲,可以首先把 $Redu\_Mc(A_i)$ 规范到 0 到 1 之间,本文规范的公式为:

$$Redu\_Mc(A_i)_{norm} = Redu\_Mc(A_i) \times \frac{\max(Redu\_Mc(A_j))}{\max(Redu\_Mc(A_j)) - \min(Redu\_Mc(A_j))};$$
$$j = 1,\cdots,n,\ i = 1,\cdots,n, \quad (13)$$

然后再把 $Redu\_Mc(A_i)_{norm}$ 规范到函数 $g(DE)$ 的数量级范围。通常可以把已经规范到 0 到 1 之间的数值乘以 $\max(g(DE)) - \min(g(DE))$ 就行了,其中 max 代表最大值,min 代表最小值。这样,分裂属性的选择就可以依照公式进行:

$$Performance(A_i) = \frac{2^{Gain(A_i)}-1}{(Cost_i+\omega)^c} \times Redu\_Mc(A_i)_{norm} \quad (14)$$

3　实验结果分析

分裂属性选择标准确定之后,本文算法在代价敏感决策树学习中使用这个模型,建树的方法大致与文献[7-8]一样,唯一的区别是它们在选择分裂属性的时候是以代价减少的最大量作为衡量标准。本文是根据式(14)来选择分裂属性,这样就可以与现有的方法进行比较。在下面的实验中,本文从 UCI[10] 数据集中选取了代价敏感学习常用的四个数据集 Breast,Heart,Australia 和 Voting,因为决策树分类和代价敏感决策树分类通常处理的是离散值属性,这些数据集中的数据首先用最小熵的方法[11] 把各个数据集的数值离散化。

表 1　实验数据

数据集	属性个数	事例个数	类分布 (N/P)
Breast	9	683	444/239
Heart	8	161	98/163
Australia	15	653	296/357
Voting	16	232	108/124

选用这些数据集是因为这些数据集都只有两个类标签,并且都含有相当多的实例个数。每个数据集被分成两个部分:训练集(60%)和测试集(40%)。但对于这些数据集中的一些细节信息,如每项测试的代价和误分类代价,实际上无从得知。为了让实验能够顺利进行,从 0 ~ 100 中随机地选取一个数作为每个数据集中的每个测试属性的测试代价。虽然这种值不是真实的,但对于所有的试验都使用相同测试代价,则对于比较的结果来说,它们的效果是等同的。因为在本文中,所讨论的是误分类代价与测试代价具有两个不同的代价尺度,测试代价一般以货币作为单位,而误分类代价没有单位,只是由专家评定的发生错误时所受损失大小的相对值。在本文的实验中,假设错误的正例(FP)的代价为 1;错误的反例(FN)的代价为 3,即表示发生错误的反例是发生错误的正例所付出的代价的 3 倍,但是在实验中误分类代价为了避免出现偏置,是随着 $g(DE)$ 函数的数量级变化的,但是 FP 和 FN 的关系保持 FN 的代价是 FP 的代价的 3 倍不变。

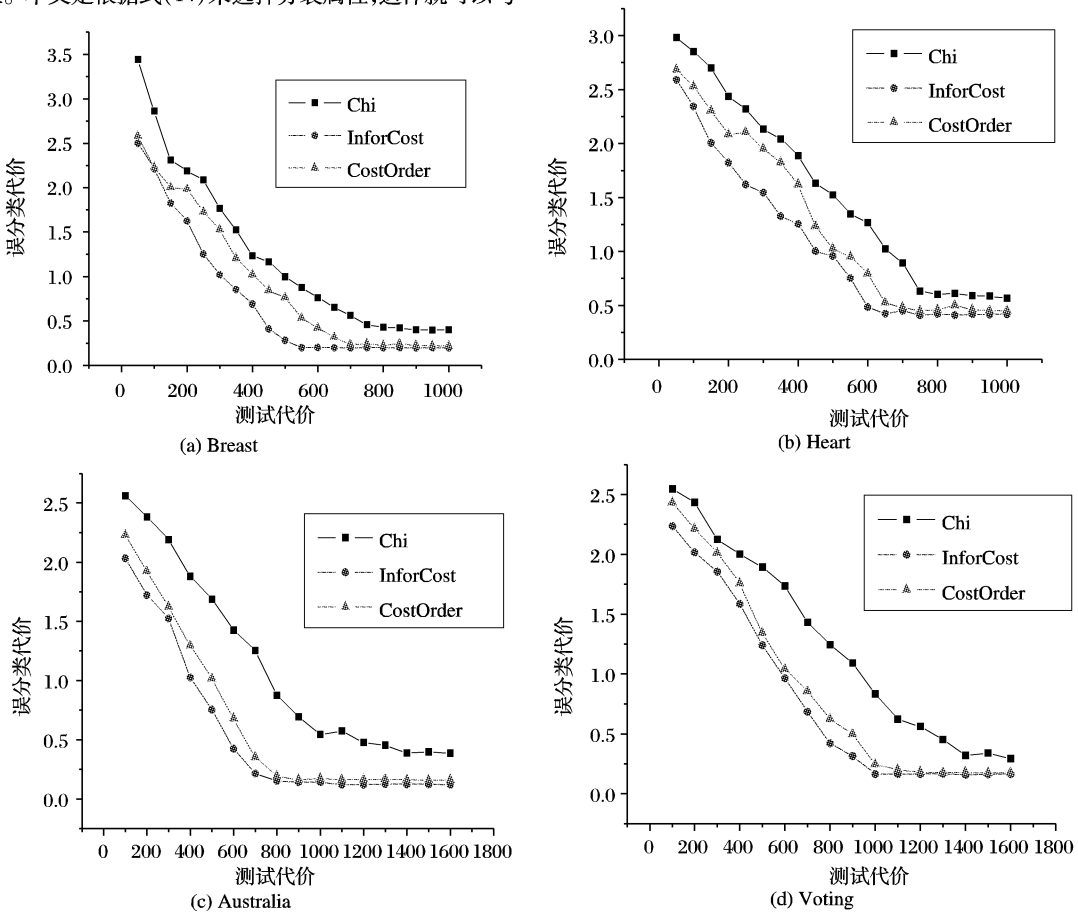

图 2　4 个数据集的误分类代价

图2中“CostOrder”代表文献[7-8]以代价为选择分裂属性衡量标准的方法,“InforCost”是本文考虑了代价和属性所含信息的方法,“Chi”方法[6]是一种使用卡方检验选择分裂属性的方法,试验中选用此方法进行比较的目的是因为卡方方法在文献[6]已经被证明优于信息熵的方法,但是文献[6]没有显示此方法是否在代价敏感决策树中也优于信息熵方法,因此,本文也构造相关试验来测试这个问题。

从实验结果可以看出:

1)当测试代价小时,三种方法误分类代价减少得很慢,但是本文方法对低测试代价没有其他两种方法敏感。

2)在相同的测试代价情况下,本文方法在四个数据中所产生的误分类错误代价都比另外两种方法低,即效果好。

3)在要降低相同的误分类代价情况下,本文方法所需的测试代价都比两种方法要少,而代价敏感分类的原则就是同等测试代价下误分类代价降低最多,这说明本文算法优于其他算法。

4)从四个试验结果可以看出,卡方算法均比信息熵算法效果要差。这说明虽然文献[6]已经证明在非代价敏感分类中,卡方方法优于信息熵方法,但是在代价敏感学习中却不是这样。

5)从图2可以发现,到了一定测试代价,误分类代价就减少得缓慢了,这个实验结果具有非常大的实用价值。比如,医疗过程中可以告诉患者,即使再多加钱,病情(即误分类代价)也不能得到成比例甚至不能再减轻,可以避免患者浪费钱,也可以避免患者受不良医务人员的欺骗;同时我们也发现,相对文献的方法,我们的方法花较少的代价就可以使误分类代价达到稳定值(因为这种分类问题通常是NP问题,没有最优解)。

6)从各图中所使用的测试代价可以看出,数据集Breast和数据集Heart有差不多个的属性,但是由于数据集事例不同,例如数据集Heart事例较少,所以实验的效果稍差,图中另外两个数据集Australia和数据集Voting也有同样的结果。

4 结语

本文利用每个属性所含的信息量和测试代价,提出了性价比的概念,并且综合了性价比和误分类减少这两个量,提出了一种新的分裂属性的选择方法。实验证明我们的方法优于文献[7-8]的方法和文献[6]的方法,并且从分析实验结果可以看出,本文方法有一定的实用价值。

参考文献:

[1] QUINLAN J R. Induction of decision trees[J]. Machine Learning, 1986, 1(1): 81-106.

[2] NUNEZ M. The use of background knowledge in decision tree induction[J]. Machine Learning, 1991, 6(3): 231-250.

[3] TAN M. Cost-sensitive learning of classification knowledge and its applications in robotics[J]. Machine Learning, 1993, 13(1): 7-33.

[4] LING C X, YANG Q. Decision trees with minimal costs[C]// Proceedings of 2004 International Conference on Machine Learning (ICML'2004). New York: ACM, 2004: 69.

[5] CHAI XIAOYONG, DENG LIN, YANG QIANG, *et al*. Test-cost sensitive naive Bayes classification [C]// Proceedings of The 2004 IEEE International Conference on Data Mining (ICDM'2004). Washington, DC: IEEE Computer Society, 2004: 51-58.

[6] 刘星毅. 一种新的决策树分类属性选择方法[J]. 计算机技术与发展, 2008, 18(5): 70-72.

[7] QIN ZHENGXIN, ZHANG SHICHAO, ZHANG CENGQI. Cost-sensitive decision trees with multiple cost scales [C]// Australian Conference on Artificial Intelligence, LNCS 3339. Berlin: Spring, 2005: 380-390.

[8] QIN ZHENXING, ZHANG CHENGQI, XIE XUEHUI, *et al*. Dynamic test-sensitive decision trees with multiple cost scales [C]// Proceedings of FSKD 2005, LNCS 3613. Berlin: Springer, 2005: 402-405.

[9] GASCHNIG J, DUDA R O, HART P E. Model design in the prospector consultant system for mineral exploration [C]// Expert Systems in the Microelectronic Age. Edinburgh, Scotland: Edinburgh University Press, 1979.

[10] BLADE C L, MERZ C J. UCI repository of machine learning databases[Z]. Irvine, CA: Department of Information and Computer Science, University of California, 1998.

[11] FAYYAD U M, IRANI K B. Multi-interval discretization of continuous-valued attributes for classification learning [C]// Proceedings of IJCAI. San Francisco: Morgan Kaufmann, 1993: 1022-1027.

[12] TURNEY P. Types of cost in inductive concept learning [C]// Proceedings of the Cost-Sensitive Learning Workshop at the 17th ICML-2000 Conference. San Francisco: Morgan Kaufmann, 2000: 15-21.

(上接第838页)

[10] SAITOU N, NEI M. The neighbor-joining method: A new method for reconstructing phylogenetic trees[J]. Molecular Biology and Evolution, 1987, 4(4): 406-425.

[11] SOBER E. Reconstructing the past: Parsimony, evolution and inference[M]. Cambridge: MIT Press, 1988.

[12] SOURDIS J, NEI M. Relative efficiencies of the maximum parsimony and distance matrix methods in obtaining the correct phylogenetic tree [J]. Molecular Biology and Evolution, 1988, 5(3): 298-311.

[13] HOLDER M, LEWIS P O. Phylogeny estimation: Traditional and Bayesian approaches[J]. Journal of Nature Reviews Genetics, 2003(4): 275-284.

[14] LI W-H. Evolutionary change of restriction cleavage sites and phylogenetic inference[J]. Genetics, 1986, 113(1): 187-213.

[15] ROCH S. A short proof that phylogenetic tree reconstruction by maximum likelihood is hard [J]. ACM Transactions on Computational Biology and Bioinformatics, 2006, 3(1): 92-94.

[16] 谢季坚, 刘承平. 模糊数学方法及其应用[M]. 3版. 武汉: 华中科技大学出版社, 2007.

[17] LIAO BO, SHAN XINZHOU, ZHU WEN, *et al*. Phylogenetic tree construction based on 2D graphical representation[J]. Chenical Physics Letters, 2006, 422(1/3): 282-288.

[18] WANG WEIPING, LIAO BO, WANG TIANMING, *et al*. A graphical method to construct a phylogenetic tree[J]. International Journal of Quantum Chemistry, 2006, 106(9): 1998-2005.

[19] 李刚成, 廖国华. 基于4D表示的DNA序列分析方法[J]. 科学技术与工程, 2008, 8(6): 1405-1409

[20] CAO ZHI, LIAO BO, LI RENFA. A group of 3 D graphical representation of DNA sequences based on dual nucleotides[J]. International Journal of Quantum Chemistry, 2008, 108(9): 1485-1490.