有根系统发生树的精确有效比较

李曙光[1,2] 陈姝颖[2] 朱丽波[3]
(山东省高校智能信息处理重点实验室(山东工商学院) 烟台 264005)[1]
(山东工商学院计算机科学与技术学院 烟台 264005)[2]
(内蒙古师范大学计算机与信息工程学院 呼和浩特 010020)[3]

摘 要 系统发生树代表了不同物种之间进化关系的历史。生物信息学中的一个基本问题是对系统发生树进行比较。一种比较方法是通过定义树空间中两棵系统发生树之间的相似度或相异度来测定这两棵树的同异。Robinson-Foulds距离是目前使用最广泛的相异度。定义了一个用于有根系统发生树比较的新的相异度,该相异度考虑了子类间更精细的相似,而不是如Robinson-Foulds距离那样仅考虑子类相同与否,因此能够提供更精确、清晰的测量。给出了两个能有效计算这个相异度的算法。简单修改之后,这些结果适用于其他5个相关的比较指标。

关键词 系统发生树,树比较,相异度,Robinson-Foulds距离,子类
中图法分类号 TP301 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.12.061

Accurately and Efficiently Comparing Rooted Phylogenetic Trees

LI Shu-guang[1,2] CHEN Shu-ying[2] ZHU Li-bo[3]
(Key Laboratory of Intelligent Information Processing in Universities of Shandong, Shandong Institute of Business and Technology, Yantai 264005, China)[1]
(College of Computer Science and Technology, Shandong Institute of Business and Technology, Yantai 264005, China)[2]
(School of Computer & Information Engineering, Inner Mongolia Normal University, Hohehot 010020, China)[3]

Abstract Phylogenetic trees represent the historical evolutionary relationships among different species. Comparing phylogenetic trees is a fundamental task in bioinformatics. One way for tree comparison is to define a pairwise measure of similarity or dissimilarity in the tree space to determine how different two trees are. Robinson-Foulds distance is by far the most widely used dissimilarity measure. A new pairwise dissimilarity measure for comparing rooted phylogenetic trees was defined which takes into account not only the identity of clusters in the case of Robinson-Foulds distance, but also more subtle similarities between clusters, and thus may provide more accurate and cleaner measurement than Robin-Foulds distance. Two algorithms to compute this measure efficiently were presented. With slight modifications of these results, they can be applied to five other related comparison indices.

Keywords Phylogenetic trees, Tree comparison, Dissimilarity measure, Robinson-Foulds distance, Cluster

1 引言

系统发生(phylogeny)是指生物形成和进化的历史。系统发生学研究生物进化规律及物种间的亲缘关系,其基本思想是根据现有生物基因或物种多样性来重建生物的进化史。由于能够揭示不同物种和基因之间的关系,系统发生分析广泛应用于分子遗传学、医学及药物设计等领域中[1,2]。

表示物种之间系统发生关系的树状图称为系统发生树(phylogenetic tree),用图论语言来描述,系统发生树是一棵树,每个叶子节点(度为1的节点)分配了唯一的标签(物种,分类单元)。度大于1的节点称为内部节点。有根树指定了某个内部节点作为树根,除树根之外其他内部节点的度均不小于3;无根树所有内部节点的度均不小于3[3]。

系统发生树的构建是一个NP完全问题[4],目前主要方法有:距离法(如非加权平均法UPGM[5]、邻接法(neighbor joining)[6]以及Fitch-Margoliash方法[7])、最大简约法(maximum parsimony)[8]、最大似然法(maximum likelihood)[9,10]、贝叶斯推断法(Bayesian inference)[11,12]等。国内学者也提出了较好的几种建树方法,如基于遗传算法的最大似然法[13]、基于代谢路径的方法[14]、基于k-mer组分信息的方法[15]等。具体请参阅文献[16,17]。

由于各种建树方法采用的优化标准、搜索树空间的方法以及对输入不确定性的灵敏度不同,依据同一组数据,采用各种方法往往会构建出不同的系统树。另一方面,最大简约法和最大似然法构建的树可能不唯一,尤其是当叶子数目较大时,贝叶斯推断法的目标就是要得到一个树的集合。因此,在

到稿日期:2014-07-21 返修日期:2014-12-05 本文受国家自然科学基金(61173173,61272430,61373079,61379019),教育部科技研究重点项目(212101),山东省自然科学基金项目(ZR2013FM015,ZR2011FL004)资助。
李曙光(1970—),男,博士,副教授,CCF会员,主要研究方向为组合最优化、生物信息学,E-mail:sgliytu@hotmail.com;陈姝颖(1979—),女,硕士,讲师,CCF会员,主要研究方向为生物信息学;朱丽波(1972—),女,硕士,副教授,硕士生导师,主要研究方向为计算机应用。

后期处理时，需要进行系统发生树的比较[18]。系统发生树的比较在生物学领域还有很多其他应用，如用来挖掘系统发生信息数据库[19]、研究寄主和宿主的关系[20]等；其在非生物学领域的应用也比较广泛，比如用于语言学[21]或计算机科学[22]。

直接分析原始数据集产生的系统发生树称为源树(source tree)，总结一组源树所得到的一棵树称为合一树(consensus tree)。合一树用来总结对同一数据集应用不同方法产生的源树的共同部分，或概括由不同数据集产生的源树的共同特征。如果不同数据集所包含的分类单元集合部分重叠，则这些源树的系统发生树称为超树(super tree)[2]。

系统发生树的比较研究主要关注两个问题：构建合一树及计算合一指标(consensus index)[23,24]。这两个问题密切相关：构建合一树要基于某个合一指标的优化，合一指标刻画了诸源树之间的合一程度。设计两棵源树的合一指标，即两棵树的比较指标，如相异度(dissimilarity)和相似度(similarity)，以构成这两个问题的研究基础。相异度是两棵树差异程度的数值度量，两棵树越相似，它们的相异度就越低，相似度则反之。

有根树的每个节点对应一个子类，即以该节点为根的子树的叶子标签构成的集合。树根及叶子节点所对应的子类称为平凡的，其他的称为非平凡的。所有子类构成的集合称为该有根树的子类表达，它唯一确定这棵树，线性时间精确算法可根据子类表达还原这棵树[3]。子类表达在有根树的比较中有重要作用。文献[25]比较了两棵有根树，列出了11个基于子类表达的比较指标，并详细说明了这些指标的出处。第一个指标是目前使用最广泛的Robinson-Foulds距离[26]，其值是两棵树子类表达的对称差的基数(更精确地说，是Robinson-Foulds距离的两倍)；接下来6个可视为Robinson-Foulds距离的衍生指标，剩余4个为基于两棵树的合一树的子类表达。

Robinson-Foulds距离对树拓扑结构的微小变化反应过于灵敏。例如，在一棵毛毛虫树(caterpillar tree，所有内部节点形成一条路的树)上，交换离树根最远和最近的两个叶子节点的标签，则新树和原树的Robinson-Foulds距离立刻达到最大值。这一指标的分布函数曲线的偏度过大，理论分析表明其渐进分布近似为泊松分布[27]。这些缺陷影响了该指标的使用效果。Bogdanowicz和Giaro对这一指标进行了改进，定义了匹配子类距离(matching cluster distance)，其在分布、分辨力、灵敏度等方面均优于Robinson-Foulds距离[28]。计算匹配子类距离的时间复杂度为$O(n^{2.5}\log n)$，其中n表示每棵树上的叶子数目。这个时间复杂度很难进一步降低，因为调用了求解完全二分图中的最小费用完美匹配的经典算法[29]，其时间复杂度已经是$O(n^{2.5}\log n)$。对于无根树也有类似推广[30,31]。

本文的主要贡献如下。定义了比较两棵有根树的一个新的相异度：子类相异度。这个指标考虑了子类间更精细的相似，而不是如Robinson-Foulds距离那样仅考虑子类相同与否，因此能够提供更精确清晰的测量。初步研究了子类相异度的性质，并给出了计算它的两个算法，其时间复杂度分别为$O(n^3)$和$O(n^2)$。利用这些结果得到了5个衍生指标。

本文第2节介绍有关术语和记号；第3节给出子类相异度的定义，并初步研究它的性质；第4节给出计算子类相异度的两个多项式时间精确算法；第5节利用所得结果推广文献[25]中的5个Robinson-Foulds距离的衍生指标；最后总结全文。

2 预备知识

令$|A|$表示集合A的基数。称$A\Delta B=(A\cup B)-(A\cap B)$为集合$A$和$B$的对称差。令$G=(V,E)$表示顶点集和边集分别为$V$和$E$的图。若一个图的顶点集可分解为不相交的两个集合$V_1$和$V_2$，使得同一集合中的任意两个顶点不相邻，则称该图为二分图，记为$G=(V_1,V_2,E)$。如果每两个顶点$v_1\in V_1$和$v_2\in V_2$都相邻，则称这个二分图为完全二分图。树是一个连通无圈图，树上的顶点习惯上称为节点。路是有两个节点度为1其余节点度为2的树。

令L表示标签集合，$|L|=n$，L中的每个标签代表一个现存物种(分类单元)。令$T=(V,E)$表示一棵有根系统发生树，其叶子节点与L中的标签一一对应，非叶子节点都没有标签。每个非叶子节点代表由此分出的叶子节点的假想祖先，树根$r(T)$是所有叶子节点的假想祖先。为方便起见，T的叶子和叶子标签不加区分，即$L(T)$既表示T的叶子集合，也表示叶子标签集合。叶子集合为L的所有有根系统发生树构成的集合记为$\mathscr{R}_L$。若除树根之外其他内部节点的度均等于3，则称为二歧树，否则称为多歧树。

设$T\in\mathscr{R}_L$，$v\in V(T)$。树T中以节点v为根的子树记为T_v。称$L(T_v)$为节点v对应的子类。树T所有子类构成的集合记为$\beta(T)$，所有非平凡子类构成的集合记为$\beta_*(T)$。称$\beta(T)$为树T的子类表达。显然，树T有$n+1$个平凡子类，非平凡子类的个数不超过$n-2$。若树T为二歧树，则恰有$n-2$个非平凡子类。

定义1[26] 两棵有根树$T_1,T_2\in\mathscr{R}_L$之间的Robinson-Foulds距离定义为：

$$RF(T_1,T_2)=|\beta_*(T_1)\Delta\beta_*(T_2)|/2 \tag{1}$$

例如，如图1所示，T_1和T_2是两棵有根树，$L=\{a,b,c,d\}$，则有$\beta_*(T_1)=\{\{a,b\},\{c,d\}\}$，$\beta_*(T_2)=\{\{a,b,c\}\}$。故有$RF(T_1,T_2)=1.5$。

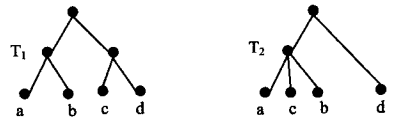

图1 两棵有根系统发生树T_1和T_2

3 子类相异度的定义及性质

本节将给出子类相异度的定义，并研究它的性质。

定义2[28] 两个子类C_1和C_2之间的权重定义为：

$$wt(C_1,C_2)=|C_1\Delta C_2| \tag{2}$$

权重$wt(C_1,C_2)$刻画了两个子类的相似程度：C_1和C_2的权重越小，则越相似。

设$C_{1,i}\in\beta_*(T_1)$，$1\leqslant i\leqslant|\beta_*(T_1)|$。定义$C_{1,i}$到$T_2$的距离$d(C_{1,i},T_2)$为$\min\limits_{C_{2,j}\in\beta(T_2)}\{wt(C_{1,i},C_{2,j})\}$。称$C'_{1,i}=\arg\min\limits_{C_{2,j}\in\beta(T_2)}\{wt(C_{1,i},C_{2,j})\}$为$C_{1,i}$在$T_2$中的最相似子类。若$C'_{1,i}$不唯一，则取使$|C_{1,i}\cap C'_{1,i}|$最大的那个子类。

类似地，设$C_{2,j}\in\beta_*(T_2)$，$1\leqslant j\leqslant|\beta_*(T_2)|$。定义$C_{2,j}$到$T_1$的距离$d(C_{2,j},T_1)$为$\min\limits_{C_{1,i}\in\beta(T_1)}\{wt(C_{2,j},C_{1,i})\}$。称$C'_{2,j}=$

$\arg\min_{C_{1,i}\in\beta(T_1)}\{wt(C_{2,j},C_{1,i})\}$为$C_{2,j}$在$T_1$中的最相似子类。若$C'_{2,j}$不唯一，则取使$|C_{2,j}\cap C'_{2,j}|$最大的那个子类。

定义3 两棵有根树$T_1,T_2\in\mathcal{R}_L$之间的子类相异度定义为：

$$Cdis(T_1,T_2)=(\sum_{C_{1,i}\in\beta_*(T_1)}d(C_{1,i},T_2)+\sum_{C_{2,j}\in\beta_*(T_2)}d(C_{2,j},T_1))/2 \quad (3)$$

例如，考虑图1所示的两棵树T_1和T_2，则有$d(\{a,b\},T_2)=wt(\{a,b\},\{a,b,c\})=1$，$d(\{c,d\},T_2)=wt(\{c,d\},\{d\})=1$；$d(\{a,b,c\},T_1)=wt(\{a,b,c\},\{a,b\})=1$。故有$Cdis(T_1,T_2)=1.5$。注意$\{a,b\}$的最相似子类是$\{a,b,c\}$，而不是$\{a\}$或$\{b\}$，虽然$\{a,b\}$与这3个子类的权重均等于1。

定义4 相异度d在集合X上的直径$\delta_d(X)$定义为X中任意两个元素关于此相异度的最大值。

定理1 子类相异度$Cdis$在$\mathcal{R}_L$上的直径$\delta_{Cdis}(\mathcal{R}_L)$为$\Theta(n^2)$。

证明：设$T_1,T_2\in\mathcal{R}_L$。由于$T_1(T_2)$的每个非平凡子类到$T_2(T_1)$的距离不超过n，并且$T_1(T_2)$的非平凡子类的个数不超过$n-2$，因此有$\delta_{Cdis}(\mathcal{R}_L)\leqslant n^2-2n$。

另一方面，如图2所示的两棵有根树，$Cdis$值不小于$(n^2-2n)/2$。

因此得到：$\delta_{Cdis}(\mathcal{R}_L)=\Theta(n^2)$。

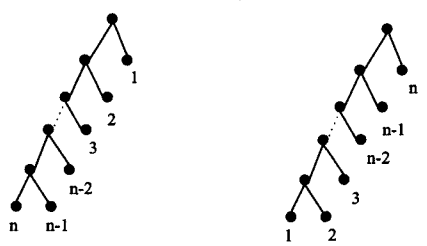

图2 两棵有根系统发生树

定义5[3] 设$T\in\mathcal{R}_L$，$X\subseteq L$。树T在X上的限制生成树$T(X)$定义为连接X中所有叶子的T的最小子图。树T在X上的限制树$T_{|X}$按如下方式得到：将$T(X)$中每条由度为2的节点所构成的极大路替换为连接该路仅有的两个邻点的一条边。

定理2 设$T_1,T_2\in\mathcal{R}_L$，$X=L\backslash\{x\}$。则有：

$$Cdis(T_1,T_2)\geqslant Cdis(T_{1|X},T_{2|X})$$
$$Cdis(T_1,T_2)\leqslant Cdis(T_{1|X},T_{2|X})+2n-3 \quad (4)$$

证明：$T_{1|X}$的每个非平凡子类C_1^X对应T_1的一个子类C_1，使得$C_1=C_1^X$或者$C_1=C_1^X\cup\{x\}$。因此有$d(C_1,T_2)\geqslant d(C_1^X,T_{2|X})$，$d(C_1,T_2)\leqslant d(C_1^X,T_{2|X})+1$。

由于$T_{1|X}$和T_1中分别有$n-4$和$n-3$个非平凡子类，T_1有一个非平凡子类在$T_{1|X}$中没有对应，记这个子类为C_1'。显然有：$d(C_1',T_2)\leqslant n$。

类似地，$T_{2|X}$的每个非平凡子类C_2^X对应T_2的一个子类C_2，使得$C_2=C_2^X$或者$C_2=C_2^X\cup\{x\}$。因此有：$d(C_2,T_1)\geqslant d(C_2^X,T_{1|X})$，$d(C_2,T_1)\leqslant d(C_2^X,T_{1|X})+1$。由于$T_{2|X}$和$T_2$中分别有$n-4$和$n-3$个非平凡子类，$T_2$有一个非平凡子类在$T_{2|X}$中没有对应，记这个子类为$C_2'$。显然有：$d(C_2',T_1)\leqslant n$。

综合以上分析，定理2得证。

定理2表明，将常数数目的叶子移位，有根树子类相异度的变化为$O(n)$，渐进意义下，这一数值远远小于定理1所示的子类相异度的最大值$\Theta(n^2)$。这说明子类相异度对树拓扑结构微小变化的反应明显优于Robinson-Foulds距离，从而能够提供比Robinson-Foulds距离更精确、清晰的测量。

4 子类相异度的计算

本节研究如何计算两棵有根树$T_1,T_2\in\mathcal{R}_L$之间的子类相异度，将给出两个多项式时间的精确算法。这两个算法都是基于有根树的后序遍历，但是计算子类之间的权重时有所不同，因此时间复杂度有差异。算法执行过程中计算了$T_1(T_2)$的每个非平凡子类在$T_2(T_1)$中的最相似子类，以为下一节做准备。

算法1

第1步 分别后序遍历T_1和T_2，得到T_1和T_2的所有子类。

第2步 对T_1的每个非平凡子类$C_{1,i}\in\beta_*(T_1)(1\leqslant i\leqslant|\beta_*(T_1)|)$，根据式(2)计算$C_{1,i}$与$T_2$的每个子类(不一定是非平凡的)之间的权重，取这些权重的最小者为$d(C_{1,i},T_2)$。在此过程中，根据上节所述得到$C_{1,i}$在T_2中的最相似子类$C'_{1,i}$。

第3步 对T_2的每个非平凡子类$C_{2,j}\in\beta_*(T_2)(1\leqslant j\leqslant|\beta_*(T_2)|)$，根据式(2)计算$C_{2,j}$与$T_1$的每个子类(不一定是非平凡的)之间的权重，取这些权重的最小者为$d(C_{2,j},T_1)$。在此过程中，根据上节所述得到$C_{2,j}$在T_1中的最相似子类$C'_{2,j}$。

第4步 根据式(3)得到T_1和T_2之间的子类相异度$Cdis(T_1,T_2)$。

我们有如下定理。

定理3 算法1的时间复杂度为$O(n^3)$，其中n表示每棵树上的叶子数目。

证明：算法1的时间复杂度取决于第2步和第3步。计算两个子类之间的权重需要$O(n)$时间。又因每棵树所有子类的个数不超过$2n-1$，故$T_1(T_2)$的每个非平凡子类到$T_2(T_1)$的距离以及它在$T_2(T_1)$中的最相似子类可以在$O(n^2)$时间之内得到。由于每棵树非平凡子类的个数不超过$n-2$，因此第2步和第3步的时间复杂度均为$O(n^3)$。因此，整个算法的时间复杂度为$O(n^3)$。

本节剩余部分说明如何修改算法1，将其时间复杂度降低为$O(n^2)$。

设$f:L\to\{1,2,\cdots,n\}$是任意一个双射。分别遍历T_1和T_2，按照f修改叶子节点。现在，每个子类可以表示为一个长度为n的向量V：若$i(1\leqslant i\leqslant n)$属于这个子类，则$V[i]=1$，否则$V[i]=0$。例如，图1中设$a,b,c,d$分别对应1,2,3,4，则$T_1$的子类$\{a,b\}$表示为$\langle1,1,0,0\rangle$，$T_2$的子类$\{a,b,c\}$表示为$\langle1,1,1,0\rangle$。

子类表示为向量后，可按如下方法在$O(n)$时间之内得到$T_1(T_2)$的每个非平凡子类到$T_2(T_1)$的距离(以及它的最相似子类)。

设$C_{1,i}\in\beta_*(T_1)$是T_1的一个非平凡子类$(1\leqslant i\leqslant|\beta_*(T_1)|)$，令$n_1$表示$C_{1,i}$中的标签的数目。设后序遍历$T_2$过程中，当前子类(节点)为$C_{2,j}$，$1\leqslant j\leqslant|\beta(T_2)|$。令$l_1$表示既在该子类中又在$C_{1,i}$中的标签的数目，$l_0$表示在该子类中但不在$C_{1,i}$中的标签的数目。则$C_{1,i}$和$C_{2,j}$之间的权重$wt(C_{1,i},C_{2,j})$为$l_0+n_1-l_1$，而$|C_{1,i}\cap C_{2,j}|=l_1$。在计算$C_{1,i}$与$T_2$的每个子类(不一定是非平凡的)的权重的过程中，取所得权重的最小者即为$d(C_{1,i},T_2)$，取得最小权重的那个子类即为$C_{1,i}$在T_2中的最相似子类$C'_{1,i}$。若$C'_{1,i}$不唯一，则取使l_1值最大的那个子类。

T_2 的叶子节点的 l_1 值和 l_0 值可在常数时间之内得到，内部节点的 l_1 值和 l_0 值可分别通过对该节点所有儿子的 l_1 值和 l_0 值求和得到。故可在 $O(n)$ 时间之内得到 T_1 的每个非平凡子类到 T_2 的距离以及它的最相似子类。

类似地，对 T_2 的每个非平凡子类 $C_{2,j} \in \beta_*(T_2)(1 \leqslant j \leqslant |\beta_*(T_2)|)$，按此方法在 $O(n)$ 时间之内得到 $d(C_{2,j}, T_1)$ 和 $C'_{2,j}$。

算法 2

第 1 步　设 f:L→{1,2,…,n}是任意一个双射。分别遍历 T_1 和 T_2，按照 f 修改叶子节点。

第 2 步　后序遍历 T_1。对所得到的 T_1 的每个非平凡子类 $C_{1,i} \in \beta_*(T_1)(1 \leqslant i \leqslant |\beta_*(T_1)|)$，按上述方法得到 $d(C_{1,i}, T_2)$ 和 $C'_{1,i}$。

第 3 步　后序遍历 T_2。对所得到的 T_2 的每个非平凡子类 $C_{2,j} \in \beta_*(T_2)(1 \leqslant j \leqslant |\beta_*(T_2)|)$，按上述方法得到 $d(C_{2,j}, T_1)$ 和 $C'_{2,j}$。

第 4 步　根据式(3)得到 T_1 和 T_2 之间的子类相异度 $Cdis(T_1, T_2)$。

定理 4　算法 2 的时间复杂度为 $O(n^2)$，其中 n 表示每棵树上的叶子数目。

5　5 个衍生指标

文献[25]所列出的 11 个指标中的前 6 个指标如表 1 所列。

表 1　Robinson-Foulds 距离及其衍生指标

指标	计算公式
$D(T_1, T_2)$	$\|\beta_*(T_1) \Delta \beta_*(T_2)\|$
$S(T_1, T_2)$	$\|\beta_*(T_1) \cap \beta_*(T_2)\|$
$d(T_1, T_2)$	$D(T_1, T_2)/[D(T_1, T_2)+S(T_1, T_2)]$
$s(T_1, T_2)$	$S(T_1, T_2)/[D(T_1, T_2)+S(T_1, T_2)]$
$d'(T_1, T_2)$	$D(T_1, T_2)/[D(T_1, T_2)+2 \cdot S(T_1, T_2)]$
$s'(T_1, T_2)$	$2 \cdot S(T_1, T_2)/[D(T_1, T_2)+2 \cdot S(T_1, T_2)]$

表 1 中的 $D(T_1, T_2)$ 和 $S(T_1, T_2)$ 分别是相似度和相异度，是非归一化(unnormalized)的基本指标，$S(T_1, T_2) = |\beta_*(T_1) \cup \beta_*(T_2)| - D(T_1, T_2)$。$d(T_1, T_2)$ 和 $s(T_1, T_2)$，以及 $d'(T_1, T_2)$ 和 $s'(T_1, T_2)$，分别是对 $D(T_1, T_2)$ 和 $S(T_1, T_2)$ 应用不同归一化策略的结果。

本文第 3 节所提出的子类相异度推广了表 1 中的第一个指标。类似于式(3)，将表 1 中的第二个指标推广如下：子类相似度定义为对所有子类与其最相似子类交集的基数求和除以 2。在此基础上，表 1 中的其余指标也相应得到了推广。

结束语　本文研究了有根系统发生树的比较问题，定义了子类相异度，研究了它的基本性质，给出了计算这个新的相异度的两个多项式时间的精确算法，并利用这些结果得到了 5 个衍生指标。后续研究可以继续研究子类相异度的其他性质，如分布、最小正值、邻域、灵敏度等；也可以尝试研究计算子类相异度的更有效的算法。

参 考 文 献

[1] Salemi M, Vandamme A-M, Lemey P. The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing [M]. Cambridge University Press, 2009

[2] 黄原. 分子系统发生学[M]. 北京：科学出版社，2012
Huang Yuan. Molecular Phylogenetics [M]. Beijing: Science Press, 2012

[3] Semple C, Steel M A. Phylogenetics [M]. Oxford University Press, 2003

[4] Bodlaender H L, Fellows M R, Warnow T J. Two strikes against perfect phylogeny [M]. Springer, 1992

[5] Sokal R R, Michener C D. A statistical method for evaluating systematic relationships[J]. University of Kansas Scientific Bulletin, 1958, 38: 1409-1438

[6] 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

[7] Fitch W M, Margoliash E. Construction of phylogenetic trees [J]. Science, 1967, 155(760): 279-284

[8] Day W H, Johnson D S, Sankoff D. The computational complexity of inferring rooted phylogenies by parsimony [J]. Mathematical Biosciences, 1986, 81(1): 33-42

[9] Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood [J]. Systema-tic Biology, 2003, 52(5): 696-704

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

[11] Huelsenbeck J P, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees [J]. Bioinformatics, 2001, 17(8): 754-755

[12] Drummond A J, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees [J]. BMC Evolutionary Biology, 2007, 7(1): 214

[13] 李军令，赵宏伟，马志强，等. 基于遗传算法的最大似然法构建系统发生树[J]. 东北师大学报(自然科学版)，2008，40(1)：36-39
Li Jun-ling, Zhao Hong-wei, Ma Zhi-qiang, et al. A maximum likehoood method based on genetic algorithms for constructing phylogenetic trees[J]. Journal of Northeast Normal University (Natural Science Edition), 2008, 40(1): 36-39

[14] 赵建邦，高琳，宋佳. 一种基于代谢路径构建系统发生树的有效方法[J]. 电子学报，2009，37(8)：1633-1638
Zhao Jian-bang, Gao Lin, Song Jia. An efficient method for constructing phylogenetic trees based on metabolic pathway[J]. Acta Electronica Sinica, 2009, 37(8): 1633-1638

[15] 刘红梅，刘国庆. 基于 k-mer 组分信息的系统发生树构建方法[J]. 生物信息学，2013，11(2)：100-104
Liu Hong-mei, Liu Guo-qing. A method for constructing phylogenetic trees based on k-mer component information[J]. China Journal of Bioinformatics, 2013, 11(2): 100-104

[16] 张树波，赖剑煌. 分子系统发育分析的生物信息学方法[J]. 计算机科学，2010，37(8)：47-51
Zhang Shu-bo, Lai Jian-huang. Bioinformatics approach for molecular evolution research[J]. Computer Science, 2010, 37(8): 47-51

[17] 张丽娜，荣昌鹤，何远，等. 常用系统发育树构建算法和软件鸟瞰[J]. 动物学研究，2013，34(6)：640-650
Zhang Li-na, Rong Chang-he, He Yuan, et al. A bird's eye view of the algorithms and software packages for reconstructing phylogenetic trees[J]. Zoological Research, 2013, 34(6): 640-650

[18] Stockham C, Wang L-S, Warnow T. Statistically based postprocessing of phylogenetic analysis by clustering [J]. Bioinformatics, 2002, 18(Suppl 1): 285-293

[19] Wang J T, Shan H, Shasha D, et al. Fast structural search in phylogenetic databases [J]. Evolutionary Bioinformatics, 2005, 21(1): 37-46

[20] de Vienne D M, Giraud T, Martin O C. A congruence index for testing topological similarity between trees [J]. Bioinformatics, 2007, 23(23): 3119-3124

[21] Pompei S, Loreto V, Tria F. On the accuracy of language trees [J]. PloS one, 2011, 6(6): e20109

[22] Hayes M, Walenstein A, Lakhotia A. Evaluation of malware phylogeny modelling systems using automated variant generation [J]. Journal in Computer Virology, 2009, 5(4): 335-343

[23] Swofford D L. When are phylogeny estimates from molecular and morphological data incongruent [M]// Phylogenetic Analysis of DNA Sequences. 1991: 295-333

[24] Bryant D. Building trees, hunting for trees, and comparing trees: theory and methods in phylogenetic analysis [D]. Dept. of Math., Univ. of Canterbury, 1997

[25] Day W H. Optimal algorithms for comparing trees with labeled leaves [J]. Journal of Classification, 1985, 2(1): 7-28

[26] Robinson D, Foulds L R. Comparison of phylogenetic trees [J]. Mathematical Biosciences, 1981, 53(1): 131-147

[27] Steel M A, Penny D. Distributions of tree comparison metrics—some new results [J]. Systematic Biology, 1993, 42(2): 126-141

[28] Bogdanowicz D, Giaro K. On a Matching Distance Between Rooted Phylogenetic Trees [J]. International Journal of Applied Mathematics and Computer Science, 2013, 23(3): 669-684

[29] Gabow H N, Tarjan R E. Faster scaling algorithms for network problems [J]. SIAM Journal on Computing, 1989, 18(5): 1013-1036

[30] Bogdanowicz D, Giaro K. Matching split distance for unrooted binary phylogenetic trees [J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2012, 9(1): 150-160

[31] Lin Y, Rajan V, Moret B M. A metric for phylogenetic trees based on matching [J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2012, 9(4): 1014-1022

(上接第 271 页)

确率大幅下降的情况；在 10%和 40%噪声条件下，C4.5 算法预测准确率下降幅度最大达到了 14.6%，下降幅度较大，而 NFPCA-in-C4.5 算法预测准确率下降幅度始终保持在 3.5%以下，下降程度明显低于 C4.5 算法。因此实验结果说明了在不同程度的噪声环境中，NFPCA-in-C4.5 算法对高维数据的预测准确率具有稳定性，具备了容噪的特性。

结束语 本文提出的 NFPCA-in-C4.5 算法对含噪声的高维数据兼具降维和容噪功能，避免了降维信息损失和噪声残留造成准确率大幅降低的问题；在不同程度噪声水平下，在高维数据集上进行了 C4.5 算法和 NFPCA-in-C4.5 的算法对比实验，结果表明在噪声水平高的环境下，NFPCA-in-C4.5 算法预测准确率的稳定性得到大幅提高，体现了 NFPCA-in-C4.5 算法对高维数据的容噪特性优势。

完成决策树的构建后，所形成的决策规则由各主成分组成，后续算法将研究决策树规则中各主成分不易理解的问题。

参 考 文 献

[1] 杨风召. 高维数据挖掘中若干关键问题的研究[D]. 上海：复旦大学，2003
Yang Feng-zhao. The Research on A Few Key Issues in High Dimensional Data Mining[D]. Shanghai: Fudan University, 2003

[2] 承文俊，沈建强，谢琪，等. 容噪学习机制及其在 Robocup 中的应用研究[J]. 计算机科学，2004，32(4)：101-103
Cheng Wen-jun, Shen Jian-qiang, Xie Qi, et al. Research on Noise Tolerance Mechanism in Robocup[J]. Computer Science, 2004, 32(4): 101-103

[3] 倪春鹏. 决策树在数据挖掘中若干问题的研究[D]. 天津：天津大学，2004
Ni Chun-peng. Research on Some Problems of Decision Tree in Data Mining[D]. Tianjin: Tianjin University, 2004

[4] Mantas C J, Abellán J. Credal-C4.5: Decision tree based on imprecise probabilities to classify noisy data[J]. Expert Systems with Applications, 2014, 41(10): 4625-4637

[5] 陈家俊，苏守宝，徐华丽. 基于多尺度粗糙集模型的决策树优化算法[J]. 计算机应用，2011，12：3243-3246
Chen Jia-jun, Su Shou-bao, Xu Hua-li. Decision tree optimization algorithm based on multiscale rough set model[J]. Computer Applications, 2011, 12: 3243-3246

[6] Breiman L, Friedman J, Stone C J, et al. Classification and regression trees[M]. CRC press, 1984

[7] 孟凡荣，蒋晓云，田恬，等. 基于主成分分析的决策树构造方法[J]. 小型微型计算机系统，2008(7)：1245-1249
Meng Fan-rong, Jiang Xiao-yun, Tian Tian, et al. Decision Tree Construction Method Based on Principal Component Analysis [J]. Journal of Chinese Computer Systems, 2008(7): 1245-1249

[8] Jolliffe I. Principal component analysis [M]. Wiley Online Library, 2005

[9] Rezghi M, Obulkasim A. Noise-free principal component analysis: An efficient dimension reduction technique for high dimensional molecular data[J]. Expert Systems with Applications, 2014, 41(17): 7797-7804

[10] Hotelling H. Analysis of a complex of statistical variables into principal components[J]. Journal of Educational Psychology, 1933, 24(6): 417

[11] 周斯斯. 谱聚类维数约简算法研究与应用[D]. 西安：西安电子科技大学，2010
Zhou Si-si. Spectral Clustering Based Dimensionality Reduction and Applications[D]. Xi'an: Xi'an University of Electronic Science and Technology, 2010

[12] Golub G. Matrix computations[M]. Johns Hopkins University Press, 1996

[13] Hanke M, Hansen P C. Regularization methods for large-scale problems[J]. Surv. Math. Ind, 1993, 3(4): 253-315

[14] 树方，平文. 数值线性代数[M]. 北京：北京大学出版社，2000
Shu Fang, Ping Wen. Numerical Linear Algebra[M]. Beijing: Beijing University Press, 2000

[15] Björck A. Numerical methods for least squares problems[M]. Siam, 1996

[16] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection[C]// International Joint Conference on Artifical Intelligence. 1995: 1137-1143

[17] 王越，万洪. 一种新的应用变精度粗糙集的决策树构造方法[J]. 重庆理工大学学报(自然科学版)，2013，27(11)：58-64
Wang Yue, Wan Hong. A New Method for Constructing Decision Tree Based on Variable Precision Rough Set[J]. Journal of Chongqing University of Technology (Natural Science), 2013, 28(11): 58-64