ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.17, No.7, July 2006, pp.1494−1500
DOI: 10.1360/jos171494

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

测试集问题的集合覆盖贪心算法的深入近似*

崔 鹏[1+], 刘红静[2]

[1](中国人民大学 信息资源管理学院,北京 100872)

[2](保定市财贸学校,河北 保定 071000)

Deep Approximation of Set Cover Greedy Algorithm for Test Set

CUI Peng[1+], LIU Hong-Jing[2]

[1](School of Information Resource Management, Renmin University of China, Beijing 100872, China)

[2](School of Finance and Trade, Baoding 071000, China)

+ Corresponding author: Phn: +86-10-62511458, E-mail: cuipeng@ruc.edu.cn

Cui P, Liu HJ. Deep approximation of set cover greedy algorithm for test set. *Journal of Software*, 2006,17(7): 1494−1500. http://www.jos.org.cn/1000-9825/17/1494.htm

Abstract: Test set problem is a NP-hard problem with wide applications. Set cover greedy algorithm is one of the commonly used algorithms for the test set problem. It is an open problem if the approximation ratio $2\ln n+1$ directly derived from the set cover problem can be improved. The generalization of set cover greedy algorithm is used to solve the redundant test set problem arising in bioinformatics. This paper analyzes the distribution of the times for which the item pairs are differentiated, and proves that the approximation ratio of the set cover greedy algorithm for test set can be $1.5\ln n+0.5\ln\ln n+2$ by derandomization method, thus shrinks the gap in the analysis of the approximation ratio of this algorithm. In addition, this paper shows the tight lower bound $(2-o(1))\ln n-\Theta(1)$ of the approximation ratio of set cover greedy algorithm for the weighted redundant test set with redundancy $n-1$.
Key words: test set; set cover greedy algorithm; derandomization method; redundant test set

摘 要: 测试集问题是一个有着广泛应用的NP难问题.集合覆盖贪心算法是测试集问题的一个常用近似算法,其由集合覆盖问题得到的近似比 $2\ln n+1$ 能否改进是一个公开的问题.集合覆盖贪心算法的推广被用来求解生物信息学中出现的冗余测试集问题.通过分析条目对被区分次数的分布情况,用去随机方法证明了集合覆盖贪心算法对测试集问题的近似比可以为 $1.5\ln n+0.5\ln\ln n+2$,从而缩小了这种算法近似比分析的间隙.另外,给出了集合覆盖贪心算法对冗余度为 $n-1$ 的加权冗余测试集问题的近似比的紧密下界$(2-o(1))\ln n-\Theta(1)$.
关键词: 测试集问题;集合覆盖贪心算法;去随机方法;冗余测试集问题
中图法分类号: TP301 文献标识码: A

测试集问题(minimum test set,简称 MTS)出现于模式识别、VLSI 测试、生物信息学等领域中.测试集问题是 NP 难的[1].在实际计算中,经常使用的近似算法是贪心算法.贪心算法按照选择测试的标准可以分为集合覆盖类型或信息类型[2].对其中信息类型的贪心算法,文献[2]没有给出精确的近似比.文献[3]设计了一种新的信息

Received 2005-10-24; Accepted 2006-01-09

贪心算法,并证明其具有近似比 $\ln n+1$.由于 MTS 的近似难度是除非 $NP\subseteq DTIME(n^{\log\log n})$,对任意 $\varepsilon>0$,不能近似到 $(1-\varepsilon)\ln n$[3,4],这个算法在主项上具有最好可能的近似比.

集合覆盖贪心算法(set cover greedy algorithm,简称 SGA)是测试集问题的一种常用算法.在实际计算中,解的大小与信息类型贪心算法相当[2,5].MTS 可以自然地归约为集合覆盖问题(minimum set cover,简称 MSC),从而得到 SGA 的近似比 $2\ln n+1$.反过来,利用一个 MSC 到 MTS 的规约,可以构造 MTS 的实例,使得 SGA 在这个实例上近似比至少为 $\ln n$[4].文献[6]猜测 SGA 的近似比可以为 $\ln n+o(1)$.文献[4]讨论了 MTS 限制测试大小不超过常数 k 的情况,给出了一种有近似比 $O(\log k)$的两阶段算法,并在 $k=2$ 的情况下给出了 SGA 的紧密近似比 11/8.

在算法分析与设计领域,基于条件概率的去随机方法常用来从概率证明中得到确定性算法.文献[7]讨论了一种去随机方法,给出了 MSC 的贪心算法的近似比 $\ln N+1$ 的另一种证明.

本文通过对条目对被区分次数分布情况的分析,用类似于文献[7]的去随机方法证明了 SGA 的近似比可以为 $1.5\ln n+0.5\ln\ln n+2$,从而缩小了这种算法在近似比分析上的间隙.

在生物信息学的探针选择问题中,由于杂交噪声的存在,往往需要多个探针区分一段序列[8].这就引出了冗余测试集问题(redundant minimum test set,简称 RMTS).RMTS 可以看作多覆盖的 MSC 的一种特殊情况.SGA 是 RMTS 的一种常用算法[5].另外,文献[9]对多覆盖的 MSC 设计了一种随机的多步取整算法(简称 RND),其近似比的期望值近似地不超过 $\ln N-\ln M$,其中 N 是底集大小,M 是覆盖次数.对 RMTS 的随机实验表明:当冗余度较小时,SGA 返回解的大小优于 RND;当冗余度较大(接近或大于 n)时,RND 返回解的大小优于 SGA[5].

文献[9]讨论了多覆盖的 MSC 问题当覆盖度 $r=N-c$ 时的情况,其中 N 是底集大小,c 是一个常数.可以证明此时贪心算法的近似比至少为 $(0.5-o(1))\left(\frac{r}{8N-2}\right)\log_2 N$.这个近似比的下界与用对偶拟合方法得到的近似比 $\ln N+1$ 之间还有一个常数的间隙.

本文讨论了 SGA 对 RMTS 的近似比的下界.为简洁起见,本文对加权的 RMTS 进行了分析,证明了 SGA 对冗余度为 $n-1$ 的加权 RMTS 的近似比至少为$(2-o(1))\ln n-\Theta(1)$,从而说明在这种情况下近似比 $2\ln n+1$ 是紧密的.

本文第 1 节是背景介绍和一些定义.第 2 节证明 SGA 对 MTS 的改进近似比.第 3 节给出 SGA 对加权 RMTS 的结果.第 4 节是结束语.

1 背 景

对最小化问题的一个近似算法 A 和实例 I,用 $m_A(I)$表示算法 A 在 I 上返回解的大小,用 $m^*(I)$表示 I 的最优解的大小.称 $m_A(I)/m^*(I)$是 A 在 I 上的近似比.用$|I|$表示实例 I 的大小.对一个定义在 Z^+上的非降函数 $r(|I|)$,称 $r(|I|)$是 A 的近似比,若存在 $n_0\in Z^+$,则对任意满足$|I|>n_0$ 的实例 I,都有 $m_A(I)/m^*(I)\le r(|I|)$.

MTS 的输入是一个条目的集合 S,$|S|=n$,以及 S 的子集(称为 S 的测试)的一个集合 T.条目对$\{i,j\}$是两个不同条目 i 和 j 的集合.一个测试 T 区分条目对$\{i,j\}$(或$\{i,j\}$被 T 区分),若 i 和 j 恰好有一个属于 T,即$|T\cap\{i,j\}|=1$.测试的一个集合 $T'\subseteq T$ 是一个 S 的测试集,若每一个条目对被 T'中至少一个测试区分(简称这个条目被 T'区分).目标是找最小的测试集.为保证公式有意义,本文规定 $n\ge 2$.S 的不同条目对的总数是 $n(n-1)/2$.令 i 和 j 是两个不同的条目,S_1 和 S_2 是 S 的两个不相交的子集.若 $i,j\in S_1$,则称$\{i,j\}$是 S_1 内部的条目对.若 $i\in S_1$ 和 $j\in S_2$,则称$\{i,j\}$是 S_1 和 S_2 之间的条目对.

RMTS 是 MTS 的推广:$T'\subseteq T$ 是一个 S 的 r-测试集,若每一个条目对被 T'中至少 r 个不同测试区分,则 r 是一个正整数,称为冗余度,目标是找最小的 r-测试集.

用$\{i,j\}\perp T$ 表示 T 区分$\{i,j\}$,用$\{i,j\}\perp T'$表示 T'区分$\{i,j\}$;用$\{i,j\}//T$ 表示 T 不区分$\{i,j\}$,用$\{i,j\}//T'$表示 T'不区分$\{i,j\}$;用$\{i,j\}\perp^1 T'$表示$\{i,j\}$仅被一个 T'中的测试区分;用$\{i,j\}\perp^2 T'$表示$\{i,j\}$至少被两个 T'中的测试区分.

引理 **1**. 设 i,j 和 k 是 3 个不同的条目,若$\{i,j\}//T'$和$\{i,k\}//T'$,则$\{j,k\}//T'$.

证明:否则,$\{j,k\}\perp T'$.设对 $T\in T'$,$\{j,k\}\perp T$.考虑两种情况:(1) $i\in T$;(2) $i\notin T$.

在情况(1),若 $j\in T$ 和 $k\notin T$,则 $\{i,k\}\perp T'$;否则 $j\notin T$ 和 $k\in T$,则 $\{i,j\}\perp T'$.这都与引理假设矛盾.类似地,情况(2)也引起矛盾. □

对 $\bar{T}\subseteq T$,定义一个 S 上的二元关系 $\sim_{\bar{T}}$ 为:对任意 i 和 j,$i\sim_{T'} j$ 当且仅当 $\{i,j\}\not\perp\bar{T}$.由引理 1,$\sim_{\bar{T}}$ 是等价关系.$\sim_{\bar{T}}$ 的等价类的集合记为 $P_{\bar{T}}$,称为 $\bar{T}$ 对 S 的划分.$\bar{T}$ 区分 $\sim_{\bar{T}}$ 的等价类之间的条目对,而不区分 $\sim_{\bar{T}}$ 的等价类内部的条目对.

信息贪心算法是:每次选择一个测试,使得这个所选测试子集对 S 的划分的信息内容函数最大(平局情况任意选择),直到所有的条目对被区分.

算法 **1**. 信息贪心算法.

输入:S,T;

输出:S 的测试集.

(1) $\bar{T}\leftarrow\varnothing$;

(2) 在 $T-\bar{T}$ 中选择 T,使得 $IC(\bar{T}\cup\{T\})$ 最小;

(3) $\bar{T}\leftarrow\bar{T}\cup\{T\}$;

(4) 如果 $\bar{T}$ 是测试集;返回 $\bar{T}$,否则转(2).

其中文献[2]使用的信息内容函数是

$$IC(\bar{T}\cup\{T\})=\frac{1}{n}\sum_{p\in P_{\bar{T}\cup\{T\}}}|p|\log_2|p| \tag{1}$$

而文献[3]使用的信息内容函数是

$$IC(\bar{T}\cup\{T\})=\sum_{p\in P_{\bar{T}\cup\{T\}}}\log_2(|p|!) \tag{2}$$

MSC 是一个得到大量研究的 NP 难问题.MSC 的实例由底集 U,U 的子集的集合 C 组成.$|U|=N$.U 的一个集合覆盖是 C 的一个子集 C',使得 U 的每个元素至少属于 C'的一个成员(或称为 U 的这个元素被 C'的这个成员覆盖).目标是找最小的集合覆盖.MSC 的常用算法——贪心算法可叙述为:每次选择一个子集,使得这个子集覆盖的底集的元素个数最大(平局情况任意选择),直到底集所有的元素被覆盖.

多覆盖的 MSC 是 MSC 的推广:$C'\subseteq C$ 是一个 U 的 M-集合覆盖,如 U 的每一个元素被 C'中至少 M 个不同成员区分,M 是一个正整数,称为覆盖次数.目标是找最小的 M-集合覆盖.

设(S,T)是 MTS 的一个实例,我们可以把它自然地归约为一个 MSC 的实例(U,C),其中:

$$\begin{aligned}U&=\{\{a,b\}\mid a,b\in S,a\neq b\}\\C&=\{c(T)\mid T\in\mathsf{T}\},c(T)=\{\{a,b\}\mid a\in T,b\in S-T\}\end{aligned} \tag{3}$$

显然,(S,T)与(U,C)的最优解的大小相等;SGA 在(S,T)上返回解的大小恰好与 MSC 的贪心算法在(U,C)上返回解的大小相等.

定义 $\bar{T}$ 的区分度量 $\#(\bar{T})$ 为不被 $\bar{T}$ 区分的条目对数,T 相对于 $\bar{T}$ 的区分度量是 $\#(T,\bar{T})=\#(\bar{T})-\#(\bar{T}\cup\{T\})$.

SGA 可以叙述为:每次选择一个测试,使得这个测试区分的条目对数最大(平局情况任意选择),直到所有的条目对被区分.

算法 **2**. SGA.

输入:S,T;

输出:S 的测试集.

(1) $\bar{T}\leftarrow\varnothing$;

(2) 在 $T-\bar{T}$ 中选择 T,使得 $\#(\bar{T}\cup\{T\})$ 最小;

(3) $\bar{T}\leftarrow\bar{T}\cup\{T\}$;

(4) 如果 $\bar{T}$ 是测试集,返回 $\bar{T}$;否则转(2).

算法 2 中的 $\bar{T}$ 称为部分测试集.因为 MSC 的贪心算法有紧密的近似比 H_N[10].根据 MTS 到 MSC 的自然规

约,SGA 对 MTS 有近似比 $H_{n(n-1)/2}<2\ln n+1$.

2 主要结果

本节首先给出两个引理,讨论条目对被区分次数的分布情况,然后用去随机方法证明 SGA 的改进近似比为 $1.5\ln n+0.5\ln\ln n+2$.

引理 **2**. 对一个测试集 $T'\subseteq T$ 和 $T\in T'$,最多 $\min(|T|,|S-T|)$个 T 和 $S-T$ 之间的条目对不被 $T'-\{T\}$区分.

证明:不失一般性,假设$|T|\le|S-T|$.我们宣称:对任意 $i\in T$,最多存在 $S-T$ 中的一个条目 j,使得$\{i,j\}//T'-\{T\}$.因为:若存在 $S-T$ 中的两个不同条目 j 和 k,使得$\{i,j\}//T'-\{T\}$和$\{i,k\}//T'-\{T\}$,则由引理 1,$\{j,k\}//T'-\{T\}$.又$\{j,k\}//\{T\}$,所以$\{j,k\}//T'$.这与 T'是测试集矛盾. □

引理 **3**. 对测试集 $T'\subseteq T$,至多 $n\log_2 n$ 个条目对仅被一个 T'中的测试区分.

证明:用 F 表示仅被一个 T'中测试区分的条目集合.我们用数学归纳法证明$|F|\le n\log_2 n$.$n=1$ 时,$|F|=0=n\log_2 n$.假设引理对任意 $n\le k-1$ 成立,下面证明引理对 $n=k$ 也成立.

任选 $T\in T'$,则$|T|\le k-1$,$|S-T|\le k-1$.实例的所有条目对可以分为 3 个不相交的部分:T 内部的条目对;$S-T$ 内部的条目对;T 和 $S-T$ 之间的条目对.

显然,一个测试 $T'\in T'-\{T\}$区分 T 内部的一个测试当且仅当 $T'\cap T$ 区分这个测试.所以,$T_T=\{T'\cap T|T'\in T'-\{T\}\}$是 T 的一个测试集.根据归纳假设,在实例(T,T_T)中,至多$|T|\log_2|T|$个条目对仅被一个 T_T 中的测试区分,而 T 内部的条目对不被 T 区分.因此,在实例 (S,T) 中,至多$|T|\log_2|T|$个 T 内部的条目对仅被一个 T'中的测试区分.类似地,至多$|S-T|\log_2|S-T|$个 $S-T$ 内部的条目对仅被一个 T'中的测试区分.

根据引理 2,至多 $\min(|T|,|S-T|)$个 T 和 $S-T$ 之间的条目对不被 $T'-\{T\}$区分.因为 T 和 $S-T$ 之间的所有条目对被 T 区分.至多 $\min(|T|,|S-T|)$个 T 和 $S-T$ 之间的条目对仅被一个 T'中的测试区分.

这样,

$$|F|\le|T|\log_2|T|+|S-T|\log_2|S-T|+\min(|T|,|S-T|) \tag{4}$$

不失一般性,假设$|T|\le|S-T|$,

$$\begin{aligned}|F|&\le|T|\log_2|T|+|S-T|\log_2|S-T|+|T|\\&\le|T|\log_2(2|T|)+|S-T|\log_2|S-T|\\&\le|T|\log_2|S|+|S-T|\log_2|S|\\&=|S|\log_2|S|\end{aligned} \tag{5}$$

□

定理 **1**. SGA 的近似比可以为 $1.5\ln n+0.5\ln\ln n+2$.

证明:令 $\#_0=n(n-1)/2$,$\#_F=n\log_2 n$.显然有部分测试集 T_1 满足$\#(T_1)\ge\#_F$,但当算法选择下一个测试 $\tilde{T}$ 后,$\#(T_1\cup\{\tilde{T}\})<\#_F$.

令算法在选择 $\tilde{T}$ 后直到结束选择的测试集合是 T_2,T^*是一个最优测试集,$l=|T^*|/2\cdot\ln\frac{2\#_0}{\#_F}$.不失一般性,假设$|T^*|>2$.

给定 $\bar{T}$,定义 $f(\bar{T})=(\#(\bar{T})-\#_F/2)(1-2/|T^*|)^{l-|\bar{T}|}$.

因为对任意 $0<x\le1$,$(1-x)^{1/x}<1/e$,

$$\begin{aligned}f(\varnothing)&\le(\#_0-\#_F/2)(1-2/|T^*|)^{|T^*|/2\cdot\ln\frac{2\#_0}{\#_F}}\\&<\#_0/(2\#_0/\#_F)=\#_F/2\end{aligned} \tag{6}$$

假设我们按照平均分布从 T^*中挑出一个测试.对 T^*中的任意测试 T,挑出 T 的概率是 $p^*(T)=1/|T^*|$,并且

$$\sum_{T\in T^*}p^*(T)=1.$$

对任意条目对 $\{i,j\}//\bar{T}$,一定有$\{i,j\}\perp T^*$,所以

$$\sum_{T:T\in T^*,\{i,j\}\perp T} p^*(T)\ge 1/|T^*| \tag{7}$$

若$\{i,j\}\perp^2 T^*$,则有

$$\sum_{T:T\in T^*,\{i,j\}\perp^2 T} p^*(T)\ge 2/|T^*| \tag{8}$$

根据 $f(\bar{T})$ 的定义以及 $p^*(T)\ge 0$ 和 $\sum_{T\in T^*} p^*(T)=1$,我们有

$$\begin{aligned}
&\min_{T\in T} f(\bar{T}\cup\{T\})\\
&\le \min_{T\in T^*} f(\bar{T}\cup\{T\})\\
&\le \sum_{T\in T^*} p^*(T) f(\bar{T}\cup\{T\})\\
&= (1-2/|T^*|)^{l-|\bar{T}|-1}\sum_{T\in T^*} p^*(T)(\#(\bar{T}\cup\{T\})-\#_F/2)\\
&= (1-2/|T^*|)^{l-|\bar{T}|-1}\left(\#(\bar{T})-\sum_{T\in T^*} p^*(T)(\#(T,\bar{T})-\#_F/2)\right)\\
&\le (1-2/|T^*|)^{l-|\bar{T}|-1}\left(\sum_{\{i,j\}//\bar{T}}\left(1-\sum_{T:T\in T^*,\{i,j\}\perp T} p^*(T)\right)-\#_F/2\right)
\end{aligned} \tag{9}$$

和

$$\begin{aligned}
&\sum_{\{i,j\}//\bar{T}}\left(1-\sum_{T:T\in T^*,\{i,j\}\perp T} p^*(T)\right)-\#_F/2\\
&=\sum_{\{i,j\}//\bar{T},\{i,j\}\perp^1 T^*}\left(1-\sum_{T:T\in T^*,\{i,j\}\perp T} p^*(T)\right)+\sum_{\substack{\{i,j\}//\bar{T},\{i,j\}\perp^2 T^*\\ T\in T}}\left(1-\sum_{T:T\in T^*,\{i,j\}\perp T} p^*(T)\right)-\#_F/2\\
&\le\sum_{\{i,j\}//\bar{T},\{i,j\}\perp^1 T^*}\left(1-\frac{1}{|T^*|}\right)+\sum_{\{i,j\}//\bar{T},\{i,j\}\perp^2 T^*}\left(1-\frac{2}{|T^*|}\right)-\#_F/2\\
&=\sum_{\{i,j\}//\bar{T},\{i,j\}\perp^1 T^*}\left(1-\frac{2}{|T^*|}\right)+\sum_{\{i,j\}//\bar{T},\{i,j\}\perp^2 T^*}\left(1-\frac{2}{|T^*|}\right)+\sum_{\{i,j\}//\bar{T},\{i,j\}\perp^1 T^*}\frac{1}{|T^*|}-\#_F/2\\
&\le\#(\bar{T})\left(1-\frac{2}{|T^*|}\right)+\#_F\frac{1}{|T^*|}-\#_F/2\\
&=(\#(\bar{T})-\#_F/2)\left(1-\frac{2}{|T^*|}\right)
\end{aligned} \tag{10}$$

其中第 2 个不等式的根据是引理 3.所以,

$$\begin{aligned}
&\min_{T\in T} f(\bar{T}\cup\{T\})\\
&\le(\#(\bar{T})-\#_F/2)(1-2/|T^*|)(1-2/|T^*|)^{l-|\bar{T}|-1}\\
&=f(\bar{T})
\end{aligned} \tag{11}$$

算法开始时,部分测试集是∅.对部分测试集 $\bar{T}$,算法选择 T,使得 $\#(\bar{T}\cup\{T\})$ 最小,即使得 $f(\bar{T}\cup\{T\})$ 最小.所以,$f(T_1)\le f(\emptyset)<\#_F/2$.

另一方面,由 T_1 的定义,

$$\begin{aligned}
f(T_1)&\ge(\#_F-\#_F/2)(1-2/|T^*|)^{l-|T_1|}\\
&=\#_F/2\cdot(1-2/|T^*|)^{l-|T_1|}
\end{aligned} \tag{12}$$

所以,$(1-2/|T^*|)^{l-|T_1|}<1$,这样$|T_1|<l=\frac{1}{2}\ln\frac{2\#_0}{\#_F}|T^*|$. □

算法选择 T_1 和 $\tilde{T}$ 后,构造如下的 MSC 实例(U_2,C_2):$U_2=\{\{i,j\}|\{i,j\}//\ T_1\cup\{\tilde{T}\}\}$, $c_2=\{c(T)|T:T\in T-T_1-\{\tilde{T}\}$,存在$\{i,j\}\in U_2$,使得$\{i,j\}\perp T\}$,$c(T)=\{\{i,j\}|\{i,j\}\in U_2,\{i,j\}\perp T\}$.

令 C^*是(U_2,C_2)的最优集合覆盖,MSC 的贪心算法在(U_2,C_2)返回解是 C'.由 T_1 和 $\tilde{T}$ 的定义,$|U_2|<\#_F$.由于 T^*

区分 U_2 中的所有条目对,$|C^*|\leq|T^*|$.

所以,

$$|T_2|+|C'|\leq(\ln\#_F+1)|C^*|\leq(\ln\#_F+1)|T^*| \tag{13}$$

算法结束时返回解的大小为

$$\begin{aligned}&|T_1|+|T_2|+1\\<&\left(\frac{1}{2}\ln\frac{2\#_0}{\#_F}+\ln\#_F+1\right)|T^*|+1\\<&(1.5\ln n+0.5\ln\ln n+2)|T^*|\end{aligned} \tag{14}$$

3 冗余测试集

加权的多覆盖 MSC 是:每个 $c\in C$ 有一个权 $w(c)\in Q^+$,解的度量是 M-集合覆盖的成员的权的总和.它的贪心算法是:称一个元素是活的,若它被已选子集覆盖的次数少于 M.每一步在没有选择子集中选择价格最小的子集.一个子集的价格定义为这个子集的权与当前它包含的活元素个数之比.若没有活元素,则算法结束.可以用对偶拟合方法证明,加权的多覆盖 MSC 的贪心算法有近似比 H_N[11].

加权 RMTS 是:每个 $T\in T$ 有一个权 $w(T)\in Q^+$,解的度量是 r-测试集中测试的权的和.SGA 可以直接推广到加权 RMTS 上:称一个条目对是活的,若它被已选测试区分的次数少于 r.每一步在没有选择测试中选择价格最小的测试.测试的价格定义为这个测试的权与当前它区分的活条目对个数之比.若没有活条目对,则算法结束.

加权 RMTS 可以自然地归约到加权的多覆盖 MSC.由此,SGA 对加权 RMTS 有近似比 $H_{n(n-1)/2}$.

定理 **2**. 存在加权 RMTS 的实例,SGA 在这个实例上的近似比至少为$(2-o(1))\ln n-\Theta(1)$.

证明:令 $n=2n'+1,r=2n',S=\{a_1,a_2,\ldots,a_n\},T=\{T_0\}\cup T_1\cup T_2,T_1=\{T_1,T_2,\ldots,T_n\},T_2=\{T_{i,j}|1\leq i\leq n',n'+1\leq j\leq 2n'+1\}$, $T_0=\{a_1,a_2,\ldots,a_{n'}\},T_i=\{a_i\},1\leq i\leq n,T_{i,j}=\{a_i,a_j\},1\leq i\leq n',n'+1\leq j\leq 2n'+1,w(T_0)=1$, $w(T_i)=\dfrac{1}{n'-i+1},1\leq i\leq n$, $w(T_{n'+1})=\alpha(n')$, $w(T_i)=\dfrac{1}{2n'-i+2},n'+2\leq i\leq 2n'+1,w(T_{i,j})=\beta(n'),1\leq i\leq n',n'+1\leq j\leq 2n'+1$,其中$\alpha(n')$和$\beta(n')$是充分小的正有理数,满足 $\alpha(n')<\dfrac{1}{n'(n'+1)}$ 和 $\beta(n')<\dfrac{\alpha(n')}{2n'}$.

每个条目对 $\{a_{i_1},a_{i_2}\},1\leq i_1\leq i_2\leq n'$ 被 $2(n'+1)$个测试 $T_{i_d,j}=\{a_{i_d},a_j\},n'+1\leq j\leq 2n'+1,d=1,2$ 区分,每个条目对 $\{a_{j_1},a_{j_2}\},n'+1\leq j_1<j_2\leq 2n'+1$ 被 $2n'$个测试 $T_{i,j_d}=\{a_i,a_{j_d}\},1\leq i\leq n',d=1,2$ 区分,每个条目对$\{a_i,a_j\},1\leq i\leq n',n'+1\leq j\leq 2n'+1$ 被 $n'-1$ 个测试 $T_{i',j}=\{a_{i'},a_j\},1\leq i'\leq n',i'\neq i$ 以及 n'个测试 $T_{i,j'}=\{a_i,a_{j'}\},n'+1\leq j'\leq 2n'+1,j'\neq j$ 区分.

由于 $\beta(n')<\dfrac{1}{2n'}\cdot\dfrac{1}{n'(n'+1)}$,算法首先在 T_2 中选择测试.由于区分条目对 $\{a_{j_1},a_{j_2}\},n'+1\leq j_1<j_2\leq 2n'+1$ 达到 $2n'$ 次需要选择所有 T_2 中的测试,算法将依次选择所有 T_2 中的测试.此时,条目对 $\{a_{i_1},a_{i_2}\},1\leq i_1\leq i_2\leq n'$和条目对 $\{a_{j_1},a_{j_2}\},n'+1\leq j_1<j_2\leq 2n'+1$ 都被区分 $2n'$次,不再是活的;而条目对$\{a_i,a_j\},1\leq i\leq n',n'+1\leq j\leq 2n'+1$ 被区分 $2n'-1$ 次,距离指定次数还差 1 次.

由于 $\alpha(n')<\dfrac{1}{n'(n'+1)}$,算法接下来选择 $T_{n'+1}$.不难证明,算法随后可能按照 T_1, $T_{n'+2}$, T_2, $T_{n'+3}$,直到 $T_{n'-1}$, $T_{2n'}$, $T_{n'}$ 的顺序选择测试.得到解 $T'=\{T_1,\ldots,T_{n-1}\}\cup T_2$.

T'中测试的权的和为 $2H_{n'}-1+\alpha(n')+n'(n'+1)\beta(n')$.因为$\{T_0\}\cup T_2$ 也是问题的解,其中测试的权的和为$1+n'(n'+1)\beta(n')$.算法在这个实例上的近似比不小于

$$\frac{2H_{n'}-1+\alpha(n')+n'(n'+1)\beta(n')}{1+n'(n'+1)\beta(n')}=(2-o(1))\ln n-\Theta(1)$$

□

4 结 论

本文给出了目前 SGA 的最佳近似比.本文的证明可以推广到加权的情况,并证明同样的近似比.SGA 是否具有与信息贪心算法在主项上同样的近似比,仍是一个有待解决的问题.

本文解决了 SGA 对冗余度至少为 $n-1$ 的加权 RMTS 的紧密分析,即由对偶拟合方法得到的近似比 $2\ln n+1$ 不能再改进.对冗余度为 n^{α}的加权 RMTS,$0<\alpha<1$,可以类似地构造实例,使得 SGA 在其上的近似比至少为 $(2-o(1))\alpha\ln n-\Theta(1)$.不加权 RMTS 的相应实例可以通过 MTS 的重复-分割技巧[4]来构造.

致谢 我们感谢 B. DasGupta 和 B. V. Halldorsson 提供的有益评论.

References:

[1] Garey MR, Johnson DS. Computers and Intractability: A Guide to the Theory of NP-Completeness. San Francisco: W. H. Freeman, 1979. 71−72.

[2] Moret BME, Shipiro HD. On minimizing a set of tests. SIAM Journal on Scientific and Statistical Computing, 1985,6(4):983−1003.

[3] Berman P, DasGupta B, Kao M. Tight approximability results for test set problems in bioinformatics. Journal of Computer and System Sciences, 2005,71(2):145−162.

[4] De Bontridder KMJ, Halldorsson BV, Halldorsson MM, Hurkens CAJ, Lenstra JK, Ravi R, Stougie L. Approximation algorithm for the test cover problems. Mathematical Programming-B, 2003,98(1−3):477−491.

[5] DasGupta B, Konwar K, Mandoiu I, Shvartsman A. Highly scalable algorithms for robust string barcoding. Int'l Journal of Bioinformatics Research and Applications, 2005,1(2):145−161.

[6] Halldorsson BV. Algorithms for biological sequence problems [Ph.D. Thesis]. Pittsburgh: Carnegie Mellon University, 2001.

[7] Young NE. Greedy algorithms by derandomizing unknown distributioms. Technical Report, T.R.1087, Ithaca: Cornell University, 1994.

[8] Borneman J, Chrobak M, Vedova GD, Figueora A, Jiang T. Probe selection algorithms with applications in the analysis of microbial communities. Bioinformatics, 2001,17(Suppl.):S39−S48.

[9] Berman P, DasGupta B, Sontag E. Randomized approximation algorithms for set multicover problems with applications to reverse engineering of protein and gene networks. In: Proc. of the 7th Int'l Workshop on Approximation Algorithms for Combinatorial Optimization Problems (APPROX 2004). LNCS 3122, Berlin: Springer-Verlag, 2004. 39−50.

[10] Johnson DS. Approximation algorithms for combinatorial problems. Journal of Computer and System Sciences, 1974,9:256−278.

[11] Rajagopalan S, Vazirani VV. Primal-Dual RNC approxmation algorithms for set cover and covering integer programs. SIAM Journal on Computing, 1999,28(2):525−540.

崔鹏(1973–),男,河北保定人,博士,讲师,主要研究领域为计算机图形学.

刘红静(1975–),女,讲师,主要研究领域为信息管理.