概率计算：复杂性的统一度量初探

姚期智

计算机科学系

斯坦福大学

斯坦福，加利福尼亚州 **94305**

1. 引言

从理论和实践的角度来看，研究算法的期望运行时间是一个有趣的课题。基本上存在两种研究途径。第一种途径（我们称之为分布途径），假设问题的输入遵循某种“自然”分布，并在此假设下寻找快速算法（参见 Knuth [8]）。例如，在排序 n 个数时，通常假设所有 n! 种初始排列的可能性相等。对这种途径的一个常见批评是，现实生活中的分布千差万别；此外，输入的真实分布常常是未知的。最近有人提出了一种试图通过允许计算中的随机移动来克服这一缺点的替代途径。这就是由 Rabin [10] 推广的随机化途径（另见 Gill[5], Solovay and Strassen [13]），尽管这个概念对统计学家来说并不陌生（例如，参见 Luce and Raiffa [9]）。注意，通过在算法中允许随机移动，输入实际上被随机化了。我们将此类算法称为随机算法。

这两种途径自然导致了问题内在复杂性的两种不同定义，我们分别称之为分布复杂性和随机复杂性。（精确的定义和例子将在第 2 节和第 3 节给出。）为了巩固这些思想，我们研究可以用决策树建模的熟悉组合问题。特别地，我们考虑（a）从邻接矩阵测试任意图属性（第 2 节），以及（b）关于 n 个数的偏序问题，包括排序、选择等（第 3 节）。我们将证明，对于这两类问题，通过著名的冯·诺依曼极小极大定理 [14]，这两种复杂性度量总是一致的。

这两种途径之间的联系使其能够应用于实际问题。通过两种不同的（在某种意义上互补的）视角来看待问题的复杂性，通常更容易推导出上界和下界。例如，对于使用邻接矩阵表示图的情况，可以证明任何随机算法无法在少于 $O(n^{2})$ 次探测的情况下确定完美匹配的存在性。之前此类随机化途径的下界是缺失的。作为另一个应用例子，我们可以证明对于（b）中的偏序问题，假设均匀分布（即所有 n! 种排列等可能）总是产生最大的复杂性。

我们也将考虑允许错误的算法（参见 Karp [6], Rabin[10]）。同样可以在分布复杂性和随机复杂性之间建立有用的联系（尽管在这种情况下不是相等）。例如，上文提到的完美匹配问题的 $O(n^{2})$ 下界仍然成立，即使我们允许随机算法出错，例如，出错概率为 5%。

由于讨论低阶计算复杂性必须基于精确的算法模型，我们将在第 2 节和第 3 节分别处理（a）类和（b）类问题。进一步的研究方向将在第 4 节讨论。

2. 测试图属性

2.1 符号与定义

令 g_{n} 为 n 个顶点上所有无向图的集合。假设 P 是 g_{n} 上的一个图属性，即一个从 g_{n} 到 {真, 假} 的映射。给定任意由邻接矩阵表示的图 $G \in g_{n}$，我们感兴趣的是通过连续探测其邻接矩阵的条目来测试 G 是否具有属性 P。在标准模型中（参见 Rivest and Vuillemin [12]），一个用于测试属性 P 的算法 A 表示为一棵二叉决策树，其中分支对应于被测试的矩阵条目是 0 还是 1。为了将此类常规算法与稍后描述的随机算法区分开，我们称 A 为纯算法。对于算法 A 的输入图 G，我们用 $r(A, G)$ 表示 A 进行的测试次数。令 A 为不做冗余测试的纯算法的族。我们现在准备讨论分布途径中的复杂性概念。给定 g_{n} 上的概率分布 d，纯算法 A 的平均代价定义为

$$C(A, d) = \sum_{G \in \mathbf{g}_n} d(G) \cdot r(A, G).$$

在输入分布 d 下，任何纯算法能实现的最佳结果显然是 $\min_{A \in A} C(A, d)$。我们现在定义测试 P 的分布复杂性为

$$F_1(P) = \sup_{d} \quad \min_{A \in A} C(A, d).$$

这个量 $\mathbf{F}_1(\mathbf{P})$ 捕捉了以下复杂性概念：**在已知输入分布的情况下，$\mathbf{F}_1(\mathbf{P})$ 是我们总能通过找到一个好算法来保证的平均代价。**

我们现在从随机化途径探索复杂性概念。一个随机算法 R 由 $\mathcal{A}$ 上的概率分布 q 指定。我们将 R 解释为以概率 q(A) 恰好按照算法 A 执行的算法。R 对于输入图 G 的期望代价是

$$\mathrm{E}(\mathrm{R}, \mathrm{G}) = \sum_{\mathrm{A}} \mathrm{q}(\mathrm{A}) \cdot \mathrm{r}(\mathrm{A}, \mathrm{G}).$$

R 的内在代价定义为对于“最坏”图 G 的 $\mathbf{E}(\mathbf{R},\mathbf{G})$，即 $\max_{G} E(R,G)$。从这个观点看，问题的一个自然的复杂性度量因此就是最佳随机算法的内在代价。也就是说，我们定义测试 $\mathbb{P}$ 的随机复杂性为

$$F_2(P) = \inf_{R} \max_{G \in g_n} E(R, G).$$

量 $\mathbb{F}_2(\mathbb{P})$ 告诉我们测试 $\mathbb{P}$ 的最佳随机算法需要多少时间。

一个新发展的概念是考虑允许最终答案有一定百分比错误的算法（参见 Karp [6], Rabin[10]）。我们现在为这些算法类定义相应的复杂性度量。让我们将纯算法族 A 扩展到一个更大的族 $A_{\mathfrak{o}}$，它包含所有对任何输入图 $G \in \mathcal{G}_{\mathfrak{n}}$ 给出“是”或“否”答案的决策树（答案不必正确）。对于决策树 $A \in A_{\mathfrak{o}}$ 和输入 $G \in \mathcal{G}_{\mathfrak{n}}$，令 $\epsilon(A, G) = 0$ 如果 A 对 G 给出了正确答案，否则 $\epsilon(A, G) = 1$。如前所述，令 $r(A, G)$ 是 A 为输入 G 进行的测试次数。

令 $\pmb{\lambda}$ 为介于 0 和 1 之间的一个数。对于任意输入分布 d，令 $\beta(\lambda)$ 是 $\mathbf{A}_0$ 的子集，由在 d 下错误概率不超过 $\pmb{\lambda}$ 的决策树组成。即，

$$\mathcal{B}(\lambda) = \left\{ A \mid A \in \mathcal{A}_0, \sum_{G \in g_n} d(G) \in (A, G) \leq \lambda \right\}.$$

具有误差 λ 的分布复杂性定义为

$$F_{1,\lambda}(P) = \sup_{d} \min_{A \in B(\lambda)} C(A,d)$$

其中 $C(A,d)$ 如前给出。

我们现在从随机化途径考虑带误差的复杂性。在族 $\lambda_{\circ}$ 上的分布 q 被称为 λ - 容忍的，如果

$$\sup_{G \in \mathbf{g}_n} \sum_{A \in \mathcal{A}_\circ} q(A) \cdot \epsilon(A,G) \leq \lambda.$$

为了确保所有输入的错误概率有界于 λ，我们只能允许其特征分布 q 是 λ - 容忍的随机算法。对于以 λ - 容忍分布 q 为特征的随机算法 R，对输入 G 的期望代价是

$$\mathrm{E}(\mathrm{R},\mathrm{G}) = \sum_{\mathrm{A} \in \mathrm{A}_\mathrm{O}} \mathrm{q}(\mathrm{A})\mathrm{r}(\mathrm{A},\mathrm{G}).$$

则具有误差 λ 的随机复杂性为

其中下确界取遍所有具有 λ - 容忍分布的随机算法 R。当 $\lambda = 0$ 时，$F_{1,\lambda}$ 和 $F_{2,\lambda}$ 分别简化为 F_{1} 和 F_{2}。

2.2 基本定理

$\mathbb{F}_1(\mathbb{P})$ 和 $\mathbb{F}_2(\mathbb{P})$ 的定义自然地提示了寻找 $\mathbb{F}_1(\mathbb{P})$ 下界和 $\mathbb{F}_2(\mathbb{P})$ 上界的方法。在第一种情况下，选择一个特定的分布 d，通过证明在该 d 下没有纯算法的平均测试次数少于 b，可以获得 $\mathbb{F}_1(\mathbb{P})$ 的下界 b。类似地，可以通过分析特定随机算法的性能来推导 $\mathbb{F}_2(\mathbb{P})$ 的上界。这些界限在实践中可能不容易获得，但至少我们知道要攻击什么。另一方面，我们如何证明 $\mathbb{F}_2(\mathbb{P})$ 的下界以了解随机化途径的局限性呢？因此，下面的定理是一座有用的桥梁。

定理 1 $\mathsf{F}_1(\mathsf{P}) = \mathsf{F}_2(\mathsf{P}).$

定理 1 可以通过使用著名的冯·诺依曼极小极大定理 [14] 来证明。它在美学上是优美的，因为两种复杂性视角实际上是相同的。它也提供了证明 $\mathbb{F}_2(\mathbb{P})$ 下界和 $\mathbb{F}_1(\mathbb{P})$ 上界的实用手段。例如，现在可以通过考虑“困难”的输入分布来确定随机化方法能力的局限性。下面的定理表明，在 g_n 中同构图上权重相等的分布是很好的候选。

定理 2 令 d 为 g_n 上的任意分布。则存在一个分布 d_0 使得
(i) $\min_{A \in \mathrm{A}} C(A,d_0) \geq \min_{A \in \mathrm{A}} C(A,d)$，并且
(ii) 如果 G_1 和 G_2 同构，则 $d_{\circ}(G_1) = d_{\circ}(G_2)$。

$\mathbf{F}_{\mathbf{i}}(\mathbf{P})$ 的精确计算可以表示为一个线性规划问题。对于小的 n，$\mathbf{F}_{\mathbf{i}}(\mathbf{P})$ 可以被有效地计算，并且定理 2 可以用于显著减少变量数量的优势。

对于具有误差容忍度的算法，下面的定理将两种途径发展的复杂性联系起来。

定理 3 对于 $0\leq \lambda \leq \frac{1}{2}$，有 $\mathbf{F}_{2,\lambda}(\mathbb{P})\geq \frac{1}{2}\mathbf{F}_{1,2\lambda}(\mathbb{P})$。

2.3 具体下界

在本小节中，我们推导测试图属性的某些具体下界。根据定理 1 和定理 3，只需证明 $\mathbf{F}_{1,\lambda}$ 的界即可。偶尔，我们可能发现可以直接为 $\mathbf{F}_{2,\lambda}$ 推导出更强的结果。

定义 如果 P(空图) = 假，我们称 P 为 $\mathcal{G}_{\mathrm{n}}$ 上的正规图属性。对于 $\mathcal{G}_{\mathrm{n}}$ 中的任意图 S，S 的大小记为 $\|S\|$，是 S 中的边数。对于一个正规属性 P，如果 S 是满足 $P(S) = \text{真}$ 的最小规模图，则图 S 是 P 的最小图。对于 $\mathcal{G}_{\mathrm{n}}$ 中的任何 G，令 $\pi(G)$ 表示 G 的自同构群。（G 的自同构是顶点集 $\{1,2,\ldots,n\}$ 的一个重标号，使得 G 保持不变。）

定理 4 令 $\mathbf{P}$ 为 $\mathcal{G}_{\mathfrak{n}}$ 上的一个正规图属性，S 是 $\mathbf{P}$ 的一个最小图，且 $0 \leq \lambda \leq 1/2$。令 $s = \| S \|$。则

定理 4 的精神类似于 Kirkpatrick [7, 定理 2.3] 的一个结果。作为一个应用，令 $\mathbf{P}$ 为图是非平面图的性质，则 S 是 $K_{3,3}$，一个 3 乘 3 的完全二分图。定理 4 告诉我们，对于任何随机算法，即使允许 25% 的错误，也需要 $\Omega(n^{2})$ 次测试。关于更多具有小最小图的属性，参见 Kirkpatrick [7]。

定理 4 可以推广到二分图属性。正如在最坏情况复杂性中 [7]，这可以与嵌入过程结合使用，以证明各种“连通性”属性的 $\Omega(n^2)$ 下界。以下定理可以通过非平凡的嵌入过程证明。

定理 5 令 T 为 g_{n} 中的任意树，P 为图包含 T 作为子树的性质。则

定理 6 令 $0 < k \leq n$，P 为包含大小为 k 的团的性质。则

任何图属性 $\mathbb{P}$ 都有一个线性下界 $\mathbb{F}_{\mathbf{i}}(\mathbb{P}) \geq \Omega(n)$。这可以如下看出。不失一般性，我们可以假设 $\mathbb{P}$ 是正规的。令 S 是 $\mathbb{P}$ 的大小为 s 的最小图。很容易证明 $\mathbb{F}_{\mathbf{i}}(\mathbb{P}) \geq s$。当我们将此与定理 4 结合时，有

这意味着 $F_{i}(P) \geq \Omega(n)$。从证明比 $\Omega(n)$ 更强的界的角度来看，具有 $s \approx \text{常数} \times n$ 的属性是最难处理的。下一个定理在这种情况下通常很有用。

定理 7 令 S 为 n 个顶点上正规图属性 P 的最小图。令 $\| S \| = s$，则

作为其应用的一个说明，考虑图是哈密顿图的性质。显然 S 现在是一个哈密顿回路，群 $\pi(S)$ 是 $\{1,2,\ldots,n\}$ 上的循环群。因此，$s = n$ 且 $|\pi(S)| = n$。所以，对于哈密顿性质 P，我们有

对于包含完美匹配的性质，可以推导出类似的结果。虽然我们尚未成功将定理 7 扩展到 $\mathbf{F}_{\mathbf{i},\lambda}$（其中 $\lambda \neq 0$）的下界，但对于是哈密顿图和包含完美匹配的特殊情况，可以修改证明来展示以下定理。

定理 8 令 $\mathbf{P}_1$ 和 $\mathbf{P}_2$ 分别为 g_n 上是哈密顿图和包含完美匹配的性质。则对于 $\lambda \in [0,0.1]$，有

3. 偏序问题

第 2 节讨论的复杂性概念可以扩展到任何具有有限可能输入集和算法集的问题。定理 1 和定理 3 仍然有效，如果问题中存在“对称性”，定理 2 也有类似物。（我们将在下面看到一个例子。）

可能研究最广泛的决策树问题是那些与偏序相关的问题（例如 [8]）。我们将集中讨论涉及元素选择的一个特殊类。（关于其他类型的偏序问题，参见 Fredman [4]。）令线性有序集 V 中元素 $\mathbf{x}$ 的秩 $\ell_{\mathrm{V}}(\mathbf{x})$ 为小于或等于 $\mathbf{x}$ 的元素个数。对于满足 $n \geq k$ 的正整数 n 和 k，令 $\mathbf{I}_{n,k} = \{(i_1, i_2, \ldots, i_k) | \text{ } i_j \text{ 是介于 } 1 \text{ 和 } n \text{ 之间的不同整数}\}$ 为所有不超过 n 的不同整数的有序 k 元组的集合。集合 V 上的一个选择问题 J 由集合 $J \subseteq I_{n,k}$ 指定，可描述如下：给定 n 个元素的集合 V，我们希望通过元素间的两两比较，找到 k 个元素 $(x_1, x_2, \ldots, x_k)$，使得 $(\ell_{\mathrm{V}}(x_1), \ell_{\mathrm{V}}(x_2), \ldots, \ell_{\mathrm{V}}(x_k))$ 在 J 中。例如，排序问题对应于选择 $J = \{(1, 2, \ldots, n)\}$；在选择最小的 k 个元素时，我们设 J 为 $(1, 2, \ldots, k)$ 的所有排列的集合。

对于集合 $V = \{x_1, x_2, \ldots, x_n\}$ 上的选择问题 J，算法 A 是一个决策树，其中每个内部节点执行形式为“$x_i : x_j$”的比较。在树的每个叶子处，指定一个输出 $(x_{il}, x_{i2}, \ldots, x_{ik})$。让我们使用 $F_i(J)$ 和 $F_{i,\lambda}(J)$ 来表示选择问题 J 的复杂性，如第 2 节所做。定理 1 仍然成立，所以我们有 $F_l(J) = F_2(J)$。关于选择问题的复杂性，有两个有趣的结果我们现在讨论。

首先，任何选择问题都有一个很好的对称性。直观地说，如果我们取任何算法 A，并以任意方式重新标记 V 的元素，A 将相应地转换为另一个有效算法。这个观察可以用来推导与定理 2 平行的以下结果。让我们仍然用 $\mathfrak{A}$ 表示算法集，用 $C(A,d)$ 表示算法 A 在输入分布 d 下的平均比较次数。

定理 9 令 J 为集合 V 上的一个选择问题，d_u 为输入的均匀分布，即 V 上的每个线性序等可能。则

非正式地说，定理 9 告诉我们均匀分布是“最难”的分布，我们可以通过研究在 d_{u} 下的最优算法来找到 $F_{i}(J)$ 的界。有许多关于各种问题的量 $\min_{A} C(A, d_{u})$ 的研究 [2,8,11,15]。

其中一些结果涉及相当困难的证明。这些结果现在通过定理 9 获得了新的意义。例如，Floyd 和 Rivest [2] 证明了对于寻找 n 个数的中位数问题，$1.375n \leq \min_{A \in \mathcal{A}} C(A, d_u) \leq 1.5n$。

在新的解释下，我们可以陈述：（1）对于任意分布 d，总存在一个（纯）算法，其平均比较次数少于 1.5 n；（2）任何随机算法对于某些输入必须至少进行期望 1.375 n 次比较。

现在，讨论第二个兴趣点。到目前为止，在我们的模型中没有看到任何随机算法在数量级上优于解决同一问题的传统（纯）算法的最坏情况复杂性。我们现在将演示一个允许小错误的选择问题中，随机算法在数量级意义上的优越性。这样一个例子来自于选择一个平庸元素（参见 F. Yao[16]），为了我们当前的目的，将其定义为一个秩位于中间三分之一的元素。即，对于任何输入集 $V = \{x_1, x_2, \ldots, x_n\}$，我们希望找到一个元素 x_j，使得 $n/3 \leq \ell_V(x_j) \leq 2n/3$。对于任何固定的 $0 < \lambda < 1$，不难构造一个随机算法，其错误概率以 λ 为界，并且仅使用常数 $\log(1/\lambda)$ 次比较。由于任何纯算法在最坏情况下必须进行 $2n/3$ 次比较，这为我们的断言提供了一个例子。然而，我们没有类似的例子，其中无错误的随机算法比最坏情况复杂性好得多。

4. 结束语

我们在一些决策树模型中发展了一个统一的平均复杂性概念，并证明了其有用性。这些结果中有许多可以推广。例如，定理 4 可以推广到任何不变群是可迁的字符串属性。我们在下面列出一些重要的未解决问题。

（1）尽管随机化已成功应用于若干问题 [1,11,13]，但几乎没有证据表明随机化途径确实能（与纯算法的最坏情况复杂性相比）在数量级上降低计算时间。除了我们在第 3 节中的例子，文献中唯一具体的演示似乎是 Gill 论文 [3, 修订版] 中引用的 Frevald 的一个结果。它指出，语言 $\{ww|w\in \{0,1\}^{*}\}$ 可以被一个 1-带图灵机以错误概率 $\lambda >0$ 在 $O(n\log^2 n)$ 步内识别（相比于传统 1-带图灵机的 $c\cdot n^2$ 步）。对于无错误的随机算法，似乎没有类似的例子。问题：在我们的模型中，有这样的例子吗？

（2）对于所有单调图属性，是否总有 $\mathbf{F}_{\dot{\mathbf{i}}}(\mathbb{P}) \geq c \cdot n^{2}$？对应的最坏情况复杂性已由 Rivest and Vuillemin [12] 回答。通常需要为 $\mathbf{F}_{\dot{\mathbf{i}}}(\mathbb{P})$ 开发更多的下界技术。

（3）关于定理 3，能否证明一个相反方向的不等式？

（4）能否在具体例子中展示 1-带概率图灵机带错误时的非线性下界？一个好的候选是识别语言 $\mathbf{L} = \{\mathbf{wcw}^{\mathbf{R}}| \quad \mathbf{w} \in \{0,1\}^{*}\}$（参见 [5] 定理 10.7）。

（5）这里研究的课题与博弈论和统计决策问题有某些共同点。了解进一步的关联将会很有趣。

参考文献

[1] J. Carter and M. Wegman, "Universal Classes of Hash Functions", Proc. 9th Annual ACM Symposium on Theory of Computing (1977), 106 - 112.

[2] R. Floyd and R. Rivest, "Expected Time Bounds for Selection", CACM 18 (1975), 165-172.

[3] J. Gill, "Computational Complexity of Probabilistic Turing Machines", Proc. 6th Annual ACM Symposium on Theory of Computing (1974), 91-95, revised Feb. 1977.

[4] M. Fredman, "How Good Is the Information Theory Bound?", Theoretical Computer Science 1 (1976), 355-361.

[5] J. Hopcroft and J. Ullman, Formal Languages and their Relation to Automata, Addison-Wesley(1969).

[6] R. Karp, "Probabilistic Analysis of Combinatorial Search", in Algorithms and Complexity, edited by J. Traub, Academic Press (1976), 1-20.

[7] Kirkpatrick, "Topics in the Complexity of Combinatorial Algorithms", University of Toronto Technical Report no. 74(1974).

[8] D. Knuth, The Art of Computer Programming vol 3, Addison-Wesley (1973).

[9] R. Luce and H. Raiffa, Games and Decisions, Wiley (1957).

[10] M. Rabin, "Probabilistic Algorithms", in Algorithms and Complexity, edited by J. Traub, Academic Press (1976), 21-40.

[11] I. Pohl, "Minimean Optimality in Sorting Algorithms" Proc. 16th Annual Symposium on Foundations of Computer Science (1975), 71-74.

[12] R. Rivest and J. Vuillemin, "On Recognizing Graph Properties From Adjacency Matrices", Theoretical Computer Science 3 (1976), 371-384.

[13] R. Solovay and V. Strassen, "Fast Monte-Carlo Test for Primality", SIAM J. on Computing 6(1977), 84-85.

[14] Von Neumann, "Zur Theorie der Gesellschaftsspiele", Math Annalen 100(1928), 295-320.

[15] A. Yao and F. Yao, "On the Average-Case Complexity of Selecting the t-th Best", to appear.

[16] F. Yao, "On Lower Bounds for Selection Problems", MIT Project Report TR-121 (1974).

专业术语中英对照表

英文术语	中文翻译
Probabilistic Computations	概率计算
Unified Measure of Complexity	复杂性的统一度量
Expected Running Time	期望运行时间
Distributional Approach	分布途径
Randomized Approach	随机化途径
Randomized Algorithm	随机算法
Distributional Complexity	分布复杂性

英文术语	中文翻译
Randomized Complexity	随机复杂性
Decision Tree	决策树
Pure Algorithm	纯算法
Binary Decision Tree	二叉决策树
Adjacency Matrix	邻接矩阵
Graph Property	图属性
Average Cost	平均代价
Expected Cost	期望代价
Intrinsic Cost	内在代价
Error Tolerance / Error Probability	误差容忍度 / 错误概率
Minimax Theorem	极小极大定理
Lower Bound / Upper Bound	下界 / 上界
Normal Graph Property	正规图属性
Minimal Graph	最小图
Automorphism	自同构
Linear Programming Problem	线性规划问题
Partial Order Problems	偏序问题
Selection Problem	选择问题
Rank	秩
Uniform Distribution	均匀分布
Worst-case Complexity	最坏情况复杂性
Order-of-magnitude	数量级
Probabilistic Turing Machine	概率图灵机
Monotone Graph Property	单调图属性