第四章 现实生活中的NP难度问题及其现实处理方法

金人超

华中科技大学计算机科学与技术学院
jrc@hust.edu.cn

0-1背包问题的动态规划近似算法（伪多项式时间算法）

假定物品价值$p_1, ..., p_n$都是正整数。定义$W(i,p)$为在前i个物品中挑选若干价值和为p的物品时能得到的最小重量和；若无法使价值和为p，则$W(i,p)=+\infty$。则：

$$W(i+1,p)=\begin{cases}\min\{W(i,p),\ w_{i+1}+W(i,p-p_{i+1})\}, & \text{if } p_{i+1}\le p;\\ W(i,p), & \text{otherwise.}\end{cases}$$

找出满足$W(n,p)\le W$的最大的p既为最终结果。采用动态规划算法，需要的计算时间和空间都为

$$O(n\cdot\Sigma p_i)=O(n^2 p_{max})$$

注意$p=\Omega(2^{|p|})$，其中$|p|$为p的表示规模。

0-1背包问题的多项式时间近似策略(FPTAS)

对任意$\varepsilon>0$，令$K=\varepsilon\, p_{max}/n$

$p_i' = \lfloor p_i/K \rfloor \le n/\varepsilon$，$i=1,2,\ldots,n$

使用前述伪多项式算法求解整数0-1背包问题。设其得到的解的物品集为S，而原问题的最优解的物品集为O。则

$$P(S) \ge P'(S)K \ge P'(O)K \ge P(O)-nK$$

$$=P(O)-\varepsilon\, p_{max}$$

$$\ge(1-\varepsilon)P(O) \qquad (\text{不妨设}\, p_{max} \le P(O))$$

NP难问题的近似算法

1. 最优解的可近似程度

若一个求解最优化问题的一个近似算法A求得的实例I的近似最优解目标函数值为c，实例I相应的最优解值为c^*，则A是一个ε-近似算法（$0 < \varepsilon < 1$)）当且仅当对任意实例I：

$$\frac{|c^* - c|}{\max(c^*, c)} \le \varepsilon$$

Four Classes of NP Optimization Problems

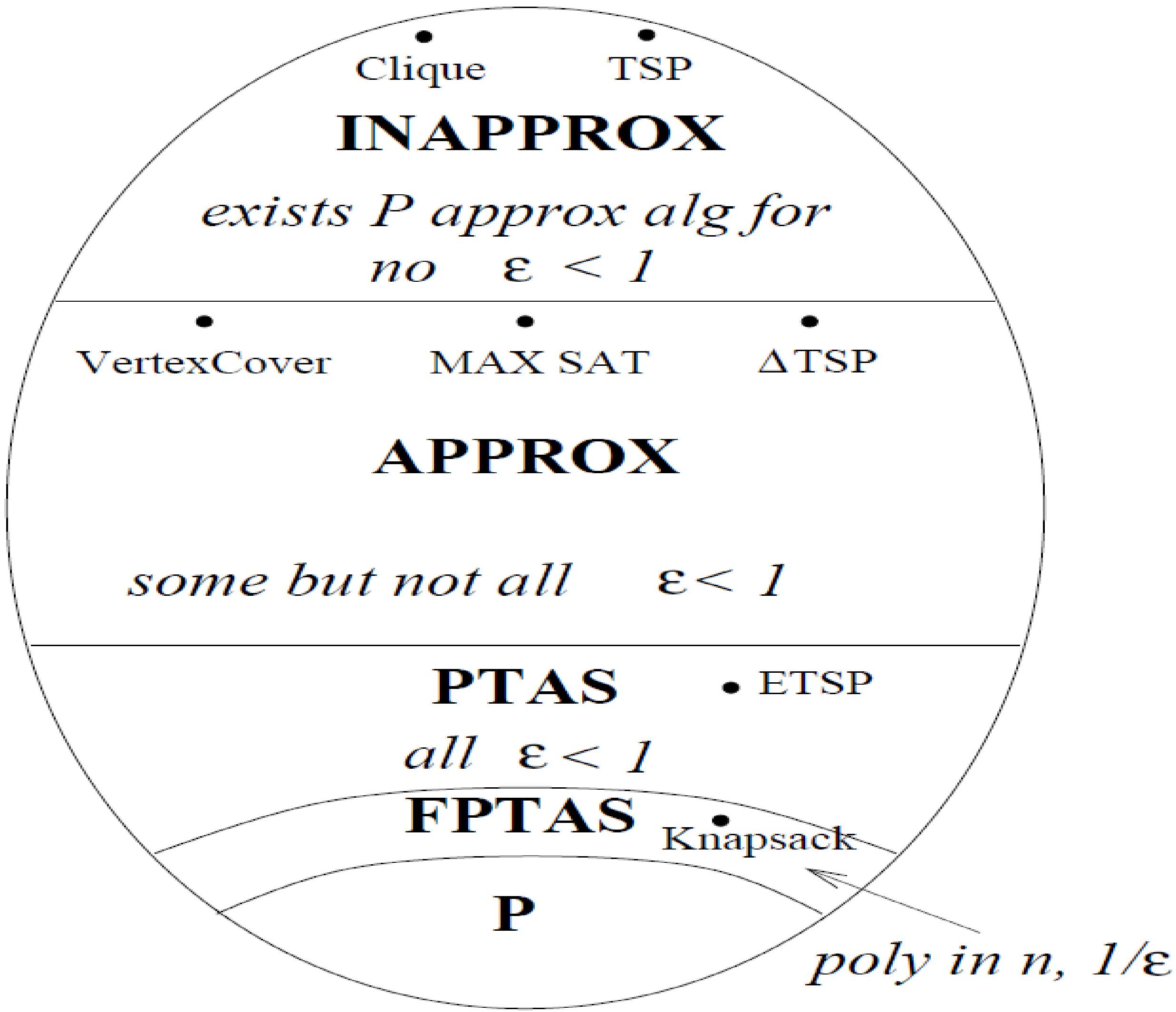
$$\text{INAPPROX} \equiv \text{no PTIME } \epsilon\text{-approx alg if P} \neq \text{NP}$$

$$\begin{aligned}\text{APPROX} \equiv\ & (\exists \epsilon_1 \epsilon_2 \,.\, 0 < \epsilon_1 < \epsilon_2 < 1) \\ & \text{exists PTIME } \epsilon_2\text{-approx alg} \\ & \text{no PTIME } \epsilon_1\text{-approx alg if P} \neq \text{NP}\end{aligned}$$

$$\text{PTAS} \equiv (\forall \epsilon > 0)\text{exists PTIME } \epsilon\text{-approx alg}$$

$$\begin{aligned}\text{FPTAS} \equiv\ & (\forall \epsilon > 0)\text{exists uniform } \epsilon\text{-approx alg} \\ & \quad \text{running in time poly}(n, \tfrac{1}{\epsilon})\end{aligned}$$

(F)PTAS stands for *(Fully) Polynomial-Time Approximation Scheme.*


Clique
TSP
INAPPROX
exists P approx alg for
no ε < 1
VertexCover
MAX SAT
ΔTSP
APPROX
some but not all ε< 1
PTAS
ETSP
all ε< 1
FPTAS
Knapsack
P
poly in n, 1/ε

面对寻求绝对形式化完美方法的失败，科学家不得不“返朴归真”，从自然界、人类社会寻求灵感和智慧。

“师法自然”，“外师造化，内得心源”

拟物方法：到物理世界中寻找出于原始数学问题等价的自然现象，观察其中物质运动的演化规律，从中受到启发，得出对数学问题的求解算法。

物理状态的演化天然地是使能量函数最小化，最后往往陷入局部极小陷阱，难以求得问题的全局最优解。

拟人方法：人类在最近几千年的社会生活中形成了丰富的社会经验，利用这些经验往往可以启发出好的跳出局部极小陷阱的策略，将这些策略形式化为算法，称为拟人途径。

遗传进化算法、人工神经网络、模拟淬火算法、蚁群算法等是近年来发展出来的“师法自然”的算法，它们与拟物拟人算法的不同在于针对性较差。照搬照套地使用这些算法时往往不能取得好的结果。

拟物拟人算法的优点在于针对问题对症下药。

§3 求解SAT问题的拟物拟人方法

对于一般的 CNF

$$\bigwedge_{i=1}^{l}(P_{i,1} \vee P_{i,2} \vee \cdots \vee P_{i,k_i} \vee \overline{P}_{ri,1} \vee \overline{P}_{ri,2} \vee \cdots \vee \overline{P}_{ri,K_{ri}}), \tag{2.7}$$

其中 $P_{i,1},\cdots,P_{i,k_i},P_{ri,1},\cdots,P_{ri,k_{ri}}$ 为命题变元集 $\{P_1,\cdots,P_m\}$ 中两两不同的命题变元．相应地写出总势能函数

$$U(x_1,x_2,\cdots,x_m) = \sum_{i=1}^{l} U_i(x_1,x_2,\cdots,x_m), \tag{2.8}$$

其中

$$U_i(x_1,x_2,\cdots x_n) = \begin{cases} (1-x_{i,1})(1-x_{i,2})\cdots(1-x_{i,k_i})x_{ri,1}x_{ri,2}\cdots x_{ri,k_{ri}}, \\ \qquad \text{若 } x_{i,1} \leqslant 1 \wedge \cdots \wedge x_{i,k_i} \leqslant 1 \wedge x_{ri,1} \geqslant 0 \wedge \cdots \wedge x_{ri,k_{ri}} \geqslant 0; \\ 0, \quad \text{否则}. \end{cases}$$

(2.9)[1)]

这里 $U_i(x_1,x_2,\cdots,x_m)$ 的物理意义为 m 维 Euclid 空间中带负电的导体

$$R^m - x_{i,1} \leqslant 1 \wedge \cdots \wedge x_{i,k_i} \leqslant 1 \wedge x_{ri,1} \geqslant 0 \wedge \cdots \wedge x_{ri,k_{ri}} \geqslant 0$$

所诱导出的电场的静电势能，其中 $x_{i,1},\cdots,x_{i,k_i},x_{ri,1},\cdots,x_{ri,k_{ri}}$ 为实变元集 $\{x_1,\cdots,x_m\}$ 中两两不同的实变元．

CNF(2.7)的 SAT 问题等价于对(2.8)式中总势能函数 $U(x_1, x_2, \cdots, x_m)$求最小值点的问题. 设 $x_1^*, x_2^*, \cdots, x_m^*$ 为最小值点, 若 $U(x_1^*, x_2^*, \cdots, x_m^*)>0$ 则 SAT 问题无解, 若 $U(x_1^*, x_2^*, \cdots, x_m^*)=0$, 则依变换(2.10)

$$P_v^* = \begin{cases} 1, \text{若 } x_v^* \geqslant 1; \\ 0, \text{若 } x_v^* \leqslant 0; \\ \text{任意地给定 0 或 1, 若 } x_v^* \in (0,1); \end{cases} \quad \text{对于 } v = 1,2,\cdots,m. \tag{2.10}$$

$P_1^*, P_2^*, \cdots, P_m^*$ 为 CNF(2.7)的成真指派. 反之若 $P_1^*, P_2^*, \cdots, P_m^*$ 为 CNF(2.7)的成真指派则总势能函数

$$U(P_1^*, P_2^*, \cdots, P_m^*) = 0. \tag{2.11}$$

参考文献:

1. 黄文奇，金人超. 求解SAT问题的拟物拟人算法——Solar. 中国科学(E辑), 1997, 27(2): 179-186.
2. 金人超，黄文奇. 并行计算: 提高SAT 问题求解效率的有效方法. 软件学报，2000, 11 (3) : 398～ 400

- 启发式算法（heuristic algorithm）
- 定义1. 基于直观或经验构造的算法，在可接受的花费（时间、空间）下，给出待解组合优化问题的每个实例的一个可行解，该可行解与最优解偏差事先不一定可以预计.
- 定义2. 启发式算法是一种技术，在可接受的计算费用内寻找最好解，但不保证该解的可行性与最优性，无法描述该解与最优解的近似程度。
- 特点（与传统优化方法不同）：凭直观和经验给出算法；不在理论上证明所得解与最优解的最大偏离程度.

现代优化算法

- 局部搜索（Local search）
- 禁忌搜索（tabu search）
- 模拟退火（simulated annealing）
- 蚁群算法 （Ant Colony optimization）
- 遗传算法（genetic algorithms）
- 群体（群集）智能（Swarm intelligence）
- 拉格朗日松弛算法（lagrangean relaxation）
- 人工神经网络（artificial neural networks）
- 深度学习（deep learning）

优点：

（1）有可能比简化数学模型方法求得的解的误差小；

（2）计算时间可接受；

（3）可用于某些最优化算法（如分支定界算法）之中的估界；

（4）直观易行；

（5）速度较快；

（6）程序简单，易修改。

不足：

（1）不能保证求得全局最优解；

（2）解的精度不稳定，有时好有时坏；

（3）算法设计与问题、设计者经验、技术有关，缺乏规律性；

（4）不同算法之间难以比较。

现代优化方法的计算时间评价

(1)概率分析 （probability analysis）

用最坏情况分析，会因一个最坏实例影响总体评价.

在实例数据服从一定概率分布情形下，研究算法复杂性和解的效果.

(2)大规模计算分析

通过大量实例计算，评价算法效果.

- 注意数据的随机性和代表性.
- Benchmark

机器学习的复杂度

- PAC（Probably Approximately Correct可能近似正确）模型。 Valiant，1984

S：一个样本空间（例如，平面上所有点的集合）

D：样本空间中样本的概率分布（例如，点在平面上均匀分布）

$c: S \to \{0,1\}$：一个真实概念，接受或拒绝样本空间中的每个点（例如，平面上某条直线，接受直线上方的点，拒绝下方的点）

C：一个概念类（例如，平面上所有直线的集合）

学习的目标：使用一定数量的已知样本，以至少$1-\delta$的概率学到一个模型 $h \in C$， 以高概率正确分类未知样本：

$$\Pr_{x \in D}\left[h(x) = c(x)\right] \geq 1 - \epsilon$$

例如：使用m个已知点，以至少85%的概率学到一条直线 $h \in C$， 以至少95%的概率正确划分平面上所有点。

机器学习的复杂度

计算学习理论的关键问题之一是样本复杂性

- 需要多少样本数据才能实现目标?

Valiant 提出了以下定理，用于**有限**概念类：

$$m = O\left(\frac{1}{\epsilon}\log\frac{|C|}{\delta}\right)$$

学习算法：

输入：m个已知样本$x_1,\ldots,x_m,c(x_1),\ldots,c(x_m)$；

输出：模型$h \in C$；

1. 找到一个满足所有样本的模型$h \in C$，即$h(x_i) = c(x_i)$ for all $x_1,\ldots,x_m$;
2. 输出h。

机器学习的复杂度

用反证法。 设 $h \in C$ 是任何“坏”模型：也就是说，使得 $\Pr[h(x) = c(x)] < 1 - \epsilon$ 。

那么如果独立地从样本分布 D 中挑选出 m 个点， h 在所有这些点上的正确概率最多为$(1-\epsilon)^m$. 因此，在 C 中存在与所有样本数据一致的坏模型的概率最多为 $|C|\,(1-\epsilon)^m$

$$
\begin{aligned}
\delta &= |C|\,(1-\epsilon)^m \\
m &= \log_{1-\epsilon}\frac{\delta}{|C|} \\
&= \frac{\log \delta/|C|}{\log 1-\epsilon} \\
&\approx \frac{1}{\epsilon}\log\frac{|C|}{\delta}.
\end{aligned}
$$