

2-3-SAT 问题相变现象剖析及其应用*

白 颖 卜东波

(国家智能计算机研究开发中心 北京 100080)

(北京曙光信息技术产业公司 北京 100080)

E-mail: bai@gatekeeper.ncic.ac.cn/dongbobu@usa.net

摘要 3-SAT 问题有一个非常奇妙的相变现象. 对于固定的变量数 N , 合取范式的可满足概率随着子句个数 K 的变化而发生剧烈的变化; 当 $K \approx 4.3 * N$ 时, 可满足概率急剧地从 1 变为 0. 相变现象决定了问题的难易分布, 对于快速求解算法的设计有着非常重要的意义. 文章着重讨论了 SAT 问题的更一般形式, 即 2-3-SAT 问题的相变现象, 研究了相变点处的 2-子句和 3-子句个数的关系, 发现了 2-子句和 3-子句在约未能力意义下的当量关系, 并提出了如何有效地利用 2-3-SAT 的相变现象.

关键词 NP 完全问题, 合取范式, SAT 问题, 相变现象, 可满足概率, 2-3-当量.

中图法分类号 TP301

可满足性问题(SAT. 问题)是一大类 NP 完全问题的核心, 在数据库检索、VLSI 设计、计算机视觉、机器定理证明、机器人规划、机器学习等领域有着广泛的应用. 由于理论和实践的背景, SAT 问题的复杂性分析及算法设计一直是计算机科学领域的焦点之一.

物质的三态(固态、液态、气态)变化是自然界常见的现象, 它反映了物质量变和质变的规律. 复杂的计算系统中存在着类似的现象, 而且其相变规律可以用几个宏观参数加以刻画. Cheesman 的工作说明了在 SAT 问题中也存在着相变现象, 并且猜想相变点附近包含“难解”的实例. 白颖和卜东波^[1]运用回归分析的方法逼近 3-SAT 和 2-SAT 的相变曲线, 从宏观的角度把握相变现象.

在实践中, 2-3-SAT 是一种更一般的形式, 它包含 2-SAT 和 3-SAT 作为特例. 作为文献[1]的继续, 本文使用文献[1]的结论, 对 2-3-SAT 的相变现象进行研究, 发现了 2-子句和 3-子句的当量关系, 证明了相变点处 2-子句和 3-子句的线性关系, 这个结论和实验数据非常接近.

1 SAT 问题及其相变现象简介

SAT 问题是第 1 个 NP 完全问题.^[2]我们将描述这个问题时所用的术语定义如下:

设 $U = \{u_1, u_2, \dots, u_n\}$ 表示布尔变量集.

定义 1.1. 真值赋值

关于 U 的真值赋值是一个函数 $t: U \rightarrow \{T, F\}$. 如果 $t(u) = T$, 我们称 u 在赋值 t 下取真值; 如果 $t(u) = F$, 我们称 u 在赋值 t 下取假值. 真值赋值可以简单地用一个 N 元布尔向量组表示.

定义 1.2. 文字

如果 u 是 U 的一个变量, 那么, u 和 $\neg u$ 是 U 上的文字. 文字 u 在赋值 t 下取真值, 当且仅当变量 u 在赋值 t 下取真值; 文字 $\neg u$ 在赋值 t 下取真值, 当且仅当变量 u 在赋值 t 下取假值.

定义 1.3. 子句

U 上的子句是 U 上的一些文字的析取. 子句在赋值 t 下取真值, 当且仅当子句包含的文字至少有一个在赋值 t 下取真值. k 子句表示子句包含 k 个文字.

定义 1.4. 合取范式(Conjunctive Normal Formula)

* 本文研究得到国家 863 高科技项目基金资助. 作者白颖, 1956 年生, 博士, 研究员, 博导, 主要研究领域为人工智能, 计算语言学, Internet. 卜东波, 1973 年生, 博士生, 主要研究领域为计算复杂性, 算法设计, 计算机网络技术.

本文通讯联系人: 白颖, 北京 100080, 国家智能计算机研究开发中心

本文 1997-03-11 收到原稿, 1997-11-10 收到修改稿

U 上的合取范式是 U 上的一些子句的合取. 合取范式赋值 t 下取真值(也称为合取范式在赋值 t 下可满足), 当且仅当合取范式包含的所有子句在赋值 t 下皆取真值.

SAT(satisfiability)问题规定如下:

实例: 给定变量集 U 以及 U 上的合取范式 C , 问: 是否存在满足 C 的赋值 t (也称为 C 的解)?

在本文中, 我们用 N 表示布尔变量数, K 表示 3-子句个数, M 表示 2-子句个数, ρ 表示 K 和 N 的比值, $\rho = K/N$.

随机 k -SAT 问题是 SAT 问题中的一个子类, 它的实例中的子句都是随机独立产生的, 并且每个子句都由互不相同的 k 个变量形成的文字组成. 2-3-SAT 是指每个子句都包含 2 个或 3 个互不相同的变量形成的文字.

一个由大量相互弱作用的个体组成的复杂系统, 随着个体数目的增加常常会有从有序到无序的转变. Cheeseman 的实验结果首先证实了 SAT 问题相变现象的存在, Mitchell^[3]的工作进一步讨论了这一现象, 并且说明了困难的实例大部分出现在相变的过渡区域内. 图 1 是我们的实验数据.

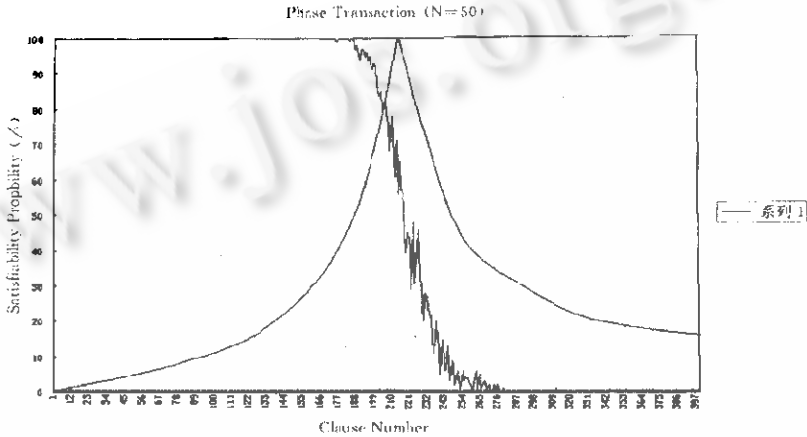


图 1 3-SAT 问题的可满足概率的相变曲线和难度曲线 ($N=50, K=1 \dots 600$)

在图 1 中, 锯齿状曲线表示可满足概率曲线, 钟型曲线是难度曲线. x 轴表示子句数 K , y 轴表示可满足的概率 P . 容易看出: 当 K 相对较小时, 实例大部分可以满足; 当 K 相对较大时, 实例大多不可以满足; 而在 $K/N \approx 4.25$ 附近, 可满足概率急剧地由 1 变为 0. 可以看出, 相变区域的难度较大.

2 3-SAT 问题可满足概率的布尔筛模型

在文献[1]中, 白硕和卜东波提出了 3-SAT 问题的布尔筛模型. 即: 任一布尔赋值看作一个通透率为 $7/8$ 的布尔筛, 对 3-子句进行过滤. 布尔赋值是子句的解, 则子句可以通过, 否则被阻碍. 比如, 对于赋值 $t_i = \langle l_1, l_2, \dots, l_N \rangle$, 任一 3-子句可以 $7/8$ 的概率通过这个布尔筛. 其中, 被阻碍的子句的形式为 $C = (\neg l_p \vee \neg l_q \vee \neg l_r)$. 有以下定理成立.

定理 1. 在理想情况下, 即全体 2^N 个布尔赋值皆相互独立, 则 CNF 可满足概率 P_r 为 $P_r = 1 - (1 - (7/8)^3)^{2^N}$.

证明: \because 对于任一赋值形成的布尔筛, 任一 3-子句通过的概率为 $7/8$

$\therefore K$ 个子句全通过的概率为 $(7/8)^K$, 即赋值是 CNF 的解

\therefore 至少有一个子句被阻碍的概率为 $1 - (7/8)^K$, 即赋值不是 CNF 的解

\therefore CNF 对所有 2^N 个赋值形成的布尔筛, 均被阻碍的概率为 $(1 - (7/8)^3)^{2^N}$, 即 CNF 无解

\therefore CNF 可满足概率 P_r 为 $P_r = 1 - (1 - (7/8)^3)^{2^N}$ □

由于全体 2^N 个布尔赋值皆相互独立的假定不成立, 因此, 在非理想情况下, 我们使用以下概率表达式:

$$P_r(\rho) = 1 - (1 - \alpha^K)^{\beta^N} = 1 - (1 - \alpha^\rho \cdot N)^{\beta^N} \quad (0 < \alpha < 1, \beta > 1) \quad (1)$$

定理 2. 如果 $0 < \alpha < 1$ 且 $\beta > 1$, 则存在 ρ_0 满足下式:

$$\lim_{N \rightarrow +\infty} P_r(\rho) = \begin{cases} 1 & \text{若 } \rho < \rho_0 \\ 1 - 1/e & \text{若 } \rho = \rho_0 \\ 0 & \text{若 } \rho > \rho_0 \end{cases}$$

证明:

$$\begin{aligned}
\lim_N P_r(\rho) &= \lim_N (1 - (1 - \alpha^{\rho \cdot N}) \beta^N) \\
&= 1 - \lim_N (1 - \alpha^{\rho \cdot N}) \beta^N \\
&= 1 - \lim_N (1 - \alpha^{\rho \cdot N})^{\alpha^{-\rho \cdot N} \cdot \alpha^{\rho \cdot N} \cdot \beta^N} \\
&= 1 - \lim_N ((1 - \alpha^{\rho \cdot N})^{\alpha^{-\rho \cdot N}})^{\alpha^{\rho \cdot N} \cdot \beta^N} \\
&= 1 - (\lim_N (1 - \alpha^{\rho \cdot N})^{\alpha^{-\rho \cdot N}}) \lim_N (\alpha^{\rho \cdot N} \cdot \beta^N) \\
&= 1 - \lim_N e^{-(\alpha^{\rho \cdot N})^N}
\end{aligned}$$

∴
令

$$\lim_{N \rightarrow +\infty} P_r(\rho) = \begin{cases} 0 & \text{若 } \alpha^\rho \cdot \beta > 1 \\ 1 - 1/e & \text{若 } \alpha^\rho \cdot \beta = 1 \\ 1 & \text{若 } \alpha^\rho \cdot \beta < 1 \end{cases}$$

$$\rho_0 = -\frac{\log \beta}{\log \alpha}$$

则 ρ_0 满足上式.

定理 2 说明了在极限意义下,可满足概率仅有 3 种可能取值. 取值的不同仅仅依赖于子句数和变量数比值的不同.

定理 3. $P''_r(\rho_0) = 0$

证明:对 $P_r(\rho)$ 取二阶导数,将 ρ_0 代入即可证明. 详细证明略.

函数的一阶导数表明了函数的变化率,二阶导数等于零的点表明了函数在此点的变化率最大. 因此,定理 2 说明了可满足概率在 ρ_0 点的变化率最大,即 ρ_0 是可满足概率的相变点. 并且, $P_r(\rho_0) = 1 - 1/e \approx 0.63... > 0.5$, 因而,我们认为相变点应当在可满足概率等于 $(1 - 1/e)$ 处,而不是可满足概率等于 0.5 处. [4-5] 这和文献 [5,7] 中提到的以可满足概率等于 0.5 处作为相变点,会发生相变点左漂移现象是一致的.

推导出可满足模型之后,可以应用回归分析的方法确定 α 和 β . 即在采集了大量实验数据之后,用式 (1) 所表示的概率模型逼近实验数据,求在最小二乘意义下的最优. 求得的结果为 $\alpha = 0.96...$ 和 $\beta = 1.16...$ 直观上讲, α 表示随机产生一个 3-子句通过布尔筛的概率,也就是随机产生一个 3-子句的统计平均可满足概率;而 β 则表示每个变量的统计平均自由度,即每个变量在统计意义下,可以取多少值.

图 2 是用上述概率表达式分别逼近 $N=100, 125, 150, 175$ 四条相变曲线的结果.

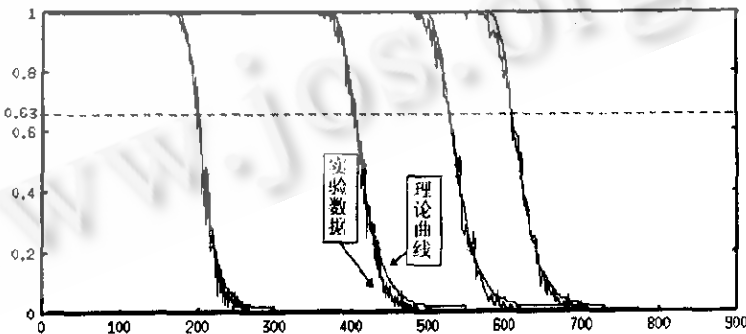


图 2 $P(k) = 1 - (1 - \alpha^{k/N}) \beta^N$ 的逼近效果 ($N=50, 100, 125, 150$)

在图 2 中, x 轴表示子句数 K , y 轴表示可满足的概率 P ,带毛刺的曲线表示实验数据,平滑曲线是 $P(k) = 1 - (1 - \alpha^{k/N}) \beta^N$ 的理论预测曲线 ($\alpha = 0.9604, \beta = 1.1624$). 从图中可以看出,两条曲线很接近,表明了非常好的拟和效果.

3 2-3-SAT 问题的可满足概率

类似于第 2 节中介绍的 3-SAT 问题的可满足概率模型,我们可以得到 2-3-SAT 的可满足概率模型.

定理 4. 设任一 3-子句通过布尔筛的概率为 α_3 ,任一 2-子句通过布尔筛的概率为 α_2 ,任一变量的自由度为 β ,则 2-3 SAT 的可满足概率 P_r 为:

$$P_r(M, K) = 1 - (1 - \alpha_2^M * \alpha_3^K)^{\beta^N}$$

证明:类似于定理 1,易证上式成立,详证略.

定理 5. 2-3-SAT 的相变点处的 2-子句个数 M 和 3-子句个数 K 满足以下线性关系:

$$K + \lambda * M = \rho_0 * N.$$

证明:由定理 3 可知:相变点在二阶导数等于 0 处,由此可得:

$$\begin{aligned} \frac{\partial P_r}{\partial M} &= -\beta^N * (1 - \alpha_2^M * \alpha_3^K)^{\beta^N - 1} * (-\alpha_2^M * \alpha_3^K * \log \alpha_2) \\ &= \beta^N * \alpha_3^K * \log \alpha_2 * \{(1 - \alpha_2^M * \alpha_3^K)^{\beta^N - 1} * \alpha_2^M\} \\ &= C * \{(1 - \alpha_2^M * \alpha_3^K)^{\beta^N - 1} * \alpha_2^M\} \quad (C \text{ 为常数}) \end{aligned}$$

$$\begin{aligned} \frac{\partial^2 P_r}{\partial M^2} &= C * \{\alpha_2^M * (\beta^N - 1) * (1 - \alpha_2^M * \alpha_3^K)^{\beta^N - 2} * (-\alpha_2^M * \alpha_3^K * \log \alpha_2) + (1 - \alpha_2^M * \alpha_3^K)^{\beta^N - 1} * (\alpha_2^M * \log \alpha_2)\} \\ &= C * (1 - \alpha_2^M * \alpha_3^K * \beta^N) \end{aligned}$$

令 $\frac{\partial^2 P_r}{\partial M^2} = 0$, 可得:

$$\begin{aligned} 1 - \alpha_2^M * \alpha_3^K * \beta^N &= 0 \\ \alpha_2^M * \alpha_3^K &= 1 / \beta^N \end{aligned} \tag{2}$$

$$M * \log \alpha_2 + K * \log \alpha_3 = -N * \log \beta$$

上式表明, M 和 K 满足线性关系. 对 K 求二阶偏导数,可以得到同样的结论.

将 2-SAT 和 3-SAT 的相变点作为特殊解代入式(2),可得 $K + \lambda * M = \rho_0 * N$ 成立.

其中

$$\lambda = \log \alpha_2 / \log \alpha_3 \approx 3.12 \dots$$

定理 5 表明, 2-3-SAT 的相变点处的 2-子句个数和 3-子句个数满足线性关系. 其中 λ 是 2 子句和 3 子句的当量, 它的直观含义是指在统计意义下, 一个 2-子句有相当于 λ 个 3-子句的否定能力. 图 3 是对变量数等于 100 的大量 2-3-SAT 实例相变点的统计结果. 图中显示的是可满足概率等于 0.63 的等值线, x 轴表示相变点处的 2-子句数目, y 轴表示相变点处的 3-子句数目. 从图中可以看出很好的线性, 斜率大约为 3.13, ..., 和我们理论上的预测非常接近.

2 3 SAT 相变点的线性关系

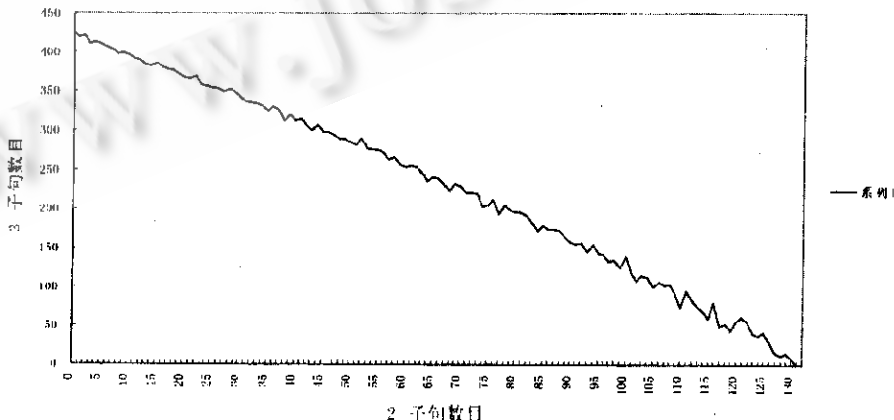


图 3 2-3-SAT 问题相变点处 2-子句个数和 3-子句句数的线性关系

4 2-3-当量的应用

在完备性算法中,最关键处在于如何选取分枝变量.应用2-3-当量可以更好地指导分枝变量的选取,使得形成的子问题避开相变区域.

5 结论

本文在文献[1]的基础上,提出了2-3-SAT的可满足概率的布尔筛模型,进而证明了相变点处的2-子句个数和3-子句个数的线性关系,得到了在统计意义下的一个2-子句有相当于几个3-子句的否定能力的当量.实验结果充分验证了上述结果的正确性.

致谢 在本文的写作过程中,得到了国家智能计算机研究开发中心的张中博士、清华大学的贺思敏博士、柳常青博士的鼓励和建议,在此一并致谢!

参考文献

- 1 Bai Shuo, Bu Dong-bo. Modeling phase transition 3-SAT problem. In: Proceedings of the 2nd International Competition and Symposium on SAT Problem. Beijing, 1996
- 2 Garey M R, Johnson D S. Computers and intractability: a guide to the theory of NP-completeness. San Francisco: Freeman W H and Company, 1979
- 3 Mitchell D, Selman B, Levesque H. Hard and easy distributions of SAT problem. In: Proceedings of the 10th Conference on Artificial Intelligence(AAAI'92). San Jose, CA, July 1992. 459~465
- 4 Dubois O, Carlier J. Probabilistic approach to the satisfiability problem. Theoretical Computer Science, 1991,81
- 5 Selman B. Stochastic search and phase transitions: AI meets physics. In: Proceedings of the International Joint Conference on Artificial Intelligence'95, Extended Abstract. 1995
- 6 Gent I P, Walsh T. Easy problems are sometimes hard. In: Proceedings of the DIMACS Challenge'93. 1993
- 7 Dubois O, André P, Boufkhad Y *et al.* SAT versus UNSAT. In: Proceedings of the DIMACS Challenge'93. 1993

Analysis for Phase Transition of the 2-3-SAT Problem

BAI Shuo BU Dong-bo

(National Research Center for Intelligent Computing System Beijing 100080)

(Dawning Information Technology Industry Corporation Beijing 100080)

Abstract There exists a very interesting feature in SAT (satisfiability) problem. With a fixed numbers of variables, the satisfactory probability of an SAT instance change sharply from 1 to 0 while the number of clauses increasing, and the phase transition point is estimated to be $K \approx 4.3 * N$. The phase transitions are of great importance to the efficient algorithms designing to solve the SAT problem. More generally, the phase transition of 2-3-SAT problem was discussed in this paper. The analysis of the location of the phase transition point of 2-3-SAT shows that there is an linear ratio between a 2-clause and a 3-clause in the sense of the constraint power which could help to design a more powerful heuristic for algorithms designing.

Key words SAT problem, 2-3-SAT, phase transition, 2-3-ratio, constraint power.