

基于关联约束非二元弧一致性的约束满足问题求解*)

袁际军 单汨源 王克喜

(湖南大学工商管理学院 长沙 410082)

摘要 弧一致性算法在二元约束满足问题中取得了成功的应用,但并不能被有效泛化至预处理非二元约束满足问题(NCSP)。本文提出了处理 NCSP 的关联约束非二元弧一致性算法。通过随机 NCSP 生成器产生问题实例,分别采用关联约束非二元弧一致性算法和非二元弧一致性算法进行预处理,并对预处理后的问题实例应用回溯算法进行求解。对比分析采用两种预处理算法和不采用预处理下回溯算法的求解性能,仿真实验结果表明关联约束非二元弧一致性算法可以有效地剔除冗余的约束元组和变量域值,使关联约束非二元弧一致性回溯算法具有更良好的鲁棒性。

关键词 非二元约束满足问题,回溯算法,关联约束非二元弧一致性,随机 NCSP 生成器

Associated Constraint Based Non-binary Arc-consistency for Constraint Satisfaction Problems

YUAN Ji-jun SHAN Mi-yuan WANG Ke-xi

(Management School, Hunan University, Changsha 410082, China)

Abstract Arc consistency has been successfully applied to binary constraint satisfaction problem but cannot be generalized to effectively preprocess non-binary constraint satisfaction problem (NCSP). Associated constraint based non-binary arc-consistency (nACBA) for preprocessing NCSP is proposed in this paper. Some NCSP instances generated by random NCSP generator are first preprocessed by nACBA and non-binary arc-consistency (nAC) respectively, and then solved by backtracking algorithm. Performance of backtracking, nACBA based backtracking and nAC based backtracking are evaluated and compared. The computational results indicate that nACBA based backtracking is even more robust than the other two algorithms, which means associated constraint based non-binary arc-consistency can more effectively eliminate redundant variable values and constraint tuples.

Keywords Non-binary constraint satisfaction problem, Backtracking, Associated constraint based non-binary arc-consistency, Random NCSP generator

1 引言

约束满足问题(CSP)是人工智能领域的热点问题,具有广泛的应用领域。目前对约束满足问题的研究大多数集中在二元约束满足问题(BCSP)^[1-3],但一些现实问题,如 N-皇后问题和产品配置问题,可以更自然地采用非二元约束满足问题(NCSP)进行建模^[4]。NCSP 是一个典型的 NP-完全问题,原则上总可以被归约为 BCSP,如采用对偶图法(dual graph)和隐藏变量法(hidden variable)^[5,6]。将 NCSP 归约为 BCSP 的优点是可以利用大量现成的启发式搜索算法以及当前仅适用于 BCSP 的算法来间接求解 NCSP。然而,把 NCSP 转化为 BCSP 常常会引入新的变量和增加变量的值域,导致问题求解的搜索空间急剧增大;并且强制二元化还会产生不自然的陈述和理解上的困难。由于非二元约束满足问题是二元约束满足问题的泛化,二者在求解方法上具有较强的联系,许多二元约束满足问题的求解方法经过改进后可以直接用来求解非二元约束满足问题,如回溯算法以及变量排序和变量值排序等启发式算法^[7]。一致性技术作为一种预处理机制在二元约束满足问题领域中得到了广泛成功的应用,如二元约束前向检测算法和维持弧一致性算法(MAC)^[7]等。因此很多学者把应用于二元约束的弧一致性概念经过简单的扩展来处理非二元约束满足问题^[8,9]。这种简单的扩展显现出了较多的问题,它并不能像在二元约束满足问题中能有效地剔除不良解。同时,非二元约束满足问题中还分布地存在许多矛盾的或冗余的约束元组,对这些约束的处理可以进一步促进对变量域

值的剔除,从而加速对问题的求解搜索效率。

为了解决泛化弧一致性技术在非二元约束满足问题中存在的问题,本文提出了关联约束非二元弧一致性算法(associated constraint based non-binary arc-consistency; nACBA)。nACBA 作为一种预处理技术,既可以在非二元约束满足问题的搜索求解前对变量的域值和约束的元组进行预处理,也可以在搜索的过程中进行预处理。为了验证关联约束非二元弧一致性算法的预处理性能以及这种预处理带来的时间成本,本文在求解前先分别用非二元约束弧一致性算法和关联约束非二元弧一致性算法预处理同样的一组随机 NCSP 实例,然后采用回溯算法求解;同时对没有经过预处理的相同实例也采用回溯算法进行求解,并分析比较不同预处理算法带来的效果。

本文第 2 部分提出非二元约束满足问题的相关模型和概念;第 3 部分分析传统非二元约束弧一致性方法的不足,提出 nACBA 算法,并分析算法的复杂性;第 4 部分介绍仿真实验,开发一个非二元约束满足问题的随机生成器,并提出比较不同算法的相关指标;最后得出相关的结论。

2 约束满足问题的相关概念

2.1 约束满足问题的描述

一个约束满足问题可以定义为一个三元组 (X, D, C) : X 表示一组变量的集合,记 $X = \{X_i | i \in \{1, 2, \dots, n\}\}$, n 表示不同变量的个数; D 表示相应变量值域的集合, $D = \{D_i | i \in \{1, 2, \dots, n\}\}$, D_i 表示变量 X_i 的值域,本文仅研究变量值域是离

*)基金项目:国家自然科学基金项目(70671037);高等学校博士学科点专项科研基金项目(20050532005)。袁际军 博士研究生,主要从事大规模定制、产品配置和约束编程等方面的研究。

散的情况; C 表示作用在不同变量上的约束集合, 记 $C = \{C_{S_j} | j \in \{1, 2, \dots, m\}\}$, m 表示约束的个数, S_j 表示第 j 个约束 C_{S_j} 的作用范围, S_j 具有唯一性, 且 $S_j = \{X_{i_k} | k \in \{1, 2, \dots, |S_j|\}\}$, $C_{S_j} \subseteq D_{i_1} \times D_{i_2} \times \dots \times D_{i_{|S_j|}}$ 且 $t = |S_j|$ 。约束作用范围中的各个变量相应域值的一个可行组合称为该约束的一个元组。给定一个约束满足问题, 如果 $\forall i$, 都有 $|S_i| = 2$, 则称该约束满足问题为二元约束满足问题 (BCSP), 否则该约束满足问题为非二元约束满足问题 (NCSP)。

2.2 约束满足问题的可行解

定义 1 给定一个约束 C_{S_i} , 将作用范围 S_i 中的某些变量及其在约束中对应的域值组合加以保留, 并将其他变量及其在约束中对应的域值组合去掉, 形成一个作用在保留变量集合上新的约束关系, 这种新的约束关系被称之为约束 C_{S_i} 在保留变量集合上的约束投影。假如保留变量集合为 S_j , 则可以将 C_{S_i} 在 S_j 上的约束投影运算记为 $\pi_{S_j}(C_{S_i})$, 经过投影运算后形成一个新的约束, 新约束的作用范围 S_j 是原有约束作用范围 S_i 的子集。

设存在一个约束满足问题 $P = (X, D, C)$ 及对一个变量集合 X 的赋值 I_X , 如果赋值 I_X 满足约束集合 C 中的所有约束, 则称赋值 I_X 是约束满足问题 P 的一个可行解。而将所有满足约束集合中的所有约束的变量集合 X 的赋值称为约束满足问题 P 的所有可行解, 记为 $\text{solution}(P)$; 则约束满足问题 P 的所有可行解用投影运算定义为: $\text{solution}(P) = \{I_X = (d_1, d_2, \dots, d_n) | d_i \in D_{X_i}, \forall C_{S_i} \in C, \pi_{S_i}(I_X) \in C_{S_i}\}$ 。

3 关联约束非二元弧一致性 (nACBA)

3.1 非二元弧一致性 (non-binary arc-consistency: nAC)

在约束满足问题中, 经常存在一些变量的某些域值决不会在最终的可行解中出现, 预先剔除这些域值会显著缩小问题求解时的搜索空间。弧一致性技术作为一种常用的方法, 在二元约束满足问题中得到了广泛的研究, 如 AC-1, AC-2, AC-3, AC-4, AC-5, AC-6, AC-7^[10]。对二元约束满足问题而言, 若一个约束弧一致, 则说明这个约束中的任何一个变量的现存任何一个域值都可以扩展到另一个变量中的某个域值, 使值组满足这个约束或和这个约束一致, 也就是在另一个变量中存在一个域值支持当前试图扩展的域值。

类似地, 非二元弧一致性定义^[8]如下: 存在一个非二元约束满足问题 $P = (X, D, C)$, 设约束集合中的任意一个约束 $C_{S_i} \in C$ 和约束的作用范围 $S_i \subseteq X$ 中的任意一个变量 $X_i \in S_i$, 对变量 X_i 中的某个域值 $d_i \in D_i$, 如果存在 C_{S_i} 所允许的一个域值组合 $t_i \in C_{S_i}$, 使 $d_i \in t_i$, 则称变量 X_i 的域值 d_i 相对于约束 C_{S_i} 是非二元弧一致的; 如果变量 X_i 中的每个域值都相对于约束 C_{S_i} 是非二元弧一致的, 则称变量 X_i 相对于约束 C_{S_i} 是非二元弧一致的; 如果约束 C_{S_i} 作用范围中所有的变量 X_i 都相对于约束 C_{S_i} 是非二元弧一致的, 则称约束 C_{S_i} 是非二元弧一致的; 若非二元约束满足问题中所有的约束都是非二元弧一致的, 则称该约束满足问题是非二元弧一致的。

然而上述这种从二元弧一致性简单泛化至非二元弧一致性存在较多的不足, 它不能像二元弧一致性处理二元约束满足问题那样能有效地剔除变量中的冗余域值, 更不能剔除约束中的冗余约束元组。事实上, 剔除冗余的约束元组, 可进一步促进变量中的冗余域值的剔除; 而且, 约束中冗余约束元组的剔除, 还能进一步影响约束满足问题求解时的约束检查次数、回溯次数等, 间接地减少问题求解时的搜索空间, 增加对问题的求解效率。

例 1: 如图 1 所示, 给定一个非二元约束满足问题 $P = (X, D, C)$, 其中 $X = \{X_1, X_2, X_3, X_4\}$, $D_1 = \{1, 2, 3\}$, $D_2 = \{1, 2\}$, $D_3 = \{1, 2\}$, $D_4 = \{1, 2\}$, $C = \{C_{S_1}, C_{S_2}\}$, $S_1 = \{X_1, X_2, X_3\}$, $S_2 = \{X_2, X_3, X_4\}$, $C_{S_1} = \{\langle 1, 1, 1 \rangle, \langle 2, 2, 2 \rangle\}$, $C_{S_2} = \{\langle 1, 2, 1 \rangle, \langle 2, 1, 2 \rangle\}$ 。

当采用非二元弧一致 (nAC) 预处理该问题时, 由于 $X_1 \in S_1$, 在 $X_1 \leftarrow 3$ 时, C_{S_1} 中不存在任何元组支持 X_1 取 3 的赋值, 因此域值 3 被从 X_1 的值域中剔除。当所有的约束非二元弧一致时, 所有变量中只有域值 3 被从 X_1 的值域中剔除。

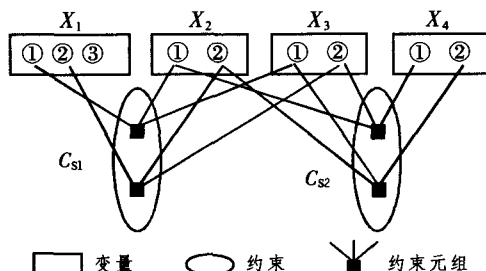


图 1 非二元约束满足问题的一个实例

实际上, 例 1 是一个无解的非二元约束满足问题。如当 $X_1 \leftarrow 1$ 时, 尽管在 C_{S_1} 中存在 $(X_1 \leftarrow 1) \wedge (X_2 \leftarrow 1) \wedge (X_3 \leftarrow 1)$ 被满足, 但 $(X_2 \leftarrow 1) \wedge (X_3 \leftarrow 1)$ 不能被 $\pi_{(X_2, X_3)}(C_{S_2})$ 同时满足, 因此 X_1 的域值 1 不可能在问题的可行解中出现, 预先剔除 X_1 的域值 1 有利于减少问题求解时的搜索空间; 同时可知, C_{S_1} 中的 $\langle 1, 1, 1 \rangle$ 是一个多余的约束元组, 从 C_{S_1} 中剔除约束元组 $\langle 1, 1, 1 \rangle$ 不会影响约束满足问题的求解, 但却有利于减少求解时的约束检查次数; 同理可以剔除 X_1 的域值 2 和 C_{S_1} 中元组 $\langle 2, 2, 2 \rangle$, 此时 X_1 的值域为空并且 C_{S_1} 为空集, 两者均可提前判断该约束满足问题无解。

3.2 关联约束非二元弧一致性 (nACBA)

鉴于非二元弧一致性在预处理非二元约束满足问题时的不足, 本文提出了关联约束非二元弧一致方法。

定义 2 $\exists C_{S_i}, C_{S_j} \in C$, 如果 $S_i \cap S_j \neq \emptyset$, 则称 C_{S_i} 和 C_{S_j} 互为关联约束。

定义 3 $\exists C_{S_i}, C_{S_j} \in C$, 如果 C_{S_i} 和 C_{S_j} 互为关联约束, 令 $C_{S_i \cap S_j} = (\pi_{S_i \cap S_j}(C_{S_i})) \cap (\pi_{S_i \cap S_j}(C_{S_j}))$, 若 $C_{S_i \cap S_j} \neq \emptyset$, 则称 C_{S_i} 和 C_{S_j} 互为支撑约束, 否则称 C_{S_i} 和 C_{S_j} 互为非支撑约束。

定义 4 $\exists C_{S_i}, C_{S_j} \in C$, 且 $\exists t \in C_{S_i}$, 若 $(\pi_{S_i \cap S_j}(t)) \in C_{S_i \cap S_j}$, 则称约束元组 t 被 C_{S_j} 支撑, 否则称约束元组 t 不被 C_{S_j} 支撑。

定义 5 给定一个非二元约束满足问题 $P = (X, D, C)$, $\forall C_{S_i} \in C$, 存在一个约束 $C_{S_j} \in C$ 使得 C_{S_i} 和 C_{S_j} 互为支撑约束。如果任一约束元组 $t \in C_{S_i}$, 若约束元组 t 不被 C_{S_j} 支撑, 则从 C_{S_i} 中剔除元组 t , 称为约束 C_{S_i} 相对元组 t 的一次修改。对于 $\forall C_{S_k} \in C$, 若 C_{S_k} 不能再被修改, 则称 P 满足关联约束一致性; 此时若 $\exists C_{S_i} \in C$ 且 $C_{S_i} = \emptyset$, 则问题 P 无解, 否则再对满足关联约束一致性的 P 实施非二元弧一致性, 若问题 P 达到非二元弧一致性, 则称约束满足问题 P 此时满足关联约束非二元弧一致性。

关联约束非二元弧一致性算法描述如图 2。

定理 1 给定一个非二元约束满足问题 $P = (X, D, C)$, 若 P 满足关联约束一致性, 对任意两个互为支撑约束的 C_{S_i} 和 C_{S_j} , 则必有:

$C_{S_i} = C_{S_i} \times C_{S_j}$ 且 $C_{S_j} = C_{S_j} \times C_{S_i}$ 。(注: $\times$ 表示“半连接”^[11])。

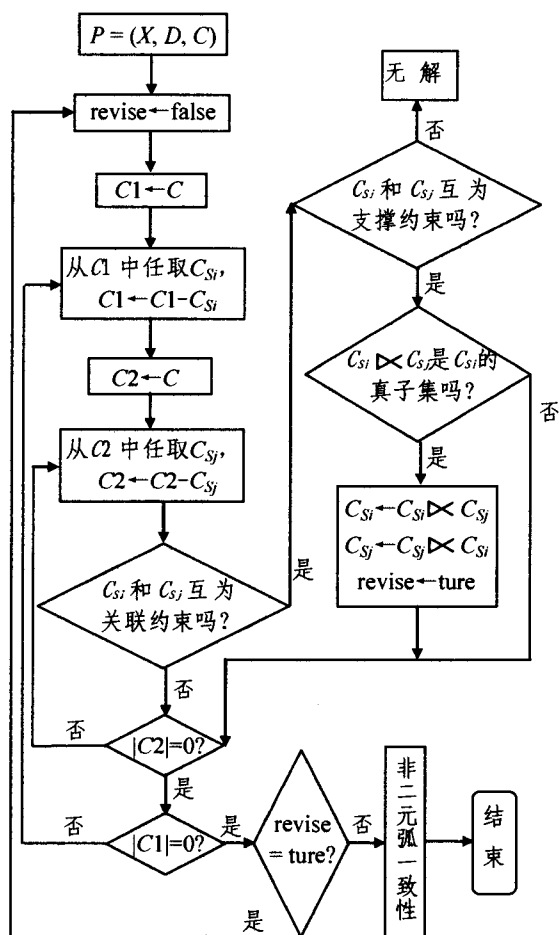


图 2 关联约束非二元弧一致性

证明略, 根据定义 5 直接可证。

3.3 关联约束非二元弧一致性的复杂性分析

给定一个非二元约束满足问题 $P = (X, D, C)$, 令 $|X| = n$, $\max_{1 \leq i \leq n} (|D_i|) = k$, $|C| = e$, $\max_{1 \leq j \leq e} (|C_{S_j}|) = t$, $\max_{1 \leq j \leq e} (|S_j|) = r$, 则关联约束一致性的时间复杂性为 $O(te^3)$, 非二元约束一致性的时间复杂性为 $O(nek^{r+1})$, 而关联约束非二元弧一致性算法由关联约束一致性和非二元弧一致性算法组成, 因此关联约束非二元一致性的时间复杂性为 $O(te^3 + nek^{r+1})$ 。

4 仿真实验

求解约束满足问题时通常需要得到问题的一个解或所有的解,而关联约束非二元弧一致性作为一种预处理算法时并不适宜求解出问题的可行解,它经常需要与其他的搜索算法相结合才能发挥强大的作用;因此,本文在对预处理后的问题搜索求解时主要应用回溯算法(BT)。回溯算法是在对问题进行搜索的过程中,从第一个变量开始赋值,通过对新变量的赋值由当前的部分赋值状态延长并转变到一个新的由延长的部分赋值所表示的状态,在对每一个变量进行赋值后,并不立即对新的变量进行赋值,而是检查当前状态的部分赋值是否满足约束,如果满足,则继续对下一个新变量进行赋值;如果不满足,则对当前变量进行新的赋值,产生新的状态并进行约束检查,如果当前赋值变量的所有域值都不能使当前的部分赋值满足所有约束,那么这个部分赋值(不包括当前变量)将不会出现在任何问题的解中,而其相应的状态也称为无解状态,这时搜索将返回上一个进行赋值的变量并进行新的赋值

以试图避免无解状态的产生。

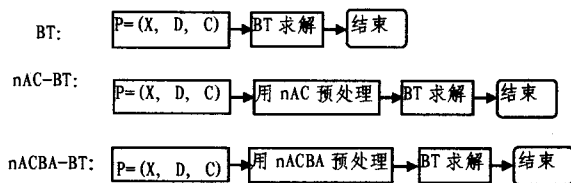


图3 BT, nAC-BT 和 nACBA-BT 三种算法的处理过程

对非二元约束满足问题进行预处理,其目的在于使预处理时间和预处理后的搜索求解时间总和比无预处理而直接求解的时间更少。为了验证关联约束非二元弧一致性算法的预处理性能是否优越于非二元弧一致性算法,本文分别采用回溯算法(BT)、非二元弧一致性回溯算法(nAC-BT)和关联约束非二元弧一致性回溯算法(nACBA-BT)分别对比求解相同的非二元约束满足问题。三种算法的处理过程如图3。

4.1 仿真实验数据的产生

广泛验证比较本文各算法的求解性能需要使用随机生成的非二元约束满足问题实例,因此,作者根据如下参数用 MATLAB 6.5 开发了一个随机非二元约束满足问题生成器(随机 NCSP 生成器): (1) n : 表示变量的个数; (2) k : 表示变量值域的规模,该随机 NCSP 生成器假定产生的 NCSP 中的所有变量均有相同的值域规模; (3) r : 表示约束维数,该随机 NCSP 生成器产生的 NCSP 中的所有约束均有相同的维数; (4) d : 表示约束密度, $d = |C| / \binom{n}{r}$; (5) $tightness$: 表示约束的松紧度, $tightness = (k^r - \text{约束中允许的元组个数}) / k^r$ 。区别于其他的随机 NCSP 生成器,该生成器能保证随机产生 NCSP 的约束中的每个元组都不重复。

4.2 评价指标

本文采用了如下常用的评价指标^[9]:(1)求得第一个解时 CPU 的平均运行时间(fst-CPU);(2)求得所有解时 CPU 的平均总运行时间(all-CPU);(3)求得所有解时的平均回溯次数(bkts);(4)求得所有解时的平均约束检查次数(ccks)。由于篇幅限制,本文省略了其他的一些评价指标,如预处理时的平均约束元组剔除比率、预处理时平均域值剔除比率、预处理时平均 CPU 时间以及相关的标准差等。

4.3 仿真实验条件设定与实验结果

本文分别对上述三种算法的性能进行了测试,根据非二元约束满足问题的求解特性^[12],本文设定实验参数为: $n=8$, $k=6$, $r=3$,同时分别固定密度 d 为 0.3, 0.5 和 0.7, 在每种密度下, 分别固定松紧度 $tightness$ 为 0.3, 0.35, 0.4, 0.45, 0.475, 0.5, 0.525, 0.55, 0.575, 0.6, 0.625, 0.65, 0.7, 0.75, 0.8。在每种密度和相应的松紧度下分别随机产生 15 个 NC-SP 实例, 共产生 675 个实例进行实验。实验条件为 Windows XP, Celeron(R) CPU 2.5GHz, 内存 512 MB, 编程语言为 MATLAB 6.5。仿真实验结果如图 4 所示。

4.4 仿真实验结果分析

从上面的仿真实验结果可知, BT 和 nAC-BT 的求解时间曲线几乎重合, 从放大的图形可知 nAC-BT 所描绘的曲线总是略高于 BT 所描绘的曲线, 说明在对非二元约束满足问题求解之前用 nAC 预处理所耗费的时间代价, 超过了预处理导致的搜索空间的减少而引起搜索时间的减少量; 另外, BT 与 nAC-BT 的平均回溯次数和平均约束检查次数的数据描绘曲线几乎重合, 说明 nAC 对 NCSP 预处理时几乎没有剔除冗余的赋值, 无法达到其在 BCSP 中良好的预处理性能。由此可

见,在对 NCSP 求解之前用 nAC 预处理并不能有效地改进问题的求解搜索性能,说明 nAC-BT 相对于 BT 的改进是失效的。

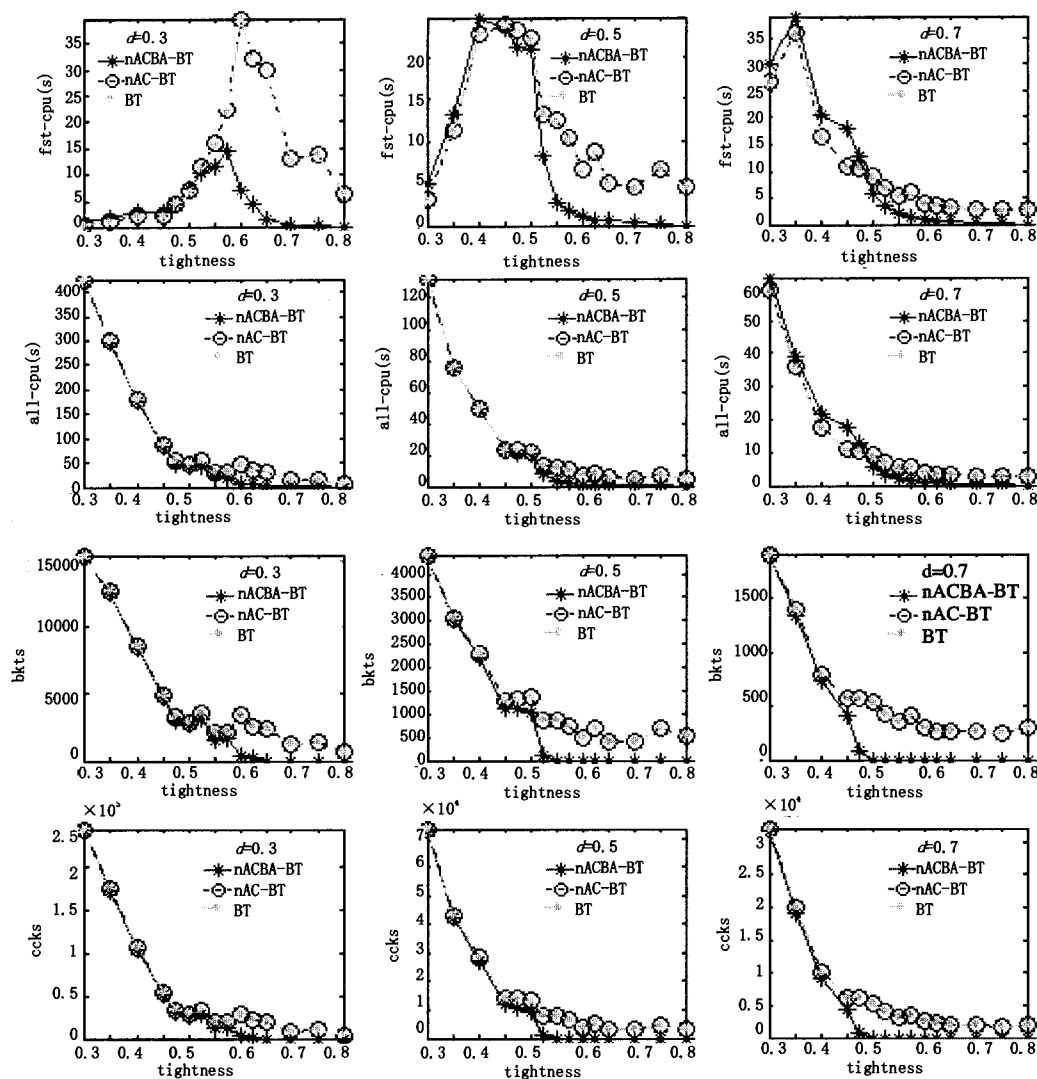


图4 nACBA-BT, nAC-BT 和 BT 三种算法实验结果比较

在松紧度高于 0.45, 密度分别为 0.3, 0.5 和 0.7 时, nACBA-BT 的求解时间曲线明显地位于 BT 和 nAC-BT 的求解时间曲线的下方, 说明在这个区间 nACBA-BT 算法所花的时间要明显小于 BT 和 nAC-BT 所花的时间。这是由于在该区间 nACBA 的预处理导致了多余冗余的域值和约束元组被剔除, 从而使搜索求解时 nACBA-BT 引起的平均回溯次数和平均约束检查次数均明显地小于相应的 BT 和 nAC-BT, 从而使 nACBA-BT 算法所花的时间要明显小于 BT 和 nAC-BT 所花的时间。

在松紧度小于 0.45, 密度分别为 0.3, 0.5 和 0.7 时, nACBA-BT 的求解时间曲线位于 BT 和 nAC-BT 的求解时间仿真曲线的上方, 同时从上面的仿真结果可知, nACBA-BT 引起的平均回溯次数和平均约束检查次数均几乎相等与相应的 BT 和 nAC-BT, 说明此时几乎没有域值或约束元组被剔除。这是因为松紧度越小, 约束允许的约束元组组合越多, 相应可行解的个数急剧增多, 域值或约束元组的剔除也变得越困难, nACBA 预处理所花费的时间代价增大, 从而使 nACBA-BT 算法所花的时间大于 BT 和 nAC-BT 所花的时间。进一步分析可知, nACBA 预处理所花费的时间要稍大于 nAC 预处理所花费的时间, 这与前述复杂性分析的结果是相吻合的, 但相对于搜索求解时间而言, 预处理时间几乎可以忽略不计。

根据上述分析可知, 在对非二元约束满足问题求解之前进行预处理, 松紧度越大, nACBA 相对于 nAC 有着更良好的预处理性能, 能够剔除更多冗余域值和约束元组, 而密度的变化对预处理的性能影响不大; 同时也说明, 简单地泛化二元约束弧一致性方法至非二元弧一致性(nAC)存在较大的不足。

结束语 针对二元约束弧一致性被简单地泛化至非二元弧一致性的不足, 本文提出了关联约束非二元弧一致性。同时比较了回溯算法、非二元弧一致性回溯算法和关联约束非二元弧一致性回溯算法, 实验表明关联约束非二元弧一致性回溯算法较前二者有更良好的性能。这说明关联约束非二元弧一致性能显著地剔除冗余的域值和约束元组, 明显地改进了非二元弧一致性预处理非二元约束满足问题的不足。

本文提出的算法并不仅限于实验中三元约束的情况, 对于更高维数的约束满足问题该方法亦具有通用性, 并能得到相似的实验结果。本文并没有研究在非二元约束满足问题的搜索求解过程中运用一致性技术预处理的情况, 如对比研究非二元前向检测算法和基于关联约束非二元弧一致性的改进非二元前向检测算法, 这是作者下一步进行研究的方

参考文献

- [1] Craenen B G W, Eiben A E, Van Hemert J L Comparing evolution-

- ary algorithms on binary constraint satisfaction problems. IEEE Transactions on Evolutionary Computation, 2003, 7(5): 424-444
- [2] Manuel B, Daniel K. Locally consistent constraint satisfaction problems with binary constraints. Lecture Notes in Computer Science, 2005, 3787: 295-306
- [3] 杨轻云,孙吉贵,张居阳. 最大度二元约束满足问题粒子群算法. 计算机研究与发展, 2006, 43(3): 436-441
- [4] Michel A, Khaled H-H, Guillaume M, et al. Mass customization and configuration: Requirement analysis and constraint based modeling propositions. Integrated Computer-Aided Engineering, 2003, 10(2): 177-189
- [5] Fahiem B, Chen Xingquang, Peter V B, et al. W Binary vs. n-binary constraints. Artificial Intelligence, 2002, 140(1/2): 1-37
- [6] Nikolaos S, Kostas S. Binary encodings of non-binary constraint satisfaction problems: Algorithms and experimental results. Journal of Artificial Intelligence Research, 2005, 24(7): 641-684
- [7] Bessiere C, Reagin J C. MAC and combined heuristics: Two reasons

- to forsake FC (and CBJ?) on hard problems//Proceedings of Principles and Practice of Constraint Programming - CP96, 1996, 61-75
- [8] Bessiere C, Messeguer P, Freuder E C, et al. On forward checking for non-binary constraint satisfaction. Artificial Intelligence, 2002, 141(1/2): 205~224
- [9] Lal A. Neighborhood interchangeability for non-binary CSPs & application to databases. Ph. D. Dissertation. Lincoln: University of Nebraska, 2005
- [10] Liu Zhe. Algorithms for Constraint Satisfaction Problems (CSPs). Master's Thesis. Canada: University of Waterloo, 1998
- [11] 施伯乐,丁宝康,汪卫. 数据库系统教程(第2版). 北京: 高等教育出版社, 2005
- [12] Lal A, Choueiry B Y, Freuder E C. Neighborhood interchangeability and dynamic bundling for non-binary finite CSPs//Proceedings of the 20th National Conference on Artificial Intelligence and the 17th Innovative Applications of Artificial Intelligence Conference, 2005: 397-404

(上接第126页)

的特征向量;

(2)根据距离函数计算待分类文本向量 d 和训练文本向量的相似度,可以使用两向量之间的欧氏距离 $|d-d_i|$ 进行计算;选择与 d 相似度最大(距离最小)的 k 个文本作为 d 的 k 个最近邻;

(3)根据 d 的 k 个最近邻依次计算文本类别 $c_1, c_2, \dots, c_r$ 相应的权重,计算公式为:

$$p(x, c_j) = \sum_{i=1}^k \text{sim}(d_i, d) \delta(d_i, d)$$

其中, $\text{sim}(d_i, d)$ 表示文本 d_i 与文本 d 之间相似度,

$$\delta(d_i, d) = \begin{cases} 1, & d_i \text{ 是属于 } c_j \text{ 的文本} \\ 0, & d_i \text{ 不属于 } c_j \end{cases} \text{ 为示性函数。}$$

(4)将待分类文本 d 归入权重最大的类别。

考虑到 KNN 分类器对训练集数据倾斜问题的敏感性,我们使用如下距离公式^[18]对 KNN 算法进行改进,即以

$$d(d_i, d) = \frac{|d_i - d|}{\sqrt{M(d_i)} \sqrt{M(d)}} \quad (12)$$

替换算法中步骤(2)的欧氏距离。这里 $M(d_i)$ 、 $M(d)$ 表示 d_i 、 d 与其它文本之间欧式距离的平均值。

(12)式使得样本分布密集区域的样本之间距离被放大,而稀疏区域的样本之间距离被缩小,从而有效降低了由于样本点分布不均而对分类器的影响。

3.2 试验结果及其分析

本文对上述方法的分类效率进行了试验。试验数据为新浪网上下载的 3460 篇文本,其中分为文学(560 篇)、金融(579 篇)、体育(651 篇)、计算机(572 篇)、新闻(495 篇)以及军事(603 篇)。试验时采用 3 分交叉试验法,将 3460 篇文本平均分为 3 份,2 份为训练集,1 份为测试集;每份轮流作为测试集循环测试 3 次,取平均值为测试结果。经过文本预处理后的原始特征集特征维数为 1321,使用(7)式进行特征抽取,使用(11)式进行投影变换,取投影空间维数 $s=85$,效果评估函数使用常用的查准率、查全率和 F_1 测试值:

表 1 试验结果统计

类别	文学	金融	体育	计算机	新闻	军事
查准率	86.1%	89.2%	91.3%	84.1%	85.7%	89.2%
查全率	87.3%	85.4%	82.5%	88.6%	90.2%	84.7%
F_1 测试值	86.7%	87.3%	86.7%	85.8%	87.9%	86.9%

查准率 = 分类的正确文本数 / 实际分类文本数;

查全率 = 分类的正确文本数 / 应有文本数;

$$F_1 = \frac{\text{查准率} \times \text{查全率} \times 2}{\text{查准率} + \text{查全率}}$$

分类结果统计如 1 表所示。

平均查准率为 87.6%,查全率为 86.5%; F_1 测试值为 87.1%,效果还是令人满意的。

结束语 特征降维问题是文本处理所必须面对的主要问题之一,是制约文本分类效率的瓶颈。本文讨论了 PCA 在文本分类中的应用。论证了投影特征的选取方法及其相关的理论依据。讨论了 PCA 与 SVD 这两种基于特征抽取的降维方法异同点以及在分解矩阵为正定阵的条件下两者之间的等价性。特征选择和特征抽取从不同的立足点出发对文本特征进行压缩,两种模式各有其优点和不足。如何将两种模型进行合理的结合,构造组合型的特征降维模型应该是提高特征降维效率的思路之一,这也是我们今后要进行的研究工作。

参考文献

- [1] 廖莎莎,等. 中文文本分类中基于概念屏蔽层的特征提取方法. 中文信息学报[J], 20(3): 22-28
- [2] 苏金树,等. 基于机器学习的文本分类技术研究进展. 软件学报[J], 2006(9): 1848-1859
- [3] Cover T M. The best two independent measurements are not the two best. IEEE Transactions on Systems, Man, and Cybernetics, 1974, 4: 116-117
- [4] Dumais S, Platt J, Heckerman D, et al. Inductive learning algorithms and representations for text categorization//Proceedings of the CIKM-98, 7th ACM International Conference on Information and Knowledge Management. Bethesda, MD, 1998: 148-155
- [5] 宋枫溪,等. 统计模式识别中的维数削减与低损降维. 计算机学报[J], 2005(11): 1915-1922
- [6] 陈伏兵,等. 小样本情况下 Fisher 线性鉴别分析的理论及其验证. 中国图象图形学报[J], 10(8): 984-991
- [7] 吴春国,等. 关于 SVD 与 PCA 等价性研究. 计算机学报[J], 2004(2): 286-288
- [8] 张生亮,等. 基于类间散布矩阵的二维主分量分析. 计算机工程[J], 2006(6): 44-46
- [9] 张志佳,等. 基于线性投影的代数空间降维分析. 计算机工程[J], 2005(11): 25-27
- [10] 刘海峰,等. 基于潜在语义空间的文本检索问题研究. 情报科学[J], 2007(5): 748-753
- [11] 刘海峰,等. 基于 Web 的文本检索位置加权模型研究. 情报科学[J], 2007(3): 451-455
- [12] 王文胜,等. 一种基于奇异值分解的特征抽取方法. 电子与信息学报[J], 2005(2): 294-297
- [13] Duda R O, Hart P E, Store D G. Pattern Classification, 2nd Edition. New York: John Wiley & Sons, 2001: 117-123
- [14] 张明淳. 工程矩阵理论[M]. 南京: 东南大学出版社, 1995: 116-133
- [15] 王粤芳,等. 高等代数[M]. 北京: 高等教育出版社, 2003: 377-381
- [16] 周水庚,等. 隐含语义索引及其在中文文本信息处理中的应用[J]. 小型微型计算机系统, 2001(22): 240-243
- [17] Yang Yiming, Liu Xin. A re-examination of text categorization methods//Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval. 1999: 42-49
- [18] 王和勇,等. 基于聚类和改进距离的方法在数据降维中的应用. 计算机研究与发展[J], 2006, 43(8): 1485-1489