证据

A.5 命题 8 的证明

鉴于X = +∞,即问题1中没有模型约束，推导出附录A.2中的相似比对。

当同时给定 \ = 0 时，时间约束 Θ(r)≤\=0 在对齐的元组中强制执行相同的时间戳，即r[U1个] = r[U2个] = • • • =r[U米]，这正是附录 A.2 中等式对齐所要求的。

A.6 定理 1 的证明

我们首先表明问题出在 NP 中。给定一个对齐的实例
,这三个条件都可以在多项式时间内得到验证。条件 (1) 通过将元组相互比较来验证 O(|R|2个)时间。回想一下 |R|≤分钟{|吨1个|, |吨2个|，。. . , |吨<|} =G。对于条件（2），时间约束\可以通过遍历中的元组来检查O(|R|)时间和模型约束X通过搜索每个元组的最近邻居来检查r∈R在O(|R|日志 |米|) 时间。条件 (3) 可以简单地通过比较 | 来检查。R|和^。

接下来，我们将从 < = 3 的场景开始，并从最大3维匹配[17、19]，卡普的 21 个 NP 完全问题之一 [20]，到我们的对齐问题。最后，我们可以证明缩减也适用于 <≥4，从而证明NP-硬度。结合问题在NP中，我们可以得出问题的NP完全性。

让甲、乙、丙是有限的不相交集，并且Pc⊆A×乙×C，IE，Pc= {(a, b, c)|一个∈一个，乙∈乙丙∈C}。为了p1个（A1个,b1个,C1个),p2个（A2个,b2个,C2个) ∈, 我们说p1个和p2个在某个坐标上相交，表示为 p1个↔p2个, 如果A1个=A2个,b1个=b2个或者C1个=C2个. 我们用P⊆Pc
表示一个 3 维匹配，如果没有元素P与其他人相交，即∀p1个,p2个∈, p1个=p2个. 这最大3维匹配就是找到匹配P∗在所有可能的值中P拥有最多的三元组。判定问题是，给定一个整数^，判定是否存在3维匹配P这样 |P|≥^。

举个例子最大3维匹配问题Pc，为了在模型约束下构建相似性对齐问题，我们首先设置 \ = +∞,即，任何两个对齐的元组都满足时间约束。然后，我们创建吨1个,吨2个,吨3个

在我们的问题中，通过为它们分配唯一的时间戳。同时，让每个吨1个8个∈吨1个相当于A我∈A，每个吨2个9∈吨2个
相当于bj∈乙和每个吨3：∈吨3个相当于Ck∈C。接下来，如果（A我,bj,Ck) ∈Pc, 我们指定吨1个8个[V1个],吨2个9[V2个]和吨3：[V3个] 值的独特组合。然后，我们设置回归模型以满足公吨1个8个[V1个],吨2个9[V2个]) =吨3：[V3个]. 让模型约束X = 0，因为每个 (吨1个8个[V1个],吨2个9[V2个],吨3：[V3个])是独一无二的，我们有Rc= {(吨1个8个,吨2个9,吨3：)|公吨1个8个[V1个],吨2个9[V2个]) =吨3：[V3个]}。因此，我们得到（吨1个8个,吨2个9,吨3：）∈Rc当且仅当 (A我,bj,Ck) ∈Pc.

接下来，我们将证明，对于每个满足的 3 维匹配,我们有 |P|≥^,当且仅当对齐的实例R = {(吨1个8个,吨2个9,吨3：)|公吨1个8个[V1个],吨2个9[V2个]) =吨3：[V3个]}对应于P 在我们的问题中也是满足（1）三个候选键的集合 ü1个,ü2个,ü3个, (2) 时间约束 Θ(r)≤ \和模型约束 Δ(R M）≤X每个都满意r∈,和 (3) |R|≥^。

首先，根据定义，我们假设 |P|≥^。自从 ∀p1个,p2个∈, p1个=p2个, 对于相应的,我们将有∀r1个,r2个∈
，1个-r2个，因此满足条件（1）。对于条件（2），首先

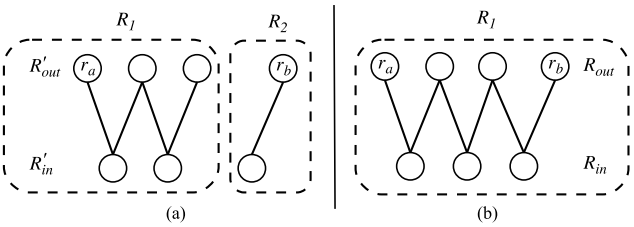

图 10：oa 路径的 (a) 断开连接图和 (b) 连接图示例

回想一下，我们通过假设时间约束足够大来构建对齐问题。此外，我们生成Rc通过满足模型约束，从而自动满足条件（2）。最后，由于我们有 (吨1个8个,吨2个9,吨3：) ∈R当且仅当 （A我,bj,Ck) ∈P, |R| = |P|≥^成立，满足条件（3）。

相反，假设 |R|满足条件（1）、（2）、（3）。按照类似的步骤，根据条件（1），∀r1个,r2个∈，1个-r2个， 相应的P有p1个=p2个,∀p1个,p2个∈P.接下来，根据条件（3），|R|≥^。自从 （吨1个8个,吨2个9,吨3：) ∈R当且仅当 (A我,bj,Ck) ∈,我们会得到 |P| = |R|≥^。满足3维匹配问题的条件。

A.7 命题 2 的证明

我们首先表明模型约束下问题 1 相似性对齐的过程等同于 < = :的 :-Set Packing (k-SP) 问题。给定一个地面集 + 和一个集合小号的集合，每个集合包含：来自 + 的元素，k-SP 问题是找到的最大子集合小号所以它们成对不相交。在我们的场景中，让来自原始数据的所有单元组的集合为

+ , 然后让 '2个相当于S.这两个问题的等价性是直观的，因为它们都最大化了目标集。

接下来，我们的对齐搜索（算法 2）遵循以下框架d-:-SP 问题 [18, 24] 的最优局部搜索算法，它被证明具有有界近似比。根据 [18]中的证明，对于 <-维时间序列和参数 d对于对齐搜索，近似比率的界限b获得。

A.8 命题 3 的证明

条件（4）要求任何元组A>直流电∈'>直流电必须至少与 ' 中的一个元组重叠8=. 假设一个元组A>直流电不与 ' 中的任何元组重叠8=. 结合条件（3），A>直流电不与 ' 中的任何元组重叠乙<. 这是不可能的，因为 '乙<已确保 ' 中的每个元组D=2与 ' 中的至少一个元组重叠乙<. 此外，每次交换也将确保这一点，因为它只交换有冲突的元组。因此，任何元组 A>直流电∈'>直流电必须至少与 ' 中的一个元组重叠8=.

条件（5）要求任何元组A8=∈'8=必须至少与 ' 中的一个元组重叠>直流电.这是因为对齐搜索算法的第 8 行确保我们遍历 '乙<通过增加子集大小?。如果A8=∈'8=不与 ' 中的任何元组重叠>直流电,在里面
以前的迭代较小？，子集 ' ‘ 8== '8=\{A8=}应该
已经与当前的 ' 交换>直流电.

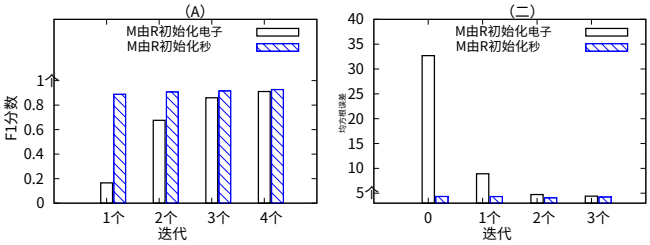

图11：时间约束下相似性对齐多次迭代的对齐精度和模型预测性能80和模型约束X =8个在房子

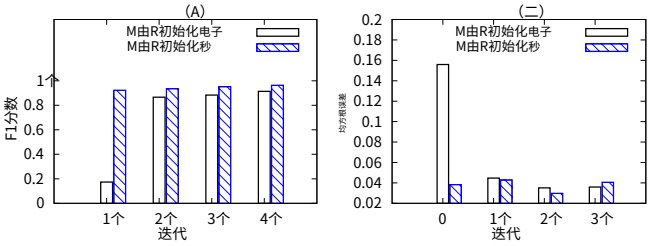

图12：时间约束下相似性对齐多次迭代的对齐精度和模型预测性能 \ =85和模型约束X =0.5在水上

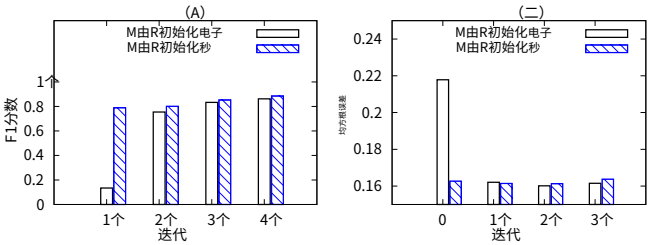

图13：时间约束下相似性对齐多次迭代的对齐精度和模型预测性能75和模型约束X =0.5关于遥测

A.9 引理 4 的证明

为了简单起见，这里我们稍微滥用吨:8表示吨:8[*:]，即时间戳吨:8.让C<0G， C<8=表示的最大和最小时间戳A8个,连续A8个和A8+1个在oa 路径中。自从A8个=A8+1个,A8个和A8+1个必须重叠一些时间戳C:,

IE，C:8=C:(8+1). 根据定义 2 中的时间约束，我们有C:8≤C<0G≤C:8+ \和C:(8+1)≤C<0G 8+1个≤C:(8+1)+ \。因此，我们得到C<0G 8+1个≤C:(8+1)+ \ =C:8+ \≤C<0G+\,IE，C<0G 8+1个 ≤C<0G+ \。通过迭代应用这个公式，我们有 C<0G≤C<0G+\ 8个 ≤C<0G+2\≤ · · · ≤C<0G+(; -1)\≤C<8= ; ;−1个 ;−2个 1个 1个 +;\。同样，我们也可以证明C<0G 1个 ≤C;<8=+; \。因此，我们得到C<0G≤C;<8=+;\和C<0G≤C<8=+;\,指示 |C<0G 1个 - C;<8=| ≤;\和 |C<0G−C<8=|≤;\. 1个 ; 1个

最后，我们有C:1-C:;|≤最大（|C<0G 1个 − C<8=|, |C<0G−C<8=; ; 1个 |) ≤ ;\,∀: ∈ {1, 2, 。. . ,<} 即 |C:1[*:] -C:;[*:]| ≤;\.

A.10 引理 5 的证明

在第 3.3.1 节的条件 (2) 和“中选候选人的要求”中乙<, 中的元组 >直流电和 '8=不应重叠

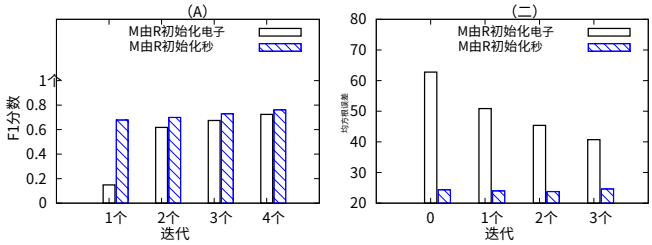

图14：时间约束下相似性对齐多次迭代的对齐精度和模型预测性能 \ =35和模型约束X =60关于空气质量

同一组中的其他元组。因此，如果A0和A1个由路径连接，路径上的元组必须从 '>直流电和 '8=，即 oa 路径。

A.11 引理 6 的证明

首先，我们将证明一个 oa-path (A0,. . . ， A1个)必须存在。自从 '8=可以被交换 '>直流电, '8=和 '>直流电应该满足命题3的条件（4）和（5），即每个元组A在 '>直流电和 '8=与至少一个其他元组重叠。如果不存在 oa-path (A0,. . . ， A1个)，这个图是不连通的，即图的至少两个顶点没有通过路径连接。不失一般性，我们可以假设断开的路径可以分为我连接部分R = {'1个, '2个,. . . , '我}, 因为每个元组A至少与另一个元组重叠，∀'8个∈ R,我们有 |'8个|≥2. 回想一下 |'>直流电| > |'8=| = ?,也就是说，必须存在 '8个∈ R,

|'8个∩'>直流电| > |'8个∩'8=| (示例参见图 10(a))。让 ' >'直流= '8个∩'>直流电和 ' 8== '8个∩'8=. 我们可以安全地交换 ' ‘ 8=和 ' >'直流,因为元组在 ' ‘> 直流不与外面的其他元组重叠 ' 8=, 和 |' ‘> 直流| > |' ‘8=|。因此，我们发现 |' ‘ 8=| < |'8=| =? 可以与'交换 ‘ > 直流.它与引理中的假设相冲突 6 算法已找到所有可能与 |' 的交换 ‘ 8=| <? 。总而言之，一个oa路径（A0,. . . ， A1个)必须存在，即这个图是一个连通图（例子见图10（b））。

接下来，我们将显示 oa-path % 的长度>0小于2? − 1. 回想一下 |'8=| =? 。最多有 ? 来自 ' 的元组8= 在 oa 路径中，即 |'8=∩%>0| ≤?。由于元组来自 '>直流电和 '8= 或者出现在 oa 路径中，以及开始和结束顶点A0 和A1个都来自 '8=, 最多有 ? − 来自 ' 的 1 个元组>直流电 在 oa 路径中，即 |'>直流电∩%>0| ≤ ?−1. 因此，结合 %>0⊆ '>直流电∪'8=, 我们有 |%>0| = |'>直流电∩%>0|+|'8=∩%>0| ≤2?−1。

总之，我们证明∀A0， A1个∈'8=, 必须存在一个 oa-path %>0 = (A0,. . . ， A1个)有长度≤2? - 1。

A.12 命题 7 的证明

通过结合引理 4 和 6，一条来自A0到A1个存在，有 |C:0[*:]−C:1[*:]| ≤ (2? −1)\,∀: ∈ {1, 2, 。. . ,<}。也就是说，对于任何元组对A0，A1个∈'8=，时间戳之间的差异A0和A1个不大于 (2? − 1)\。

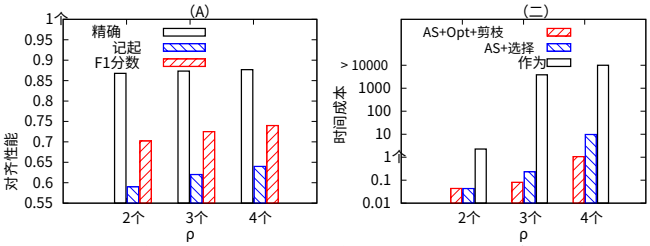

图 18：(a) 对齐性能和 (b) SAMC 的时间成本通过改变d在空气质量数据集上使用 \ =35和 X =60

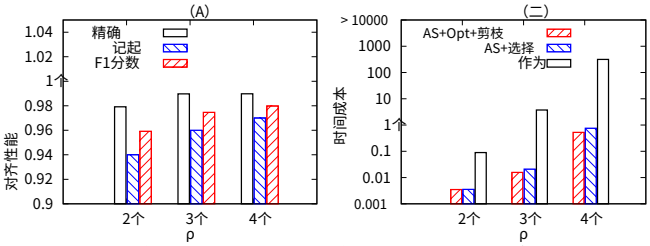

图 19：(a) 对齐性能和 (b) SAMC 的时间成本通过改变d使用 \ = 在 Fuel 数据集上120和X =35

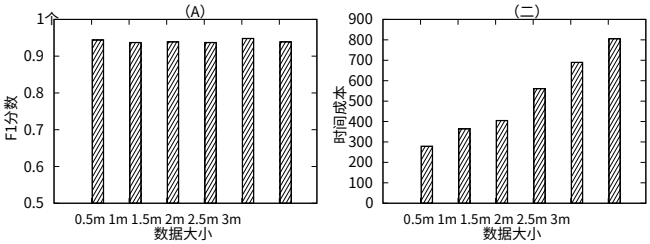

图 15：在 Apache Spark 上比 Fuel 对齐的可扩展性

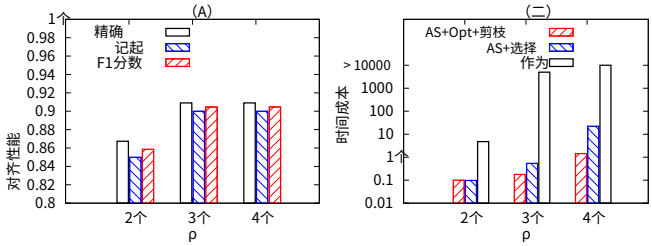

图 16：(a) 对齐性能和 (b) SAMC 的时间成本doverWater 数据集与\ =85和X =0.5

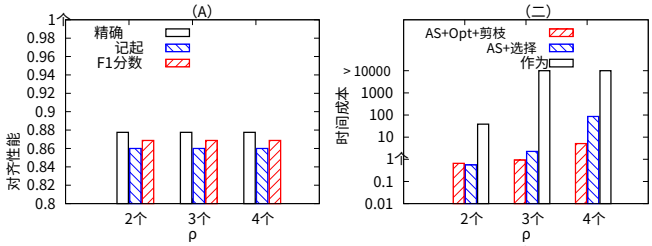

图 17：(a) 对齐性能和 (b) SAMC 的时间成本通过改变d使用 \ = 在遥测数据集上75和 X =0.5

其他结果

A.13 第 4.1 节的附加结果

为了进一步评估大规模数据的可扩展性，我们在 Apache Spark 上实施了建议的 SAMC。图 15 说明了 Fuel 数据集的结果，范围从 0.1 到 300 万行。如图所示，F1-score 总体上是稳定的，而相应的时间成本几乎呈线性增长。

A.14 第 4.2.1 节的附加结果

图 11-14 显示了在对其他数据集进行不同次数的迭代后该提议的结果。

A.15 第 4.2.2 节的附加结果

图 16-19 通过不同的方式反对提案的结果丁，与其他数据集相比，我们将考虑交换的最大集合。